

**NATIONAL TECHNICAL UNIVERSITY OF ATHENS**

**SCHOOL OF CIVIL ENGINEERING**

**DEPARTMENT OF WATER RESOURCES AND ENVIRONMENTAL ENGINEERING**

**Modelling and simulation of non-Gaussian stochastic  
processes for optimization of water-systems under  
uncertainty**

**Ph.D. Thesis**

**Ioannis Tsoukalas**

**Athens, December 2018**



**NATIONAL TECHNICAL UNIVERSITY OF ATHENS**

**SCHOOL OF CIVIL ENGINEERING**

**DEPARTMENT OF WATER RESOURCES AND ENVIRONMENTAL ENGINEERING**

**Modelling and simulation of non-Gaussian stochastic  
processes for optimization of water-systems under  
uncertainty**

**Thesis submitted for the degree of Doctor of Philosophy  
at the National Technical University Of Athens**

**Ioannis Tsoukalas**

**Athens, December 2018**

## **THESIS COMMITTEE**

### **THESIS SUPERVISOR**

Christos Makropoulos – Associate Professor, N.T.U.A.

### **ADVISORY COMMITTEE**

1. Christos Makropoulos – Associate Professor, N.T.U.A (Supervisor)
2. Nikos Mamassis – Associate Professor, N.T.U.A.
3. Dragan Savic – Professor, University of Exeter.

### **EVALUATION COMMITTEE**

1. Christos Makropoulos – Associate Professor, N.T.U.A (Supervisor)
2. Nikos Mamassis – Associate Professor, N.T.U.A.
3. Dragan Savic – Professor, University of Exeter.
4. Demetris Koutsoyiannis – Professor, N.T.U.A.
5. Evangelos Baltas – Professor, N.T.U.A.
6. George P. Karatzas – Professor, Technical University of Crete
7. Demetris F. Lekkas – Associate Professor, University of Aegean

Tsoukalas Ioannis, *Modelling and simulation of non-Gaussian stochastic processes for optimization of water-systems under uncertainty*, A Thesis submitted for the degree of Doctor of Philosophy, NTUA, Athens, 2018.

# ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΥΔΑΤΙΚΩΝ ΠΟΡΩΝ & ΠΕΡΙΒΑΛΛΟΝΤΟΣ

## MODELLING AND SIMULATION OF NON-GAUSSIAN STOCHASTIC PROCESSES FOR OPTIMIZATION OF WATER-SYSTEMS UNDER UNCERTAINTY

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ΠΡΟΣΟΜΟΙΩΣΗ ΜΗ-ΓΚΑΟΥΣΙΑΝΩΝ ΣΤΟΧΑΣΤΙΚΩΝ  
ΑΝΕΛΙΞΕΩΝ ΓΙΑ ΤΗ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΩΝ ΝΕΡΟΥ ΥΠΟ  
ΑΒΕΒΑΙΟΤΗΤΑ

**ΙΩΑΝΝΗΣ ΤΣΟΥΚΑΛΑΣ**

Πολιτικός Μηχανικός, Α.Π.Θ.

ΜΔΕ Επιστήμη Και Τεχνολογία Υδατικών Πόρων, Ε.Μ.Π.

**ΑΘΗΝΑ**

**20.12.2018**

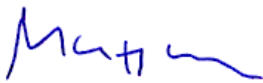
### ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ



Επιβλέπων


**ΧΡΗΣΤΟΣ ΜΑΚΡΟΠΟΥΛΟΣ**

Αναπληρωτής Καθηγητής Ε.Μ.Π.



**ΝΙΚΟΣ ΜΑΜΑΣΗΣ**

Αναπληρωτής Καθηγητής Ε.Μ.Π.



**ΔΗΜΗΤΡΗΣ ΚΟΥΤΣΟΓΙΑΝΝΗΣ**

Καθηγητής Ε.Μ.Π.



**ΓΕΩΡΓΙΟΣ ΚΑΡΑΤΖΑΣ**

Καθηγητής, Πολυτεχνείο Κρήτης




**DRAGAN SAVIC**

Professor of Hydroinformatics,  
University of Exeter



**ΕΥΑΓΓΕΛΟΣ ΜΠΑΛΤΑΣ**

Καθηγητής Ε.Μ.Π.



**ΔΗΜΗΤΡΗΣ Φ. ΛΕΚΚΑΣ**

Αναπληρωτής Καθηγητής  
Πανεπιστήμιο Αιγαίου

Copyright © Ioannis Tsoukalas, 2018

Copying, storage and distribution of this work, wholly or partly, is forbidden for commercial purposes. Reproduction, storage and distribution for non-profit purposes, educational or research activities is permitted, provided the source is indicated and the existing message is maintained.

The views and conclusions contained in this document reflect the author's view, and do not necessarily represent the views of the National Technical University of Athens.



Modelling and simulation of Non-Gaussian stochastic processes for optimization of water-systems under uncertainty by Ioannis Tsoukalas is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

To my parents,  
Kyriakos and Erasmia

## ACKNOWLEDGEMENTS

If you are reading this, a journey of mine has come to an end. It was an adventurous quest full of experiences, knowledge, and moments. Its path was not a straight line, it never is. There is no path worth walking down, paved with roses. So sailors, beware! The path has peaks, valleys and sharp turns. It entails crossings with noble men, *knights* or even *dragons*, lurking in the dark. After all, what a sort of a quest would be without these? One may ask, is it worth it? I don't know! But I can tell you this, now that I am reaching *Ithaca*<sup>1</sup>, I can see. I realize that *Ithaca*, doesn't really exist. It is a state of mind, an artifact made by men. It's all about the journey. The journey never ends, there is no *graduation* from it, just another destination. But trust me on this, there is no *Ithaca* if you walk alone. Now that I am *here*, I owe to express my sincere acknowledgments to those who walked down this path with me, the *knights* who stand by me. In their recognition, let me introduce the notion of *memorable* moments (hereafter *m*-moments). By definition, *m*-moments, are snapshots of this journey imprinted in my memory, memories that will never be forgotten.

Let me start from the *First Knight* of this quest, Associate Professor Christos Makropoulos, my Thesis supervisor. Without him, this journey would never have started. I am grateful for his continuous trust, support, guidance and belief, in both bright and dark moments, as well as for the decisive 2012 *m*-moment that made me realize that research is fun. Thanks Professor!

I would also like to thank Associate Professor Nikos Mamassis, member of my advisory committee, for the encouragement, kind and positive attitude, as well as willingness to support my efforts. An *m*-moment that I recall is one of the many that made me realize that teaching is fun.

I wish thank Professor Dragan Savic, member of my advisory committee, for his support in my efforts, belief in my research, as well as the always timely response to my requests. An *m*-moment that I recall is a 2017 moment, where Professor Savic accepted the invitation to become a member of this journey.

I also feel important to acknowledge the honorable members of my evaluation committee, Professor Demetris Koutsoyiannis, Professor Evangelos Baltas, Professor George Karatzas, and Associate Professor Demetris F. Lekkas. An *m*-moment that I recall is a 2018 moment where they all joined this journey. Specifically, I wish to thank Professor Koutsoyiannis and Professor Baltas for being supportive and encouraging in moments of need. An *m*-moment that I recall is a 2015 moment, where Professor Koutsoyiannis gave me the necessary courage and motivation to explore uncharted waters. Another *m*-moment that I recall is a 2018 moment, where Professor Baltas gave me generously some useful guidelines regarding this Thesis.

This journey, as well as its destination, wouldn't be the same without the fellowship of Dr. Andreas Efstratiadis and Panagiotis Kossieris, to whom I am grateful for the unreserved exchange of ideas and the many fruitful discussions that made studying fun. Thank you both, and I wish you the best!

---

<sup>1</sup> Please devote a moment and read the poem of Constantine P. Cavafy, *Ithaca* (1911). A reading far more inspiring than the next ten Chapters.



Particularly, I wish to thank Andreas, first of all for being a friend, for the constant support and belief, as well as the many invaluable lessons. An *m*-moment that I recall is 2016 moment that made me realize that beyond research, also writing can be fun.

Also, I cannot forget the continuous support and encouragement of my friend and comrade, Panagiotis. I recall two *m*-moments, the first is a 2012 moment that made me realize that collaboration is fun, especially with friends. The second is a 2015 moment that made me understand the inner value of giving.

I would also like to thank Dr. Maggie Kossida for her belief in me and motivational spirit, as well the many professional and personal discussions that helped me throughout these years. An *m*-moment that I recall is a 2014 moment that regards guiding me with my first journal paper.

Furthermore, I thank Dr. Simon-Michael Papalexou, for his support, positive attitude and friendship, as well as the many constructive and intriguing discussions, especially during the last year. An *m*-moment that I recall is a 2017 moment that gave me courage to carry on.

Also, I wish to acknowledge my good friend and colleague Archontia Lykou, her positive attitude, warm-hearted character and willingness to help, were indispensable elements of this journey. An *m*-moment that I recall is a 2017 moment of caring.

Moreover, I thank Dr. Ifigenia Koutiva, Patricia Gourgoura, Dr. Klio Monokrousou, Dionysis Nikolopoulos, George Moraitis and Argiro Plevri, for offering me encouragement, friendship, and moral support, as well as the many NTUA *m*-moments. Special thanks go to all the members of Itia research team, NTUA.

Last but certainly not least, I wish to thank the unsung heroes that shaped this quest.

First of all, my parents, Kyriakos and Erasmia, and my sister Eftychia for their unconditional love and untold encouragement. An *m*-moment that I recall is a 2012 moment, where my father somehow convinced me to get out my comfort zone and start this journey. Thanks dad!

Of course, I owe a lot to my beloved Efi, for her non-stop encouragement and support, for her limitless patience and for always being there. Finally, I have to mention the non-measurable, by any function, contribution of my friends back at home, VR, FF, KT, GZ, SK, PP and SV. My parents, friends and Efi, offered me unaccountably infinite, non-PhD related, *m*-moments that made life fun, kept me on path, and allow me to reach *Ithaca*. Without them *Ithaca* would be just Ithaca, an island in the Ionian sea, Greece.

Thank you all for all those *m*-moments.

Ioannis Tsoukalas

18 November 2018

# CONTENTS

---

|   |           |
|---|-----------|
| ACKNOWLEDGEMENTS .....  | viii      |
| Contents.....   | x         |
| List Of Figures.....  | xvi       |
| List Of Tables .....  | xxvii     |
| Abstract.....   | xxx       |
| Περίληψη .....  | xxxii     |
| <b>1 INTRODUCTION .....</b>   | <b>1</b>  |
| 1.1 Setting the scene .....   | 1         |
| 1.2 Stochastic modelling and simulation of hydrometeorological processes .....                                    | 2         |
| 1.3 Optimization of water-system problems under uncertainty .....   | 4         |
| 1.4 Thesis overview and contribution .....  | 6         |
| <b>2 MODELLING AND SIMULATION OF HYDROMETEOROLOGICAL PROCESSES: A REVIEW OF THE STATE-OF-THE-ART .....</b>        | <b>10</b> |
| Preamble.....   | 10        |
| 2.1 Basic concepts and definitions .....  | 11        |
| 2.2 Characteristics of hydrometeorological processes.....   | 12        |
| 2.3 Simulation schemes.....   | 13        |
| 2.3.1 Linear stochastic models .....  | 13        |
| 2.3.2 Point process models.....   | 16        |
| 2.3.3 Two-part models.....  | 16        |
| 2.3.4 Resampling models.....  | 18        |
| 2.3.5 Copula-based models.....  | 19        |
| 2.4 Summary.....  | 20        |
| <b>3 ON THE REPRODUCTION OF DEPENDENCIES THROUGH LINEAR STOCHASTIC MODELS WITH NON-GAUSSIAN WHITE NOISE .....</b> | <b>21</b> |
| Preamble.....   | 21        |
| 3.1 A glimpse of history.....   | 21        |
| 3.2 The envelope behavior of linear stochastic models with non-Gaussian white noise .....                         | 23        |
| 3.2.1 The Thomas-Fiering approach .....   | 23        |
| 3.2.2 The envelope behavior in the classical univariate AR(1) model.....  | 24        |
| 3.2.3 From the univariate to the multivariate AR(1) model.....  | 28        |
| 3.2.4 The envelope behavior beyond AR models.....   | 33        |
| 3.3 Real-World case study .....   | 34        |
| 3.4 Discussion.....   | 36        |
| 3.5 Summary.....  | 38        |

|          |   |           |
|----------|---|-----------|
| <b>4</b> | <b>NON-GAUSSIAN MODELS FOR UNCONDITIONAL, CONDITIONAL AND STOCHASTIC SIMULATION OF RANDOM VARIABLES AND PROCESSES .....</b>     | <b>39</b> |
|          | Preamble.....   | 39        |
| 4.1      | On the Nataf joint distribution model.....  | 40        |
| 4.1.1    | Introduction and historical background.....   | 40        |
| 4.1.2    | Theoretical background.....   | 40        |
| 4.1.3    | Unconditional Monte Carlo simulation.....   | 48        |
| 4.1.4    | Numerical examples .....  | 50        |
| 4.1.4.1  | Continuous-type marginal distributions.....   | 50        |
| 4.1.4.2  | Discrete-continuous-type marginal distributions.....  | 51        |
| 4.1.4.3  | Discrete-type marginal distributions .....  | 54        |
| 4.2      | A Nataf-based conditional distribution model.....   | 55        |
| 4.2.1    | Theoretical background.....   | 55        |
| 4.2.2    | Conditional Monte Carlo Simulation.....   | 57        |
| 4.2.3    | Numerical examples .....  | 57        |
| 4.3      | Nataf-based stochastic processes with arbitrary marginal distributions and correlation structure.....                           | 60        |
| 4.3.1    | Multivariate and univariate cyclostationary processes.....  | 61        |
| 4.3.2    | Multivariate and univariate stationary processes.....   | 63        |
| 4.3.3    | Selection of the target marginal distributions and correlation structures.....  | 64        |
| 4.3.4    | The auxiliary Gaussian processes.....   | 66        |
| 4.3.5    | Estimation of the equivalent correlation coefficients .....   | 66        |
| 4.3.6    | Mapping auxiliary processes to the actual domain.....   | 67        |
| 4.3.7    | Brief overview via a step-by-step procedure .....   | 67        |
| 4.3.8    | Numerical examples .....  | 67        |
| 4.3.9    | A brief note on Nataf-based stochastic models.....  | 70        |
| 4.4      | The case of mixed marginal distributions.....   | 71        |
| 4.5      | Identification of equivalent correlation coefficients.....  | 72        |
| 4.5.1    | A hybrid Monte Carlo approach.....  | 72        |
| 4.5.2    | The Log-Normal case .....   | 74        |
| 4.5.3    | A cautionary note.....  | 75        |
| 4.6      | Bits and pieces of NDM in hydrology .....   | 75        |
| 4.7      | Summary.....  | 77        |
| <b>5</b> | <b>SIMULATION OF STATIONARY STOCHASTIC PROCESSES EXHIBITING ANY-RANGE DEPENDENCE AND ARBITRARY MARGINAL DISTRIBUTIONS .....</b> | <b>78</b> |
|          | Preamble.....   | 78        |
| 5.1      | Introduction .....  | 79        |
| 5.2      | Modelling the auto-dependence structure of stationary processes.....  | 81        |

|          |   |            |
|----------|---|------------|
| 5.3      | Theoretical background of the models.....   | 85         |
| 5.4      | The auxiliary Gaussian models .....   | 86         |
| 5.4.1    | The univariate SMA model .....  | 87         |
| 5.4.2    | The multivariate SMA model.....   | 88         |
| 5.4.3    | The univariate AR model .....   | 90         |
| 5.4.4    | The multivariate AR model.....  | 91         |
| 5.4.5    | A note on the computation of the square root matrix.....  | 93         |
| 5.5      | Generation procedure .....  | 94         |
| 5.6      | Hypothetical simulation studies .....   | 95         |
| 5.6.1    | SMARTA model.....   | 95         |
| 5.6.1.1  | Simulation of univariate processes.....   | 95         |
| 5.6.1.2  | Simulation of multivariate processes .....  | 99         |
| 5.6.2    | CMARTA model.....   | 102        |
| 5.7      | Real-world simulation studies .....   | 106        |
| 5.7.1    | Simulation of multivariate annual streamflow processes.....   | 106        |
| 5.7.2    | Simulation of univariate daily rainfall process.....  | 108        |
| 5.8      | Summary.....  | 111        |
| <b>6</b> | <b>SIMULATION OF CYCLOSTATIONARY STOCHASTIC PROCESSES WITH ARBITRARY MARGINAL DISTRIBUTIONS .....</b> | <b>113</b> |
|          | Preamble.....   | 113        |
| 6.1      | Introduction .....  | 114        |
| 6.2      | SPARTA at a glance.....   | 114        |
| 6.3      | The auxiliary Gaussian PAR model .....  | 116        |
| 6.3.1    | Multivariate contemporaneous PAR(1) model.....  | 117        |
| 6.3.2    | Univariate PAR(1) model.....  | 118        |
| 6.4      | Generation procedure of sparta model .....  | 118        |
| 6.5      | Case Studies.....   | 119        |
| 6.5.1    | Univariate simulation with common distribution models.....  | 119        |
| 6.5.2    | Toy simulation with seasonally-varying distribution models.....                                       | 122        |
| 6.5.3    | Multivariate simulation .....   | 125        |
| 6.6      | Summary.....  | 130        |
| <b>7</b> | <b>BUILDING A PUZZLE FOR MULTI-TEMPORAL STOCHASTIC SIMULATION .....</b>                               | <b>132</b> |
|          | Preamble.....   | 132        |
| 7.1      | Introduction .....  | 133        |
| 7.2      | Addressing multi-scale consistency .....  | 136        |
| 7.2.1    | Problem description .....   | 136        |
| 7.2.2    | The NDA approach: Step-by-step implementation.....  | 137        |
| 7.2.3    | Computational details .....   | 138        |

|          |   |            |
|----------|---|------------|
| 7.3      | Modular framework for developing multi-temporal simulation schemes .....  | 139        |
| 7.3.1    | Multi-temporal stochastic simulation as a puzzle.....   | 139        |
| 7.3.2    | Three-level configuration for annual to daily simulations.....  | 140        |
| 7.4      | Case study A: multi-temporal simulation of daily processes.....   | 142        |
| 7.5      | Case study B: Disaggregation of daily rainfall to hourly scale.....   | 151        |
| 7.6      | Summary.....  | 154        |
| <b>8</b> | <b>MULTI-OBJECTIVE OPTIMIZATION ON A BUDGET: EXPLORING SURROGATE MODELLING FOR ROBUST MULTI-RESERVOIR RULES GENERATION UNDER HYDROLOGICAL UNCERTAINTY .....</b> | <b>158</b> |
|          | Preamble.....   | 158        |
| 8.1      | Introduction .....  | 159        |
| 8.2      | Methodology.....  | 160        |
| 8.2.1    | Overall conceptual approach.....  | 160        |
| 8.2.2    | Models and tools.....   | 162        |
| 8.2.3    | Fundamentals of SBO algorithms.....   | 163        |
| 8.2.3.1  | Initial sampling plan .....   | 164        |
| 8.2.3.2  | Surrogate models.....   | 164        |
| 8.2.3.3  | Infill criteria.....  | 166        |
| 8.2.3.4  | Model assessment and validation.....  | 169        |
| 8.2.4    | The deployed MOSBO algorithms.....  | 169        |
| 8.3      | The study area: The hydro-system of Nestos, Greece .....  | 172        |
| 8.4      | Benchmarking the algorithms' performance .....  | 174        |
| 8.4.1    | Hypervolume indicator .....   | 175        |
| 8.4.2    | Unary $\epsilon$ -indicator (epsilon indicator).....  | 175        |
| 8.4.3    | Empirical attainment function .....   | 175        |
| 8.5      | Experimental setup.....   | 176        |
| 8.6      | Results and discussion.....   | 176        |
| 8.6.1    | Comparison and benchmarking results .....   | 176        |
| 8.6.2    | Results for the case study.....   | 184        |
| 8.7      | Summary.....  | 185        |
| <b>9</b> | <b>SURROGATE-ENHANCED EVOLUTIONARY ANNEALING SIMPLEX ALGORITHM FOR EFFECTIVE OPTIMIZATION OF WATER RECOURCES PROBLEMS ON A BUDGET .....</b>                     | <b>187</b> |
|          | Preamble.....   | 187        |
| 9.1      | Introduction .....  | 188        |
| 9.2      | Optimization methodology.....   | 189        |
| 9.2.1    | Evolutionary Annealing-Simplex.....   | 189        |
| 9.2.2    | Surrogate-Enhanced Evolutionary Annealing-Simplex.....  | 190        |
| 9.2.2.1  | Overview of SEEAS algorithm.....  | 190        |

|           |   |            |
|-----------|---|------------|
| 9.2.2.2   | Surrogate model (RBF) .....   | 191        |
| 9.2.2.3   | Acquisition function .....  | 192        |
| 9.2.2.4   | Detailed description of SEEAS .....   | 192        |
| 9.3       | Benchmarking methodology .....  | 196        |
| 9.3.1     | Benchmarking protocol.....  | 196        |
| 9.3.2     | Performance evaluation approach.....  | 197        |
| 9.3.3     | Brief description of benchmarking optimization algorithms.....                          | 198        |
| 9.3.3.1   | Dynamically Dimensioned Search (DDS) .....  | 198        |
| 9.3.3.2   | Multistart Local Metric Stochastic RBF algorithm (MLMSRBF) .....                        | 198        |
| 9.3.3.3   | DYnamic COordinate Search-Multistart Local Metric Stochastic RBF (DYCORS-LMSRBF)<br>198 |            |
| 9.4       | Test functions.....   | 199        |
| 9.4.1     | Setup of optimization problems.....   | 199        |
| 9.4.2     | Statistical evaluation of optimal solutions.....  | 199        |
| 9.4.3     | Evaluation of convergence behavior .....  | 201        |
| 9.4.4     | Sensitivity analysis against input parameters of SEEAS .....                            | 204        |
| 9.4.5     | Suitability assessment based on stochastic dominance .....                              | 205        |
| 9.5       | Hydrological calibration .....  | 206        |
| 9.5.1     | Study area, simulation model and calibration setup.....                                 | 206        |
| 9.5.2     | Model calibration with unknown parameters.....  | 207        |
| 9.5.3     | Toy calibration with synthetic runoff.....  | 209        |
| 9.6       | Optimization of multi-reservoir system performance .....                                | 210        |
| 9.6.1     | Problem statement.....  | 210        |
| 9.6.2     | The parameterization-simulation-optimization scheme.....                                | 210        |
| 9.6.3     | Results.....  | 211        |
| 9.7       | Summary.....  | 212        |
| <b>10</b> | <b>CONCLUSIONS AND DISCUSSION .....</b>   | <b>214</b> |
| 10.1      | Stochastic modelling and simulation of hydrometeorological processes .....              | 214        |
| 10.2      | Optimization of water-system problems under uncertainty .....                           | 216        |
| 10.3      | Overall conclusions and future research.....  | 217        |
|           | <b>REFERENCES.....</b>  | <b>218</b> |
| <b>A</b>  | <b>APPENDIX A .....</b>   | <b>251</b> |
| A.1       | The univariate cyclostationary Thomas-fiering model.....                                | 251        |
| A.2       | Supplementary material of Chapter 3.....  | 252        |
| <b>B</b>  | <b>APPENDIX B .....</b>   | <b>256</b> |
| B.1       | Supplementary material of Chapter 5.....  | 256        |
| <b>C</b>  | <b>APPENDIX C .....</b>   | <b>263</b> |
| C.1.      | The multivariate contemporaneous PAR(1) model .....                                     | 263        |

|          |   |            |
|----------|---|------------|
| C.2.     | Supplementary material of Chapter 6.....                                | 264        |
| <b>D</b> | <b>APPENDIX D.....</b>  | <b>269</b> |
| D.1      | Supplementary material of Section 7.4.....                              | 269        |
| D.2      | Multi-temporal simulation of multivariate daily rainfall processes..... | 272        |
| D.3      | Supplementary material of Section 7.5.....                              | 290        |
|          | <b>LIST OF PUBLICATIONS.....</b>  | <b>294</b> |

# LIST OF FIGURES

**Figure 3.1** | Comparison of the (A) January–February, (B) March–April, and (C) September–October dependence patterns between historical and synthetic monthly runoff data ( $10^9 \text{ m}^3$ ) of the Nile, at Aswan dam. Synthetic time series were generated by the cyclostationary Thomas–Fiering (TF) approach (adapted by *Tsoukalas et al. [2018e]*; the simulated negative values were not truncated to zero in order to avoid distortion of the dependence pattern). The red line (—) depicts the envelope equation of the TF model (when combined with PIII white noise. See also Appendix A). ..... 23

**Figure 3.2** | Relationship between (A) the target skewness coefficient of process  $\underline{x}_t$  and the required skewness for white noise term  $\underline{\varepsilon}_t$  for a given lag-1 autocorrelation coefficient  $\rho_1$ ; and (B) the lag-1 autocorrelation coefficient  $\rho_1$  and the required skewness coefficient of white noise term  $\underline{\varepsilon}_t$  to attain the target skewness coefficient of process  $\underline{x}_t$ . ..... 25

**Figure 3.3** | Scatter plots depicting the simulated (using the TF model, i.e., the autoregressive model of order 1 (AR(1))-PIII) lag-1 dependence pattern among consecutive time steps (i.e., pair values (•) of the previous and current time steps). The labels of each plot resemble the corresponding scenarios of **Table 3-1**. The red line (—) depicts the envelope equation shown in the title of each plot. .... 28

**Figure 3.4** | Scatter plots depicting the simulated (using the contemporaneous multivariate autoregressive model of order 1 (CMAR(1) model) with PIII white noise) for (A) and (B) lag-1 dependence patterns of the zero-autocorrelated processes  $\underline{x}_t^1$  and  $\underline{x}_t^2$ , respectively, for consecutive time steps (i.e., pair values (•) of the previous and current time steps). Panel (C) depicts the contemporaneous dependence (lag-0) of  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . The red line (—) depicts the envelope equation shown in the title of each plot. Panel (D) compares the simulated and theoretical autocorrelation function (ACF) of  $\underline{x}_t^1$  while panel (E) compares that of  $\underline{x}_t^2$ . Finally, panel (F) compares the simulated and theoretical cross-correlation function (CCF) of  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . ..... 32

**Figure 3.5** | Scatter plots depicting the simulated (using the CMAR(1) model with PIII white noise) for (A) and (B) lag-1 dependence pattern of the autocorrelated processes  $\underline{x}_t^1$  and  $\underline{x}_t^2$ , respectively, for consecutive time steps (i.e., pair values (•) of the previous and current time steps), while panel (C) depicts the contemporaneous dependence (lag-0) of  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . The red line (—) depicts the envelope equation shown in the title of each plot. Panel (D) compares the simulated and theoretical ACF of  $\underline{x}_t^1$  while panel (E) compares that of  $\underline{x}_t^2$ . Lastly, panel (F) compares the simulated and theoretical CCF of  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . ..... 33

**Figure 3.6** | Scatter plots depicting the simulated lag-1 dependence pattern among consecutive time steps (i.e., pair values (•) of the previous and current time steps) obtained by: (A) ARMA(1,1)-PIII; (B) MA(32)-PIII; and (C) SMA(32)-PIII models. Comparison of synthetic and theoretical autocorrelation function (ACF) obtained by: (D) ARMA(1,1)-PIII; (E) MA(32)-PIII; and (F) SMA(32)-PIII models. 34

**Figure 3.7** | Scatter plots showing the lag-1 dependence pattern of the daily streamflow ( $\text{m}^3/\text{s}$ ) of the Achelous river at the Kremasta dam, Greece (orange dots; •) and of a synthetic time series generated using an AR(1)- PIII model (black dots; •). The red line (—) depicts the envelope equation embedded each plot. .... 35

**Figure 4.1** | Graphical illustration of function  $F(\cdot)$  (i.e., Eq. (4.15)) that expresses the relationship between the equivalent,  $\tilde{\rho}_{1,2}$  and target  $\tilde{\rho}_{1,2}$  correlation coefficients assuming that both  $x_1$  and  $x_2$  are described by the two-parameter Gamma distribution (assuming that  $b := b_1 = b_2 = 1$ ) with a) equal shape parameters (i. e.,  $a := a_1 = a_2$ ) and b) different shape parameters by setting  $a_1 = 5$  and varying  $a_2$  from 5 to 0.01. .... 45

**Figure 4.2** | Hypothetical example of two RVs,  $\underline{x}_1, \underline{x}_2 \sim G(a = 0.5, b = 1)$  with  $\rho_{1,2} \in \{0.1, 0.4, 0.8, 0.95\}$ . Each row of subplots corresponds to a specific value of  $\rho_{1,2}$  and each column of



subplots, from left to right, depicts the joint PDF in the Gaussian and uniform domain as well the joint CDF in the actual domain. .... 47

**Figure 4.3** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) identical Gamma-distributed variables with  $G(0.7, 10)$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients (i.e., Eq. (4.23)). Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\hat{\rho}}_{12} = 0.7504$  and  $\hat{\rho}_{12} = 0.7073$  respectively. .... 51

**Figure 4.4** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) zero-inflated Gamma-distributed variables with identical continuous part  $G(0.7, 10)$  and  $p_{0;\underline{x}_1} = p_{0;\underline{x}_2} = 0.9$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\hat{\rho}}_{12} = 0.8431$  and  $\hat{\rho}_{12} = 0.6989$  respectively..... 52

**Figure 4.5** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) zero-inflated Gamma-distributed variables with identical continuous part  $G(0.7, 10)$ ,  $p_{0;\underline{x}_1} = 0.9$  and  $p_{0;\underline{x}_2} = 0.6$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\hat{\rho}}_{12} = 0.8795$  and  $\hat{\rho}_{12} = 0.7071$  respectively..... 53

**Figure 4.6** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) zero-inflated Gamma-distributed variables with identical continuous part  $G(0.7, 10)$ ,  $p_{0;\underline{x}_1} = 0.9$  and  $p_{0;\underline{x}_2} = 0$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\hat{\rho}}_{12} = 0.9422$  and  $\hat{\rho}_{12} = 0.7096$  respectively..... 53

**Figure 4.7** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) identical Poisson-distributed variables with  $Poi(0.5)$ . a) The established relationship between equivalent,  $\rho$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\hat{\rho}}_{12} = 0.8193$  and  $\hat{\rho}_{12} = 0.7054$  respectively. .... 54

**Figure 4.8** | Bivariate example of  $\underline{x}_1 \sim G(2, 10)$  and  $\underline{x}_2 \sim LN(0.10, 4)$  with  $\rho = -0.9$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ . .... 58

**Figure 4.9** | Bivariate example of  $\underline{x}_1 \sim G(2, 10)$  and  $\underline{x}_2 \sim LN(0.10, 4)$  with  $\rho = -0.7$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for for  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ . .... 58

**Figure 4.10** | Bivariate example of  $\underline{x}_1 \sim G(2, 10)$  and  $\underline{x}_2 \sim LN(0.10, 4)$  with  $\rho = -0.5$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for for  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ . .... 58

**Figure 4.11** | Bivariate example of  $\underline{x}_1 \sim G(2, 10)$  and  $\underline{x}_2 \sim LN(0.10, 4)$  with  $\rho = 0.0$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for xfor  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ . .... 59

**Figure 4.12** | Bivariate example of  $\underline{x}_1 \sim G(2, 10)$  and  $\underline{x}_2 \sim LN(0.10, 4)$  with  $\rho = 0.5$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for for  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ . .... 59

|   |     |
|---|-----|
| <b>Figure 4.13</b>   Bivariate example of $\underline{x}_1 \sim G(2, 10)$ and $\underline{x}_2 \sim LN(0.10, 4)$ with $\rho = 0.7$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for for $\underline{x}_2 = 45$ and $\underline{x}_2 = 65$ .  | 59  |
| <b>Figure 4.14</b>   Bivariate example of $\underline{x}_1 \sim G(2, 10)$ and $\underline{x}_2 \sim LN(0.10, 4)$ with $\rho = 0.9$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for for $\underline{x}_2 = 45$ and $\underline{x}_2 = 65$ .  | 59  |
| <b>Figure 4.15</b>   Effect of dependence parameter (i.e., Pearson's correlation coefficient) on the derived quantiles of $x_1 x_2$ ; visualized as a function of $\rho$ for a) $\underline{x}_2 = 45$ and b) $\underline{x}_2 = 65$ .  | 60  |
| <b>Figure 4.16</b>   Hypothetical example of a zero-inflated stationary process with $p_0 = 0.8$ and continuous Gamma-distributed part $G(0.7, 10)$ . a) Simulated time series of 5 000 time steps. b) Comparison between empirical histogram and theoretical PDF.  | 68  |
| <b>Figure 4.17</b>   a) The established relationship between equivalent, $\tilde{\rho}$ and target $\rho$ correlation coefficients. b) Theoretical and simulated CDFs (using Weibull's plotting position). c) Theoretical, equivalent and simulated autocorrelation functions (ACF). d) Scatter plot depicting the established lag-1 dependence pattern among consecutive time steps.   | 68  |
| <b>Figure 4.18</b>   Hypothetical example of a non-stationary process with Log-Normal marginal distribution. a) All 5 000 realizations (three of which are depicted with distinct colors), each consisted of 100 time steps. Comparison of theoretical and ensemble b) mean and c) variance as a function of time. Comparison between empirical histogram and theoretical PDF for time d) $t = 25$ , e) $t = 50$ and f) $t = 75$ .  | 69  |
| <b>Figure 4.19</b>   a) Theoretical and b) simulated correlation as a function of absolute time. c) Absolute difference between theoretical and simulated correlation coefficients.   | 69  |
| <b>Figure 5.1</b>   a) Autocorrelation functions and b) climacograms of HK processes exhibiting different Hurst coefficients (dashed lines) and their approximation with the CAS (continuous line).   | 84  |
| <b>Figure 5.2</b>   Graphical illustration of the relationship between the required skewness coefficient $C_{s_v}$ of innovation term $v_t$ and a) the skewness $C_{s_x}$ of an fGn process $\underline{x}_t$ for various values of $H$ and b) the value of $H$ of an fGn process $\underline{x}_t$ for various values of skewness of $C_{s_x}$ (using the SMA model with $q = 2^{10}$ ).   | 89  |
| <b>Figure 5.3</b>   a) The established relationship between equivalent, $\tilde{\rho}$ and target $\rho$ correlation coefficients. b) Comparison between the target and equivalent autocorrelation coefficients employed within the SMARTA model for HK processes with the various values of $H$ .  | 96  |
| <b>Figure 5.4</b>   Comparison between theoretical and simulated CDFs (using the Weibull's plotting position) of SMA-PIII and SMARTA models for HK processes with a) $H = 0.6$ , b) $H = 0.7$ , c) $H = 0.8$ , d) $H = 0.9$ . Comparison between theoretical (HK) and empirical ACF of SMA-PIII and SMARTA models for HK processes with e) $H = 0.6$ , f) $H = 0.7$ , g) $H = 0.8$ , h) $H = 0.9$ . Comparison between theoretical and empirical climacograms of SMA-PIII and SMARTA models models for HK processes with i) $H = 0.6$ , j) $H = 0.7$ , k) $H = 0.8$ , l) $H = 0.9$ . Scatter plots of SMA-PIII and SMARTA models for time lag $\tau = 1$ for simulated HK processes with m) $H = 0.6$ , n) $H = 0.7$ , o) $H = 0.8$ , p) $H = 0.9$ . Scatter plots of SMA-PIII and SMARTA models for time lag $\tau = 10$ for simulated HK processes with q) $H = 0.6$ , r) $H = 0.7$ , s) $H = 0.8$ , t) $H = 0.9$ . | 98  |
| <b>Figure 5.5</b>   Comparison between theoretical (red dots, $\bullet$ ) and simulated lag-1 autocorrelation and Hurst coefficient for sites A-D. Target (red dots, $\bullet$ ) and simulated lag-0 cross-correlation coefficients for all pairs of sites A-D.   | 100 |
| <b>Figure 5.6</b>   (a-d) Theoretical and simulated (SMARTA) distribution functions (using the Weibull's plotting position) for sites A-D. (e-h) The established relationships between equivalent, $\rho$ and target $\rho$ correlation coefficients given the marginal distribution of sites A-D. (i-l) Theoretical and simulated  |     |

ACFs for sites A-D. (m-p) Theoretical and simulated climacograms (CGs) for sites A-D. In all cases, the simulation intervals have been established using all 100 realizations..... 101

**Figure 5.7** | Case A – Continuous marginal distributions. Simulated realization of process a)  $x_{t1}$  and b)  $x_{t2}$ . Comparison of simulated and theoretical distribution function for process c)  $\underline{x}_t^1$  and d)  $\underline{x}_t^2$ . Simulated, equivalent and theoretical autocorrelation function (ACF) for process e)  $\underline{x}_t^1$  and f)  $\underline{x}_t^2$ . g) Simulated and theoretical climacogram for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . h) Simulated and theoretical lag-1 autocorrelation ( $\rho_1^{(k)}$ ) as a function of scale k for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of each process i)  $\underline{x}_t^1$  j)  $\underline{x}_t^2$ , as well as their k) interaction. Simulated dependence pattern for time lag 1 for process l)  $\underline{x}_t^1$  and m)  $\underline{x}_t^2$ . n) Simulated lag 0 dependence pattern among  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . ..... 103

**Figure 5.8** | Case B – Poisson marginal distributions. Simulated realization of process a)  $\underline{x}_t^1$  and b)  $\underline{x}_t^2$ . Comparison of simulated and theoretical distribution function for process c)  $\underline{x}_t^1$  and d)  $\underline{x}_t^2$ . Simulated, equivalent and theoretical autocorrelation function (ACF) for process e)  $\underline{x}_t^1$  and f)  $\underline{x}_t^2$ . g) Simulated and theoretical climacogram for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . h) Simulated and theoretical lag-1 autocorrelation ( $\rho_1^{(k)}$ ) as a function of scale k for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of each process i)  $\underline{x}_t^1$  j)  $\underline{x}_t^2$ , as well as their k) interaction. Simulated dependence pattern for time lag 1 for process l)  $\underline{x}_t^1$  and m)  $\underline{x}_t^2$ . n) Simulated lag 0 dependence pattern among  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . ..... 104

**Figure 5.9** | Case C – Bernoulli marginal distributions. Simulated realization of process a)  $\underline{x}_t^1$  and b)  $\underline{x}_t^2$ . Comparison of simulated and theoretical distribution function for process c)  $\underline{x}_t^1$  and d)  $\underline{x}_t^2$ . Simulated, equivalent and theoretical autocorrelation function (ACF) for process e)  $\underline{x}_t^1$  and f)  $\underline{x}_t^2$ . g) Simulated and theoretical climacogram for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . h) Simulated and theoretical lag-1 autocorrelation ( $\rho_1^{(k)}$ ) as a function of scale k for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of each process i)  $\underline{x}_t^1$  j)  $\underline{x}_t^2$ , as well as their k) interaction. Simulated dependence pattern for time lag 1 for process l)  $\underline{x}_t^1$  and m)  $\underline{x}_t^2$ . n) Simulated lag 0 dependence pattern among  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . ..... 105

**Figure 5.10** | Synopsis of annual streamflow simulation study at 4 stations in New South Wales region. (a-d) Historical time series. (e-h) Empirical, simulated and theoretical distribution functions (using the Weibull’s plotting position) for stations ID1-4 (i-l) The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of stations ID1-4. (m-p) Empirical, simulated and theoretical ACFs for stations ID1-4. (q-t) Empirical, simulated and theoretical climacograms (CGs) for stations ID1-4. .... 107

**Figure 5.11** | Synopsis of daily rainfall simulation at Pavlos’ station. a) Historical time series. b) Synthetic time series; randomly selected window of 60 years. Empirical, simulated and theoretical distribution function of positive precipitation amounts for c) February, d) June and e) October (using the Weibull’s plotting position); the title of each plot provides the parameters of the GG distribution, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients for the mixed and GG distribution for f) February, g) June and h) October. Empirical, simulated and theoretical ACF for i) February, j) June and k) October; the title of each plot depicts the parameters of CAS. Empirical and simulated dependence pattern for time lag 1 for l) February, m) June and n) October; the title of each plot depicts the lag-1, target  $\rho_1^{CAS}$ , simulated  $\hat{\rho}_1$ , and equivalent  $\tilde{\rho}_1$  autocorrelation coefficients. .... 110

**Figure 6.1** | Comparison of key statistics ( $\mu$ ,  $\sigma$ ,  $C_s$  and  $\rho_1$ ) between historical and simulated flow data of Nile River (PAR and SPARTA). ..... 120

**Figure 6.2** | Comparison between simulated flow data ( $10^9 \text{ m}^3$ ), through PAR-PIII and SPARTA-PIII, empirical and theoretical cumulative distribution functions (Weibull’s plotting position). Simulated negative values are also included to avoid the distortion of the established CDFs. .... 121

|   |     |
|---|-----|
| <b>Figure 6.3</b>   Month-to-month scatter plots of historical and simulated flow data ( $10^9 \text{ m}^3$ ), through PAR-PIII and SPARTA-PIII. Simulated negative values are also included to avoid the distortion of the established dependence patterns.....  | 122 |
| <b>Figure 6.4</b>   Scatter plots with histograms for a) season 12 vs. 1 b) season 1 vs. 2, c) season 5 vs. 6, and d) season 10 vs. 11. ....  | 124 |
| <b>Figure 6.5</b>   Comparison between simulated (SPARTA) and theoretical cumulative distribution functions (Weibull plotting position) of hypothetical process. Simulated negative values (season 5 and 10) are also included to avoid the distortion of the established CDFs.....   | 125 |
| <b>Figure 6.6</b>   Comparison of monthly mean values, $\mu$ , of historical and synthetic data. ....   | 127 |
| <b>Figure 6.7</b>   Comparison of monthly standard deviation values, $\sigma$ , of historical and synthetic data. .   | 128 |
| <b>Figure 6.8</b>   Comparison of monthly skewness coefficients, $C_s$ , of historical and synthetic data.....  | 128 |
| <b>Figure 6.9</b>   Comparison of month-to-month lag-1 correlations, $\rho_1$ , of historical and synthetic data. ....  | 129 |
| <b>Figure 6.10</b>   Comparison of monthly lag-0 cross-correlations, $\rho_0$ , between sites of historical and synthetic data.....   | 129 |
| <b>Figure 6.11</b>   Scatter plots of 500 000 synthetic data for sites A and C, representing monthly runoff (mm) processes at Evinos and Mornos reservoirs, respectively, for (a) January and (b) February. Simulated negative values are also included to avoid the distortion of the established dependence patterns. ....                              | 130 |
| <b>Figure 7.1</b>   The stochastic simulation framework as a puzzle, involving a chain implementation of individual NDA <i>pieces</i> . ....  | 139 |
| <b>Figure 7.2</b>   a-b) Historical daily rainfall-runoff time series (1 January 1970 to 31 December 2008). c-d) Synthetically generated time series (randomly selected window of 40 years).....  | 143 |
| <b>Figure 7.3</b>   Rainfall-runoff series: (a-b) Historical annual time series. (c-d) Empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). (e-f) Empirical, simulated and theoretical ACFs. (g-h) Synthetic annual time series (randomly selected window of 1 000 years).....                            | 144 |
| <b>Figure 7.4</b>   Comparison of monthly empirical and simulated L-Mean, L-Scale and L-Skewness, as well as historical and simulated lag-1 month-to-month correlations. ....   | 145 |
| <b>Figure 7.5</b>   Comparison of monthly historical and simulated lag-0 cross-correlations. ....   | 145 |
| <b>Figure 7.6</b>   Monthly rainfall - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry..... | 146 |
| <b>Figure 7.7</b>   Monthly runoff - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.....   | 146 |
| <b>Figure 7.8</b>   Monthly runoff (mm) - month-to-month scatter plots of historical and simulated series. The title of each subplot provides the lag-1 month-to-month target $\rho_{s,s-1}$ and simulated $\hat{\rho}_{s,s-1}$ correlation coefficients. ....  | 147 |
| <b>Figure 7.9</b>   Comparison of daily empirical and simulated L-Mean, L-Scale, L-Skewness, as well as probability dry.....  | 148 |
| <b>Figure 7.10</b>   Comparison of daily historical and simulated lag-0 cross-correlations.....   | 148 |

|   |     |
|---|-----|
| <b>Figure 7.11</b>   Daily non-zero rainfall - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. ....  | 149 |
| <b>Figure 7.12</b>   Daily non-zero runoff - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. ....  | 149 |
| <b>Figure 7.13</b>   Daily rainfall - monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF); the parameters of CAS are given on the title of each subplot. ....   | 150 |
| <b>Figure 7.14</b>   Daily runoff - monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF); the parameters of CAS are given on the title of each subplot. ....   | 150 |
| <b>Figure 7.15</b>   Empirical ( $\bullet$ ) and simulated ( $\bullet$ ) daily annual rainfall-runoff maxima, as a function of the return period. The solid red line ( $-$ ) depicts the fitted to historical data Generalized Extreme Value (GEV) distribution (parameters: location (c), scale (b) and shape (a)). The dashed blue line ( $- -$ ) represents the 95% confidence intervals (estimated using the parametric bootstrap method).....  | 151 |
| <b>Figure 7.16</b>   Historical a) daily and b) hourly rainfall series. c) Synthetic (disaggregated) hourly rainfall realization. d-f) Comparison of distribution function of non-zero amounts for hourly historical and disaggregated series for February, June and October respectively (the fitted theoretical model is shown with red line). g-i) Comparison of autocorrelation function (ACF) for hourly historical and disaggregated series for February, June and October respectively (the fitted theoretical model is shown with red line). .... | 153 |
| <b>Figure 7.17</b>   Comparison of empirical and disaggregated, a-c) L-mean ( $L_1^{(k)}$ ), d-f) L-scale ( $L_2^{(k)}$ ), g-i) L-skewness ( $L_{CS}^{(k)}$ ), j-l) probability dry ( $p_0^{(k)}$ ) and m-o) lag-1 autocorrelation coefficient ( $\rho_1^{(k)}$ ), as a function of aggregation scale k, for February, June and October. ....   | 154 |
| <b>Figure 8.1</b>   Schematic representation of the conceptual approach. ....   | 161 |
| <b>Figure 8.2</b>   Flowchart of the surrogate-based optimization process. ....   | 164 |
| <b>Figure 8.3</b>   Behavior of generalized exponential correlation function (univariate case; for simplicity the index $i$ is omitted) with, (a) varying $\theta$ and fixed $p = 2$ and (b) with varying $p$ and fixed $\theta = 1$ . ....   | 166 |
| <b>Figure 8.4</b>   Graphical representation of hypervolume and hypervolume contribution. a) The light grey depicts the hypervolume of the Pareto set (non-dominated region) b) The dark grey depicts the exclusive (contribution) hypervolume of a point $\mathbf{p}$ (adopted from: [Couckuyt et al., 2013]).....   | 168 |
| <b>Figure 8.5</b>   ParEGO algorithm flowchart. ....  | 170 |
| Figure 8.6. ....  | 172 |
| <b>Figure 8.7</b>   Geographical representation of the river basin Mesta / Nestos [Skoulikaris et al., 2008]. ....  | 172 |
| <b>Figure 8.8</b>   Hydrosystem modelled in WEAP21 and detail (inside the circle) of simulation of pump-storage. Symbol ( $\bullet$ ) represents the catchments, ( $\bullet$ ) represents the demand nodes, ( $\otimes$ ) represents in-stream flow requirements, ( $\rightarrow$ ) represents the river, ( $\bullet$ ) represents river nodes or junctions and ( $\blacktriangle$ ) represents the reservoirs. ....  | 174 |
| <b>Figure 8.9</b>   Average performance of algorithms for various computational budgets.....  | 177 |
| <b>Figure 8.10</b>   Empirical CDF of all MOSBO and MOEA for 200 and 400 function evaluations; also the CDFs of MOEA for 1000 function evaluation are depicted.....   | 177 |
| <b>Figure 8.11</b>   EAF difference plot for ParEGO and SMS-EGO for 400 function evaluations.....   | 179 |

|  |     |
|--|-----|
| <b>Figure 8.12</b>   EAF difference plot for ParEGO and SUMO for 400 function evaluations. ....  | 180 |
| <b>Figure 8.13</b>   EAF difference plot for SMS-EGO and SUMO for 400 function evaluations. ....   | 180 |
| <b>Figure 8.14</b>   EAF difference plot for SMS-EMOA and NSGAI for 1 000 function evaluations. ....   | 181 |
| <b>Figure 8.15</b>   EAF difference plot for SMS-EMOA and SUMO for different computational budget: 1 000 and 400 function evaluations respectively. ....   | 182 |
| <b>Figure 8.16</b>   Comparison of SUMO with MOEA for 2 000 and 5 000 function evaluations. Blue diamond represents the median of each column. In left panel (NHVR) higher values are preferred. In the right panel (Ie) lower values are preferred. ....  | 184 |
| <b>Figure 8.17</b>   Energy-duration curves and monthly energy characteristics of the case study for the upper left point (Ulp) and Lower right point (Lrp) of the best Pareto front. ....   | 185 |
| <b>Figure 9.1</b>   Approximated surface (RBF) in a 2-D example (Ackley function) using all available sample points (left panel). The right panel demonstrates a randomly selected simplex and the modified surrogate-enhanced reflection movement using candidate points on the line formed from the simplex centroid and the maximum reflection point. The simplex is reflected at the candidate point with the minimum function value. .... | 191 |
| <b>Figure 9.2</b>   Outline of SEEAS algorithm following the steps explained in section 9.2.2.4 (* denotes the use of the surrogate model within the associated simplex transformations). ....   | 193 |
| <b>Figure 9.3</b>   Convergence curves for test function OF1 (Sphere) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d). ....  | 201 |
| <b>Figure 9.4</b>   Convergence curves for test function OF2 (Ackley) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d). ....  | 202 |
| <b>Figure 9.5</b>   Convergence curves for test function OF3 (Griewank) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d). ....  | 202 |
| <b>Figure 9.6</b>   Convergence curves for test function OF4 (Zakharov) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d). ....  | 203 |
| <b>Figure 9.7</b>   Convergence curves for test function OF5 (Rastrigin) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d). ....   | 203 |
| <b>Figure 9.8</b>   Convergence curves for test function OF6 (Levy) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d). ....  | 204 |
| <b>Figure 9.9</b>   Initial part of convergence curves up to 100 function evaluations, for test functions Levy (a) and Griewank (b), for the case of MFE=500 and 15 variables. ....  | 204 |
| <b>Figure 9.10</b>   Convergence curves for MFE = 500 (a) and MFE = 1000 (b). ....   | 208 |
| <b>Figure 9.11</b>   Empirical CDFs of best NSE values for MFE= 500 (a) and MFE = 1000 (b). ....   | 208 |
| <b>Figure 9.12</b>   Convergence curves for MFE = 500 (a) and MFE = 1000 (b). ....   | 209 |
| <b>Figure 9.13</b>   Empirical CDFs of best NSE values for MFE= 500 (a) and MFE = 1000 (b). ....   | 210 |
| <b>Figure 9.14</b>   Convergence curves (a) and empirical CDFs (b) for MFE = 500. ....   | 212 |
| <b>Figure A.1</b>   Scenario-based (see <b>Table 3-1</b> of the main manuscript; section 3.2—“ <i>The envelope behavior in the classical univariate AR(1) model</i> ”) comparison of synthetic (using the an AR(1) with PIII white noise) and theoretical autocorrelation function (ACF). The labels of each plot resemble the corresponding scenarios of the aforementioned table (see also <b>Table A-1</b> ). ....                          | 253 |
| <b>Figure A.2</b>   Monthly-based comparison of empirical (historical), synthetic (using AR(1) with PIII white noise), and theoretical autocorrelation functions (ACFs) of the real-world case study employed in section 3.3—“ <i>Real-world case study</i> ” of the main text. ....   | 255 |

**Figure B.1** | The diagonal panels (a, f, k, p) depict, for a randomly selected realization, the dependence pattern of the synthetically generated data of each process (i.e., for each site) for time lag  $\tau = 1$ . The lower triangular panels (e, i, j, m, n, o) illustrate the dependence pattern of the synthetically generated data among the 4 processes (i.e., for each pair of sites A-D) for time lag  $\tau = 0$ . The upper triangular panels (b, c, d, g, h, l) present the established relationships between equivalent,  $\tilde{\rho}^{ij}$  and target  $\rho^{ij}$  correlation coefficients given the corresponding distributions of processes  $\underline{x}_t^i$  and  $\underline{x}_t^j$  (i.e., for each pair of sites A-D)..... 257

**Figure B.2** | The diagonal panels (a, f, k, p) depict the dependence pattern of the observed and synthetically generated data of each process (i.e., for each station ID<sub>1-4</sub>) for time lag  $\tau = 1$ . Furthermore, they depict the lag-1, target ( $\rho_1^{CAS;i}$ ), simulated ( $\hat{\rho}_1^i$ ), and equivalent ( $\tilde{\rho}_1^i$ ) autocorrelation coefficients. The lower triangular panels (e, i, j, m, n, o) illustrate the dependence pattern of the observed and synthetically generated data among the processes (i.e., for each pair of stations ID<sub>1-4</sub>) for time lag  $\tau = 0$ . Furthermore, they depict the lag-0, target ( $\rho_0^{ij}$ ), simulated ( $\hat{\rho}_0^{ij}$ ), and equivalent ( $\tilde{\rho}_0^{ij}$ ) cross-correlation coefficients. The upper triangular panels (b, c, d, g, h, l) present the established relationships between equivalent,  $\tilde{\rho}^{ij}$  and target  $\rho^{ij}$  correlation coefficients given the corresponding marginal distributions of processes  $\underline{x}_t^i$  and  $\underline{x}_t^j$  (i.e., for each pair of stations ID<sub>1-4</sub>)..... 258

**Figure B.3** | Monthly-based comparison of empirical, simulated and theoretical distribution function of positive daily rainfall at Pavlos station (using the Weibull's plotting position). The title of each plot contains the parameters of the GG distribution, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry..... 259

**Figure B.4** | Monthly-based comparison of empirical, simulated and theoretical ACF of daily rainfall at Pavlos station. The title of each plot contains the parameters of the fitted auto-dependence structure (i.e., CAS)..... 260

**Figure B.5** | Monthly-based illustration of the relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients for the mixed and GG distribution that regard daily rainfall simulation at Pavlos station..... 261

**Figure B.6** | Monthly-based comparison of empirical and simulated dependence pattern for time lag 1. The title of each plot depicts the lag-1, target ( $\rho_1^{CAS}$ ), simulated ( $\rho_1$ ), and equivalent ( $\tilde{\rho}_1$ ) autocorrelation coefficients that regard daily rainfall simulation at Pavlos station..... 262

**Figure C.1** | Comparison of monthly mean values,  $\mu$ , of historical and synthetic data (simulation length: 500 000 years)..... 264

**Figure C.2** | Comparison of monthly standard deviation values,  $\sigma$ , of historical and synthetic data (simulation length: 500 000 years)..... 265

**Figure C.3** | Comparison of monthly skewness values,  $C_s$ , of historical and synthetic data (simulation length: 500 000 years)..... 265

**Figure C.4** | Comparison of month-to-month lag-1 correlations,  $\rho_1$ , of historical and synthetic data (simulation length: 500 000 years)..... 266

**Figure C.5** | Comparison of monthly lag-0 cross-correlations,  $\rho_0$ , between sites of historical and synthetic data (simulation length: 500 000 years)..... 266

**Figure D.1** | Rainfall - Monthly-based summary of L-scale ( $L_2$ ) as a function of aggregation scale  $k$ . ..... 269

**Figure D.2** | Runoff - Monthly-based summary of L-scale ( $L_2$ ) as a function of aggregation scale  $k$ . 269

**Figure D.3** | Rainfall - Monthly-based summary of L-skewness ( $L_{Cs}$ ) as a function of aggregation scale  $k$ . ..... 270

|  |     |
|--|-----|
| <b>Figure D.4</b>   Runoff - Monthly-based summary of L-skewness ( $L_{Cs}$ ) as a function of aggregation scale $k$ .   | 270 |
| <b>Figure D.5</b>   Rainfall - Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale $k$ .   | 271 |
| <b>Figure D.6</b>   Runoff - Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale $k$ .   | 271 |
| <b>Figure D.7</b>   (a-d) Historical annual time series for sites A-D. (e-h) Empirical, simulated and theoretical distribution functions for sites A-D (using the Weibull's plotting position) (i-l) Empirical, simulated and theoretical ACFs for sites A-D. (m-p) Synthetic annual time series (randomly selected window of 1 000 years).                          | 273 |
| <b>Figure D.8</b>   Comparison of historical and simulated lag-0 cross-correlations at the annual time scale.  | 273 |
| <b>Figure D.9</b>   Monthly-based comparison of monthly empirical and simulated L-Mean, L-Scale and L-Skewness, as well as probability dry.  | 274 |
| <b>Figure D.10</b>   Comparison of historical and simulated lag-1 month-to-month correlations for sites A-D.   | 274 |
| <b>Figure D.11</b>   Comparison of monthly historical and simulated lag-0 cross-correlations for sites A-D.  | 275 |
| <b>Figure D.12</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site A (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. | 275 |
| <b>Figure D.13</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site B (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. | 276 |
| <b>Figure D.14</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site C (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. | 276 |
| <b>Figure D.15</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site D (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. | 277 |
| <b>Figure D.16</b>   Monthly-based comparison of daily empirical and simulated L-Mean, L-Scale and L-Skewness, as well as probability dry.   | 278 |
| <b>Figure D.17</b>   Comparison of daily historical and simulated lag-0 cross-correlations for sites A-D.  | 278 |
| <b>Figure D.18</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site A (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.   | 279 |
| <b>Figure D.19</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site B (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.   | 279 |



|  |     |
|--|-----|
| <b>Figure D.20</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site C (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry..... | 280 |
| <b>Figure D.21</b>   Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site D (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry..... | 280 |
| <b>Figure D.22</b>   Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site A; the parameters of CAS are given on the title of each subplot. ....  | 281 |
| <b>Figure D.23</b>   Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site B; the parameters of CAS are given on the title of each subplot. ....  | 281 |
| <b>Figure D.24</b>   Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site C; the parameters of CAS are given on the title of each subplot. ....  | 282 |
| <b>Figure D.25</b>   Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site D; the parameters of CAS are given on the title of each subplot. ....  | 282 |
| <b>Figure D.26</b>   Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale $k$ for site A. ....  | 283 |
| <b>Figure D.27</b>   Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale $k$ for site B. ....  | 283 |
| <b>Figure D.28</b>   Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale $k$ for site C. ....  | 284 |
| <b>Figure D.29</b>   Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale $k$ for site D. ....  | 284 |
| <b>Figure D.30</b>   Monthly-based summary of L-Skewness ( $L_{Cs}$ ) as a function of aggregation scale $k$ for site A. ....  | 285 |
| <b>Figure D.31</b>   Monthly-based summary of L-Skewness ( $L_{Cs}$ ) as a function of aggregation scale $k$ for site B. ....  | 285 |
| <b>Figure D.32</b>   Monthly-based summary of L-Skewness ( $L_{Cs}$ ) as a function of aggregation scale $k$ for site C. ....  | 286 |
| <b>Figure D.33</b>   Monthly-based summary of L-Skewness ( $L_{Cs}$ ) as a function of aggregation scale $k$ for site D. ....  | 286 |
| <b>Figure D.34</b>   Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale $k$ for site A. ....  | 287 |
| <b>Figure D.35</b>   Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale $k$ for site B. ....  | 287 |
| <b>Figure D.36</b>   Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale $k$ for site C. ....  | 288 |
| <b>Figure D.37</b>   Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale $k$ for site D. ....  | 288 |

**Figure D.38** | Empirical (●) and simulated (●) daily annual rainfall maxima of sites A-D, as a function of the return period. The solid red line (—) depicts the fitted to historical data Generalized Extreme Value (GEV) distribution (parameters: location ( $c$ ), scale ( $b$ ) and shape ( $a$ )). The dashed blue line (---) represents the 95% confidence intervals (estimated using the parametric bootstrap method)..... 289

**Figure D.39** | Disaggregated hourly rainfall (non-zero) - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry ..... 290

**Figure D.40** | Disaggregated hourly rainfall - monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF); the parameters of CAS are given on the title of each subplot. .... 290

**Figure D.41** | Disaggregated hourly rainfall - Monthly-based summary of L-mean ( $L_1$ ) as a function of aggregation scale  $k$ ..... 291

**Figure D.42** | Disaggregated hourly rainfall - Monthly-based summary of L-scale ( $L_2$ ) as a function of aggregation scale  $k$ ..... 291

**Figure D.43** | Disaggregated hourly rainfall - Monthly-based summary of L-skewness ( $L_{Cs}$ ) as a function of aggregation scale  $k$ ..... 292

**Figure D.44** | Disaggregated hourly rainfall - Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$ ..... 292

**Figure D.45** | Disaggregated hourly rainfall - Monthly-based summary of lag-1 autocorrelation coefficient ( $\rho_1$ ) as a function of aggregation scale  $k$ . .... 293

# LIST OF TABLES

|   |     |
|---|-----|
| <b>Table 3-1</b>   Summary of target statistics for all scenarios (in all cases, $\mu_{\underline{x}} = 0.5$ and $\sigma_{\underline{x}}^2 = 1$ ).....  | 27  |
| <b>Table 4-1</b>   Comparison between numerical integration and the algorithm of section 4.5.1 for the numerical example illustrated in <b>Figure 4.2</b> . Panels a) and b) correspond to those of <b>Figure 4.2</b> .....   | 74  |
| <b>Table 5-1</b>   Summary of theoretical and simulated statistics as reproduced by SMA and SMARTA models.....  | 96  |
| <b>Table 5-2</b>   a) Synopsis of theoretical distribution models and their moments, as well as, of CAS parameters for each variable of the multivariate simulation study. b) The upper triangle (grey cells) contains the target lag-0 cross-correlation coefficients ( $\rho_0^{ij}$ ) between sites A-D, while the lower triangle depicts the corresponding estimated equivalent correlation coefficients ( $\rho_0^{ij}$ )..... | 100 |
| <b>Table 6-1</b>   Theoretical distributions and associated parameters of hypothetical process across seasons, as well as MLE estimation of simulated data.....   | 123 |
| <b>Table 6-2</b>   Simulated and theoretical values of key statistical characteristics of hypothetical process.....   | 124 |
| <b>Table 7-1</b>   Summary of employed Nataf-based models ( $p$ and $q$ , denote the order of the model).....   | 140 |
| <b>Table 8-1</b>   Present and future irrigation demand below Toxotes reservoir.....  | 173 |
| <b>Table 8-2</b>   Summary of optimization runs and configurations.....   | 176 |
| <b>Table 8-3</b>   Results summary for 200 and 400 function evaluations.....  | 178 |
| <b>Table 8-4</b>   Comparison of MOSBO and MOEA under equal budget.....   | 178 |
| <b>Table 8-5</b>   Comparison of MOEAs for 1 000 function evaluations.....  | 181 |
| <b>Table 8-6</b>   Results summary for 1 000 function evaluations.....  | 181 |
| <b>Table 8-7</b>   Comparison of MOSBO and best MOEA under different budgets.....   | 182 |
| <b>Table 8-8</b>   Results summary for 2 000 and 5 000 function evaluations.....  | 184 |
| <b>Table 8-9</b>   Operation rules Ulp and Lrp.....   | 184 |
| <b>Table 9-1</b>   Configuration of benchmarking suite.....   | 197 |
| <b>Table 9-2</b>   Summary characteristics of test functions (see also the Appendix of <i>Tsoukalas et al. [2016]</i> ).<br>.....   | 199 |
| <b>Table 9-3</b>   Mean and standard deviation of best solutions in 15-D test problems (optimal results are highlighted).....   | 200 |
| <b>Table 9-4</b>   Mean and standard deviation of best solutions in 30-D test problems (optimal results are highlighted).....   | 200 |
| <b>Table 9-5</b>   Mean and standard deviation of best solutions in 15-D test problems for MFE = 500, for different values of the four step parameters of SEEAS (for $p_m = 0.10$ and $\xi = 2$ ).....  | 205 |
| <b>Table 9-6</b>   Mean and standard deviation of best solutions in 15-D test problems for MFE = 500, for different values of mutation probability $p_m$ (for $N_r = N_e = N_c = N_u = 20$ and $\xi = 2$ ).<br>.....  | 205 |
| <b>Table 9-7</b>   Mean and standard deviation of best solutions in 15-D test problems for MFE = 500, for different values of mutation probability $p_m$ and cooling parameter $\xi$ (for $N_r = N_e = N_c = N_u = 20$ and $p_m = 0.10$ ).<br>.....   | 205 |
| <b>Table 9-8</b>   Median of best function values obtained from all algorithms.....   | 206 |

|   |     |
|---|-----|
| <b>Table 9-9</b>   Summary results of MWU test to infer about the preferred algorithm. H-value indicates the rejection or not of the null hypothesis, i.e., if $H = 0$ , the null hypothesis is not rejected.....   | 206 |
| <b>Table 9-10</b>   Model parameters, feasible bounds and values assigned for toy calibrations. ....  | 207 |
| <b>Table 9-11</b>   Statistical characteristics of NSE values obtained from all algorithms. ....  | 208 |
| <b>Table 9-12</b>   Statistical characteristics of NSE values obtained from all algorithms. ....  | 209 |
| <b>Table A-1</b>   Scenario-based summary of theoretical (see <b>Table 3-1</b> of the main manuscript; section <b>3.2</b> —“ <i>The envelope behavior in the classical univariate AR(1) model</i> ”) and simulated (synthetically generated; using an AR(1) with PIII white noise) statistics. .... | 252 |
| <b>Table A-2</b>   Summary of theoretical and simulated statistics for the first, zero-autocorrelated, bivariate AR(1) process with PIII white noise, employed in section <b>3.2.3</b> —“ <i>From the univariate to the multivariate AR(1) model</i> ” of the main text. ....                       | 253 |
| <b>Table A-3</b>   Summary of theoretical and simulated statistics for the second, autocorrelated, bivariate AR(1) process with PIII white noise, employed in section <b>3.2.3</b> —“ <i>From the univariate to the multivariate AR(1) model</i> ” of the main text.....                            | 254 |
| <b>Table A-4</b>   Monthly-based summary of historical and simulated (synthetically generated using an AR(1) with PIII white noise) statistics of the real-world case study employed in section <b>3.3</b> —“ <i>Real world case study</i> ” of the main text. ....                                 | 254 |
| <b>Table C-1</b>   Parameters of PIII for historical and simulated data (from PAR-PIII and SPARTA-PIII); identified with the method of moments.....   | 267 |
| <b>Table C-2</b>   Root mean square error between the theoretical values, (i.e., the historical) and the distribution parameters of simulated data of PAR-PIII and SPARTA-PIII models (see <b>Table C-1</b> )..   | 268 |

This page is intentionally left blank.

## ABSTRACT

---

Hydrometeorological inputs are a key ingredient and simultaneously one of the main sources of uncertainty of every hydrological study. This type of uncertainty is referred to as hydrometeorological uncertainty and is of utmost importance in risk-based engineering works, due the high variability and randomness that is naturally embedded in physical processes. Considering hydrometeorological time series as realizations of stochastic processes allow their analysis, modeling, simulation and forecasting. Embracing the existence of randomness and unpredictability in such processes is a first step towards their understanding and the development of uncertainty-aware methodologies for water-systems optimization.

In this vein, due to the typical size of historical data, which is not (neither will ever be) sufficient to extract safe conclusions about the long-term performance of a system, the common procedure entails driving the typically deterministic water-system models (conceptual or physical-based) using stochastic inputs (that in a statistical sense resemble the parent information; typically, but not exclusively derived from the historical time series). This essentially enables the establishment of Monte Carlo experiments where the intrinsic uncertainty of the inputs (i.e., hydrometeorological processes) is propagated through a deterministic filter (i.e., a water-system simulation model) in order to derive, or assess, the probabilistic behavior of the output of interest (e.g., water supply coverage). Further to this, when the objective is the optimization of the deterministic model's control variables (i.e., model's parameters) with respect to some quantity or metric (i.e., objective), this procedure can (and should) be embed within an iterative scheme driven by an optimization algorithm (i.e., establishing uncertainty-aware simulation-optimization frameworks).

An important step of this procedure is the realistic simulation of hydrometeorological processes, since they are the main drivers of the whole procedure, and eventually determine its accuracy, as well as the probabilistic behavior of the output of interest. This in turn, poses an intriguing challenge that arises from a series of unique peculiarities that characterize such processes, namely, non-Gaussianity, intermittency, auto-dependence (short- or long-range), cross-dependence and periodicity. Despite the significant amount of research during last decades, these challenges remain partially unresolved. To a large extent, this is due to the standard hypothesis of most simulation schemes that does not lie in the reproduction of a specific distribution, but on the reproduction of low-order statistics (e.g., mean, variance, skewness) and correlations in time and space. This is a problem because, a) for a given set of low-order statistics multiple distributions may be represented, thus making the simulation problem only partially defined, and b) as shown herein, this practice may lead to bounded, and thus unrealistic dependence forms among consecutive time steps and/or processes.

Further to this, driving water-system simulation models with long stochastically generated sequences, thus accounting for input (hydrometeorological) uncertainty, inevitably increases the required computational effort, especially within the context of simulation-optimization frameworks. This in turn, poses the challenge of addressing and ensuring the practical implementation of water-system optimization problems under uncertainty.

Thereby, the main research objectives and contributions of this Thesis are related to:

a) The development of novel non-Gaussian stochastic simulation models, able to account also for the other peculiarities typically encountered in hydrometeorological processes, such as,

intermittency, auto- and cross- dependence, periodicity, as well as their scale-varying probabilistic and stochastic behavior.

b) The development of surrogate-based optimization methodologies and algorithms that can efficiently and effectively confront water-system simulation-optimization problems under uncertainty, i.e., when using stochastic inputs to drive the simulation-optimization procedure.

Specifically, herein a by building upon copula concepts, probability laws and the theory of stochastic processes, a theoretically justified *family* of univariate and multivariate non-Gaussian stationary and cyclostationary models is defined and thoroughly investigated. This type of models have been unknown to the hydrological community, and this Thesis is the first attempt to align them with hydrological stochastics. The developed models are shown to be able to account for all the typical characteristics of hydrometeorological processes and simultaneously exhibit a simple and parsimonious character. Furthermore, these models are then coupled, using a disaggregation approach, thus eventually enabling the development of a modular stochastic simulation framework that allows the simultaneous reproduction of the probabilistic and stochastic behavior (including non-Gaussian distributions) of hydrometeorological processes at multiple time scales (from annual to daily; as well as finer time scales). The advantages of this class of stochastic processes and models, as well as of the modular stochastic simulation framework for multi-scale simulations, are demonstrated and verified through numerous hypothetical and real-world simulation studies.

Finally, in order to ensure the effective exploitation and practical implementation of these new developments in the stochastic simulation of hydrometeorological processes within the uncertainty-aware, engineering design and management of water-systems (i.e., driven by stochastic inputs), this Thesis develops appropriate surrogate-based computationally-efficient methodologies and algorithms, that effectively handle water-system simulation-optimization problems under hydrometeorological uncertainty, thus alleviating the associated computational barrier.

# ΠΕΡΙΛΗΨΗ

---

## ΘΕΤΟΝΤΑΣ ΤΟ ΠΛΑΙΣΙΟ

Τα υδρομετεωρολογικά δεδομένα αποτελούν βασικό συστατικό και ταυτόχρονα μια από τις κύριες πηγές αβεβαιότητας κάθε υδρολογικής μελέτης. Αυτού του τύπου η αβεβαιότητα είναι γνωστή ως *υδρομετεωρολογική*. Λόγω της μεγάλης μεταβλητότητας και τυχαίας φύσης αυτών των διεργασιών, η αντιμετώπιση της αποτελεί ύψιστης σημασίας προτεραιότητα σε έργα και μελέτες μηχανικού οι οποίες λαμβάνουν υπόψιν τους τις έννοιες του ρίσκου και της διακινδύνευσης. Η παραδοχή πως οι παρατηρημένες υδρομετεωρολογικές χρονοσειρές αποτελούν πραγματοποιήσεις στοχαστικών ανελίξεων (ή αλλιώς διεργασιών) επιτρέπει την ανάλυση, μοντελοποίηση, προσομοίωση και πρόβλεψη τους ως τέτοιες. Η αναγνώριση και η παραδοχή ύπαρξης τυχαιότητας και μη προβλεψιμότητας σε αυτού του τύπου διεργασίες αποτελεί το πρώτο βήμα προς την κατανόησή τους, τη μελέτη και την ανάπτυξη μεθοδολογιών για τη βελτιστοποίηση υδατικών συστημάτων υπό αβεβαιότητα.

Συνήθως, το περιορισμένο μέγεθος των ιστορικών δεδομένων (χρονοσειρών), δεν επιτρέπει (ούτε πρόκειται ποτέ) την εξαγωγή ασφαλών συμπερασμάτων για την μακροπρόθεσμη επίδοση ενός συστήματος. Για αυτό το λόγο, η συνήθης πρακτική κάνει χρήση στοχαστικών δεδομένων εισόδου (τα οποία είναι στατιστικά συνεπή με τις ιστορικές χρονοσειρές ή γενικότερα με την όποια διαθέσιμη υδρολογική πληροφορία) σε συνδυασμό με ντετερμινιστικά μοντέλα υδατικών συστημάτων (φυσικής ή εννοιολογικής βάσης). Αυτός ο συνδυασμός ουσιαστικά επιτρέπει την ανάπτυξη *πειραμάτων* τύπου Monte Carlo, όπου η αβεβαιότητα των δεδομένων εισόδου (π.χ., υδρομετεωρολογικές μεταβλητές) μεταφέρεται μέσω ενός ντετερμινιστικού φίλτρου (π.χ., μοντέλα προσομοίωσης υδατικών συστημάτων) στις μεταβλητές εξόδου (π.χ., αξιοπιστία κάλυψης υδατικών αναγκών) για τη εξαγωγή και διερεύνηση της πιθανοτικής συμπεριφοράς των τελευταίων. Επιπλέον, όταν ο στόχος της μελέτης είναι η βελτιστοποίηση των μεταβλητών ελέγχου του ντετερμινιστικού μοντέλου, με γνώμονα κάποια αντικειμενική συνάρτηση, η παραπάνω διαδικασία μπορεί (και πρέπει) να μετατραπεί σε επαναληπτική, μέσω της χρήσης κατάλληλων αλγόριθμων βελτιστοποίησης (δηλ. ανάπτυξη πλαισίων προσομοίωσης-βελτιστοποίησης που λαμβάνουν υπόψιν τους την αβεβαιότητα).

Ένα σημαντικό σημείο της παραπάνω διαδικασίας είναι η ρεαλιστική προσομοίωση των υδρομετεωρολογικών διεργασιών, αφού αποτελούν βασικό *οδηγό* της όλης διαδικασίας, καθώς ταυτόχρονα καθορίζουν την ακρίβεια προσομοίωσης αλλά και την πιθανοτική συμπεριφορά των μεταβλητών εξόδου. Αυτό με τη σειρά του θέτει μια ενδιαφέρουσα πρόκληση η οποία πηγάζει από τα ιδιαίτερα χαρακτηριστικά που παρουσιάζουν αυτού του είδους διεργασίες, όπως οι μη-Γκαουσιανές κατανομές, η διαλείπουσα συμπεριφορά, η χρονική εξάρτηση (μικρής ή μακράς εμβέλειας), η χωρική αλληλεξάρτηση καθώς και η περιοδικότητα. Παρά τη σημαντική έρευνα που έχει πραγματοποιηθεί τις τελευταίες δεκαετίες, το πρόβλημα της ρεαλιστικής προσομοίωσης υδρομετεωρολογικών διεργασιών παραμένει ακόμη θέμα συζήτησης. Σε ένα μεγάλο βαθμό, αυτό οφείλεται στην συνήθη υπόθεση των περισσότερων σχημάτων προσομοίωσης, τα οποία δεν στοχεύουν στην αναπαραγωγή κάποιας πιθανοτικής κατανομής, αλλά στην αναπαραγωγή χαμηλής τάξης στατιστικών χαρακτηριστικών (π.χ., μέση τιμή, τυπική απόκλιση και συντελεστή ασυμμετρίας) και συσχετίσεων στο χρόνο και το χώρο. Κάτι



τέτοιο αποτελεί πρόβλημα γιατί, α) για δεδομένα στατιστικά χαρακτηριστικά χαμηλής τάξης πολλές κατανομές μπορεί να είναι συνεπής, κάτι που καθιστά το πρόβλημα μερικώς ορισμένο, και β) όπως αναδεικνύεται στην παρούσα διατριβή, αυτή η πρακτική μπορεί να οδηγήσει σε φραγμένες, και άρα μη ρεαλιστικές μορφές εξάρτησης μεταξύ διαδοχικών χρονικών βημάτων και/ή διεργασιών.

Πέραν των παραπάνω, η χρήση συνθετικών χρονοσειρών μεγάλου μήκους σε συνδυασμό με μοντέλα προσομοίωσης υδατικών συστημάτων, ναι μεν παρέχει τη δυνατότητα ενσωμάτωσης της (υδρομετεωρολογικής) αβεβαιότητας, αλλά από την άλλη αυξάνει τον απαιτούμενο υπολογιστικό χρόνο, ειδικά στο πλαίσιο σχημάτων προσομοίωσης-βελτιστοποίησης. Αυτό με τη σειρά του, θέτει την πρόκληση της πρακτικής εφαρμογής τέτοιων σχημάτων για τη βελτιστοποίηση υδατικών συστημάτων υπό αβεβαιότητα.

Κύριοι ερευνητικοί στόχοι και συνεισφορά της παρούσας διδακτορικής διατριβής είναι:

α) Η ανάπτυξη μη-Γκαουσιανών στοχαστικών μοντέλων προσομοίωσης, τα οποία είναι επίσης ικανά να προσομοιώσουν τα ιδιαίτερα χαρακτηριστικά που παρουσιάζουν οι υδρομετεωρολογικές διεργασίες, δηλαδή, την διαλείπουσα συμπεριφορά, την χρονική και χωρική εξάρτηση, την περιοδικότητα καθώς και την πιθανοτική και στοχαστική συμπεριφορά τους σε πολλαπλές χρονικές κλίμακες.

β) Η χρήση υποκατάστατων μοντέλων (surrogate models) για την ανάπτυξη μεθοδολογιών και αλγορίθμων που είναι σε θέση να αντιμετωπίσουν αποτελεσματικά και αποδοτικά προβλήματα βελτιστοποίησης υδατικών συστημάτων υπό αβεβαιότητα (μέσω του συνδυασμού στοχαστικών δεδομένων εισόδου και σχημάτων προσομοίωσης-βελτιστοποίησης).

## **ΣΤΟΧΑΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΚΑΙ ΠΡΟΣΟΜΟΙΩΣΗ ΥΔΡΟΜΕΤΕΩΡΟΛΟΓΙΚΩΝ ΔΙΕΡΓΑΣΙΩΝ**

Η ιδέα της χρήσης συνθετικών χρονοσειρών στις υδρολογικές μελέτες χρονολογείται, περισσότερο από 100 χρόνια πριν, από τον *Hazen* [1914] ο οποίος συνδύασε πολλές ιστορικές παρατηρήσεις σε μια χρονοσειρά με σκοπό να δημιουργήσει μια συνθετική πραγματοποίηση της ετήσιας απορροής. Αυτή η απλή προσέγγιση υπήρξε η πρώτη από τις πολλές που αναπτυχθήκαν εν συνεχεία (δες ανασκόπηση κεφαλαίου 2.3) και συνέβαλε καθοριστικά στην γένεση του επιστημονικού τομέα της συνθετικής (ή επιχειρησιακής) υδρολογίας.

Σύμφωνα με την κλασσική κατηγοριοποίηση του *Matalas* [1975], η συνθετική υδρολογία αποτελεί παρακλάδι της στοχαστικής υδρολογίας και η εμφάνισή της αποδίδεται εν πολλοίς στις καθοριστικές εργασίες που πραγματοποιήθηκαν στα πλαίσια του προγράμματος για το νερό του Harvard [*Maass et al.*, 1962] και στις σχετικές εργασίες των *Thomas and Fiering* [1962] που πιθανώς ήταν οι πρώτοι που εφάρμοσαν την θεωρία στοχαστικών ανεξίτητων για τη σύνθεση μηνιαίων χρονοσειρών απορροής. Σύμφωνα με τον *Koutsoyiannis* [2000] σχετικά καθοριστικά επιτεύγματα που συνέβαλαν στην καθιέρωση του τομέα αυτού (και της στοχαστικής υδρολογίας γενικότερα) ήταν η αξιοσημείωτη πρόοδος στους υπολογιστές κατά τη δεκαετία του 1950 καθώς και η εκτεταμένη υιοθέτηση των μεθόδων Monte Carlo σε διάφορα επιστημονικά πεδία (π.χ., φυσική, βιολογία και οικονομικά). Μια άλλη αξιοσημείωτη συνεισφορά, που προέρχεται από ένα διαφορετικό επιστημονικό τομέα, ήταν η έκδοση του πλέον κλασσικού βιβλίου ανάλυσης χρονοσειρών από τους *Box and Jenkins* [1970] που προσφέρει μια ολοκληρωμένη αντιμετώπιση του θέματος και παρέχει μια λεπτομερή κατηγοριοποίηση των γραμμικών στοχαστικών μοντέλων περιλαμβανομένων των μοντέλων

αυτό-παλινδρόμησης (AR), κινούμενου μέσου όρου (MA) καθώς και των συνδυασμό τους, μοντέλα αυτό-παλινδρόμησης κινούμενου μέσου όρου (ARMA).

Οι αρχικές εργασίες που έκαναν χρήση της έννοιας των συνθετικών χρονοσειρών στόχευαν στην αξιολόγηση της επίδοσης των συστημάτων ταμειωτήρων με χρήση πιθανοτικών όρων, δηλ., με εκτίμηση της αξιοπιστίας επί της βάσης των προσομοιωμένων χρονοσειρών [e.g., Hazen, 1914; Sudler, 1927; Barnes, 1954; Thomas and Fiering, 1962; Klemeš, 1981]. Σήμερα, συνθετικά δεδομένα χρησιμοποιούνται σε μια μεγάλη ποικιλία μελετών (με δομή όμοια με αυτή των πειραμάτων Monte Carlo που αναφέρθηκαν παραπάνω), όπως ο βέλτιστος σχεδιασμός και λειτουργία των συστημάτων ταμειωτήρων [e.g., Koutsoyiannis and Economou, 2003; Celeste and Billib, 2009; Giuliani et al., 2014; Tsoukalas and Makropoulos, 2015a, 2015b; Feng et al., 2017], η ανάλυση ρίσκου πλημμύρας [e.g., Wheeler et al., 2005; Haberlandt et al., 2011; Paschalis et al., 2014; Qin and Lu, 2014; Moustakis et al., 2017] και γεγονότων ξηρασίας [e.g., Herman et al., 2016], καθώς και η προσομοίωση υδατικών πόρων επί της βάσης μελλοντικών κλιματικών συνθηκών [e.g., Fowler et al., 2000; Baltas, 2007; Kilsby et al., 2007; Baltas and Karaliolidou, 2008; Fatichi et al., 2011; Nazemi et al., 2013].

Η βασική απαίτηση για την εξαγωγή στατιστικά συνεπών αποτελεσμάτων από τα πειράματα Monte Carlo (δεδομένου ότι το μοντέλο προσομοίωσης παρέχει μια πιστή αναπαράσταση του συστήματος που μελετάται) είναι η πιστή αναπαράσταση και προσομοίωση των υδρομετεωρολογικών διεργασιών που με τη σειρά της επιτάσσει τη χρήση στοχαστικών μοντέλων προσομοίωσης ικανών να λάβουν υπόψη τους τις ιδιαιτερότητες των υδρομετεωρολογικών διεργασιών. Οι πιο σημαντικές από αυτές είναι, η απόκλιση από την κανονική κατανομή, η διαλείπουσα φύση, η αυτό-συσχέτιση (βραχυπρόθεσμη ή μακροπρόθεσμη), η ετερο-συσχέτιση και η περιοδικότητα (δες κεφάλαιο 2.2).

Σε ένα πιο αφηρημένο επίπεδο, τα δυο πρώτα χαρακτηριστικά (η απόκλιση από την κανονική κατανομή και η διαλείπουσα φύση) σχετίζονται με τις ιδιότητες της περιθώριας κατανομής της διαδικασίας και υπαγορεύουν την ανάγκη για ένα κατάλληλο πιθανοτικό μοντέλο. Από την άλλη πλευρά, οι αυτό- και έτερο-συσχετίσεις σχετίζονται με τις στοχαστικές (από κοινού) ιδιότητες της διαδικασίας, τόσο στο χρόνο όσο και στο χώρο, και υπαγορεύουν την ανάγκη για χρήση ενός στοχαστικού μοντέλου προσομοίωσης. Στην πραγματικότητα, στην περίπτωση που οι φυσικές διεργασίες δεν είναι αυτό- ή έτερο-συσχετισμένες, το πρόβλημα της προσομοίωσης θα ήταν αρκετά απλούστερο, καθώς η γέννηση συνθετικών χρονοσειρών θα στηριζόταν στη γέννηση αριθμών από ομοιόμορφη κατανομή και στη χρήση της αντίστροφης κατανομής (δηλ., probability integral transformation). Τέλος, η περιοδικότητα εισάγει επιπλέον πολυπλοκότητα, δεδομένου ότι υπαγορεύει την αναπαράσταση της διαδικασίας ως κυκλο-στάσιμη, με διαφορετικές περιθώριες και από κοινού ιδιότητες όχι μόνο σε διαφορετικές χρονικές κλίμακες αλλά και σε διαφορετικές περιόδους (ή γενικά σε συστηματικούς επαναλαμβανόμενα χρονικά διαστήματα).

Αναμφισβήτητα, ένα κατάλληλο σχήμα στοχαστικής προσομοίωσης θα πρέπει να είναι ικανό να αναπαράγει την πιθανοτική και στοχαστική συμπεριφορά (δηλ., τις περιθώριες και από κοινού ιδιότητες) της υδρομετεωρολογικής διεργασίας, που διαφοροποιείται ανάλογα με τον τύπο της μεταβλητής (π.χ., βροχή, απορροή ή θερμοκρασία) αλλά και από την υπό μελέτη χρονική κλίμακα (π.χ., ετήσια, μηνιαία, ημερήσια ή λεπτότερη).

Η ανάγκη για γενικευμένα σχήματα προσομοίωσης που επιτρέπουν την παραγωγή συνθετικών χρονοσειρών για πολλαπλές κατανομές πηγάζει πρωτίστως από το γεγονός ότι, η πιθανοτική συμπεριφορά πολλών υδρομετεωρολογικών διεργασιών δεν αναπαράγεται ικανοποιητικά από τα κλασσικά στοχαστικά μοντέλα (δες επισκόπηση στο κεφάλαιο 2.3). Πολλά από τα μοντέλα

αυτά (δηλ., τα κλασσικά γραμμικά στοχαστικά μοντέλα, τα μοντέλα σημειακής προσομοίωσης και τα μοντέλα επαναδειγματοληψίας) δεν είναι σχεδιασμένα για να αναπαράγουν σημαντικές πιθανοτικές πτυχές της διεργασίας (π.χ. μέγιστα και ελάχιστα - που σχετίζονται με την ουρά της κατανομής), δεδομένου ότι εκ φύσεως δεν αναπαράγουν κάποια συγκεκριμένη κατανομή αλλά συγκεκριμένα χαμηλών τάξεων στατιστικά χαρακτηριστικά (π.χ., μέση τιμή, διασπορά, ασυμμετρία) και συσχετίσεις στο χρόνο και στο χώρο. Πέραν αυτού, όπως έδειξαν οι *Tsoukalas et al.* [2018a], και συζητείται στο Κεφάλαιο 3, οι τυπικές στρατηγικές μοντελοποίησης μπορούν να οδηγήσουν σε φραγμένες, και κατά συνέπεια, μη ρεαλιστικές και μη φυσικές δομές συσχέτισης, παρά το γεγονός ότι τα σημαντικά στατιστικά χαρακτηριστικά των ιστορικών δεδομένων αναπαράγονται ικανοποιητικά.

Επιπλέον, η αναπαραγωγή της συνάρτησης κατανομής της διαδικασίας κρίνεται ως υψίστης σημασίας, όπως προτάσσεται από θεωρητικά και εμπειρικά στοιχεία. Αυτό τονίζεται emphatically από τους *Klemeš and Borůvka* [1974] οι οποίοι αναφέρουν:

*Simulation of a serially correlated series with a given marginal distribution is one of the important prerequisites of synthetic hydrology and of its applications to analysis of water resource systems.*

Αξίζει να αναφερθεί ότι η βιβλιογραφία προσφέρει εναλλακτικές μεθόδους για την αναπαραγωγή συνθετικών χρονοσειρών, όπως είναι τα επονομαζόμενα μοντέλα δυο-καταστάσεων και τα πρόσφατα μοντέλα βασισμένα σε πολυμεταβλητές συναρτήσεις γνωστές ως copula. Αυτοί οι τύποι μοντέλων είναι ικανοί να παράγουν συνθετικές πραγματοποιήσεις με δεδομένη περιθώρια κατανομή, αλλά υποστηρίζουν ένα περιορισμένο εύρος από δομές συσχέτισεων (π.χ., τα μοντέλα δυο-καταστάσεων συνήθως αγνοούν την χρονική εξάρτηση, αυτοσυσχέτιση), ενώ χαρακτηρίζονται από δύσχρηστους μηχανισμούς γέννησης (δες κεφάλαιο 2.3 για περισσότερες λεπτομέρειες).

Πέραν των παραπάνω, ένα κοινό χαρακτηριστικό των περισσότερων υπάρχοντων μεθόδων προσομοίωσης είναι ότι στοχεύουν στην προσομοίωση της διαδικασίας σε μια μόνο χρονική κλίμακα και δεν λαμβάνουν υπόψη τους την ρητή αναπαραγωγή των ιδιοτήτων της διαδικασίας σε πολλαπλά χρονικά επίπεδα. Επισημαίνεται ότι η ταυτόχρονη προσομοίωση των υδρομετεωρολογικών διεργασιών σε πολλά χρονικά επίπεδα παραμένει ακόμη μια ανοιχτή πρόκληση. Για την λεπτομερή παρουσίαση του προβλήματος καθώς και τους πιθανούς τρόπους αντιμετώπισής του με βάση τα νέα μοντέλα προσομοίωσης που παρουσιάζονται στα κεφάλαια 4-6, δες Κεφάλαιο 7.

Αναμφισβήτητα, η κύρια δυσκολία στην προσομοίωση των υδρομετεωρολογικών διεργασιών πηγάζει από το γεγονός ότι, τα κλασσικά στοχαστικά μοντέλα (δες κεφάλαιο 2.3.1), που είναι ικανά να μοντελοποιήσουν και να προσομοιώσουν, μονομεταβλητές ή πολυμεταβλητές, στάσιμες ή κυκλο-στάσιμες, διεργασίες με μεγάλο εύρος δομών συσχέτισης, δεν είναι ικανά να αναπαράγουν την μη-Γκαουσιανή και διαλείπουσα φύση των υδρομετεωρολογικών διεργασιών δεδομένου ότι τα περισσότερα από αυτά έχουν κατασκευαστεί για την προσομοίωση διαδικασιών με κανονική (Gaussian) κατανομή. Αυτή η δυσκολία ίσως σχετίζεται με την παρακάτω πρόταση, η οποία αποτελεί την *προσευχή* του Chester Kisiel [1967] στον *θεωρητικό υδρολόγο* [Klemeš, 1997 p. 288]:

*Oh, Lord, please keep the world linear and Gaussian.*

## ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΥΔΑΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ ΥΠΟ ΑΒΕΒΑΙΟΤΗΤΑ

Η σύζευξη μεθόδων προσομοίωσης και βελτιστοποίησης αποτελεί μια ισχυρή τεχνική που έχει τραβήξει την προσοχή της επιστήμης και τεχνολογίας υδατικών πόρων, δεδομένου ότι παρουσιάζει μεγάλα πλεονέκτημα έναντι της παραδοσιακής μεμονωμένης χρήσης των δυο προσεγγίσεων [e.g., *Koutsoyiannis and Economou, 2003*]. Σε αυτό το πλαίσιο, ένα μοντέλο προσομοίωσης χρησιμοποιείται για να αναπαραστήσει τις δυναμικές που αναπτύσσονται στο υπό μελέτη σύστημα σε διαδοχικά χρονικά βήματα και εν συνεχεία να αξιολογήσει την συνολική του επίδοση επί της βάσης ενός ή πολλών κριτηρίων που ορίζονται από το χρήστη. Δεδομένου ότι αυτά τα κριτήρια εκφράζονται μέσω μιας στοχαστικής συνάρτησης, η προσομοίωση μπορεί να καθοδηγηθεί μέσω ενός αλγόριθμου βελτιστοποίησης που υλοποιεί μια συστηματική αναζήτηση στον χώρο των παραμέτρων με στόχο την μεγιστοποίηση της επίδοσης του συστήματος - σε κάθε δοκιμή, νέες τιμές αποδίδονται στις μεταβλητές ελέγχου του μοντέλου προσομοίωσης το οποίο και τρέχει αυτόματα για ανανεώσει την τιμή της στοχαστικής συνάρτησης.

Τα συνδυασμένα σχήματα προσομοίωσης-βελτιστοποίησης για υδατικά συστήματα μπορούν γενικά να κατηγοριοποιηθούν σε δυο γενικές κατηγορίες: (α) Προβλήματα λήψης αποφάσεων, στα οποία οι ιδιότητες του συστήματος και οι σχετικές διεργασίες είναι γνωστές εκ των προτέρων, ωστόσο κάποια μεγέθη σχεδιασμού και η διαχείρισή τους είναι άγνωστη, και (β) προβλήματα βαθμονόμησης (γνωστά και ως αντίστροφα προβλήματα) στα οποία κάποιες εσωτερικές ιδιότητες του συστήματος, είτε φυσικές είτε εννοιολογικές, είναι άγνωστες και πρέπει να προσδιοριστούν μέσω μιας διαδικασίας που προβλέπει την ελαχιστοποίησης της απόκλισης μεταξύ προσομοιωμένων και παρατηρημένων αποκρίσεων του συστήματος. Παρά την διαφορετική λογική τους, και οι δυο τύποι προβλημάτων χαρακτηρίζονται από συγκεκριμένες αβεβαιότητες και πολυπλοκότητες, και συχνά υπόκεινται σε πολλαπλά (και συχνά αντικρουόμενα) κριτήρια και αρκετούς περιορισμούς.

Η ανάγκη για προχωρημένα εργαλεία ολικής βελτιστοποίησης (π.χ., εξελικτικοί αλγόριθμοι) έχει αναγνωριστεί ωρίς από την υδρολογική κοινότητα η οποία έχει σημαντική εμπειρία στην χρήση τους καθώς και καθοριστική συμβολή στην ανάπτυξή τους. Στην βιβλιογραφία είναι διαθέσιμες πολλές επισκοπήσεις μεθόδων βελτιστοποίησης σε τέτοια προβλήματα. Για παράδειγμα, στα πλαίσια σχεδιασμού και διαχείρισης συστημάτων νερού, διακρίνουμε τις εργασίες των *Labadie [2004]*, *Fowler et al. [2008]*, *Nicklrow et al. [2010]*, *Reed et al. [2013]* (που εστιάζουν στην πολυκριτηριακή εφαρμογή των μεθόδων) και των *Ahmad et al. [2014]*. Η βιβλιογραφία που αφορά την υδρολογική βαθμονόμηση είναι ακόμα πιο εκτενής. Για διευκόλυνση, διακρίνουμε τις πρόσφατες εργασίες των *Duan [2013]* και *Efstratiadis and Koutsoyiannis [2010]*, που παρέχουν μια πλήρη ανασκόπηση των ολικών και πολυκριτηριακών μεθόδων, αντίστοιχα. Επίσης, αξίζει να αναφερθεί, η εργασία των *Maier et al. [2014]*, που συνοψίζει την τρέχουσα κατάσταση των εξελικτικών αλγορίθμων και άλλων μετα-ευρετικών μεθόδων, και ορίζει νέες κατευθύνσεις για μελλοντική έρευνα όσον αφορά στην εφαρμογή τους σε προβλήματα υδατικών πόρων.

Στην όλη υπολογιστική διαδικασία, η προσομοίωση είναι με διαφορά η συνιστώσα με τον μεγαλύτερο υπολογιστικό φόρτο. Καθώς τα μοντέλα γίνονται όλο και πιο πολύπλοκα και απαιτητικά όσον αφορά τα δεδομένα, η απαιτήσις τους σε υπολογιστικό φόρτο και ισχύ (δηλ., CPU) αυξάνει ραγδαία [e.g., *Tolson and Shoemaker, 2007*; *Keating et al., 2010*; *Razavi et al., 2010*; *Efstratiadis et al., 2015*; *Tsoukalas and Makropoulos, 2015b, 2015a*, *Tsoukalas et al., 2015b, 2015a, 2016*]. Ένα τυπικό παράδειγμα αποτελούν τα υδρολογικά μοντέλα φυσικής βάσης μικρής χρονικής και χωρικής κλίμακας, σε αντίθεση με τα συγκεντρωτικά εννοιολογικά μοντέλα βροχής-απορροής.

Σε άλλες εφαρμογές, που αναφέρονται ως στοχαστικά προβλήματα προσομοίωσης-βελτιστοποίησης (πειράματα Monte Carlo που εμπεριέχουν την χρήση μεθόδων βελτιστοποίησης - δες κεφάλαιο 1.1), ο υπολογιστικός φόρτος αυξάνει πολλές τάξεις μεγέθους λόγω της χρήσης συνθετικών (αντί ιστορικών) χρονοσειρών πολύ μεγάλου μήκους (π.χ., χιλιάδες χρόνια) έτσι ώστε να εκτιμηθούν τα πιθανοτικά μεγέθη (π.χ., αξιοπιστία, διακινδύνευση) με την απαραίτητη ακρίβεια. Ανάλογα με τον αριθμό των παραμέτρων και την πολυπλοκότητα της επιφάνειας απόκρισης, ο αλγόριθμος βελτιστοποίησης θα πρέπει να καλέσει το μοντέλο προσομοίωσης χιλιάδες φορές για να συγκλείνει σε μια καλή λύση. Συνεπώς, ο υπολογιστικός φόρτος της προσομοίωσης θέτει ένα πρακτικό εμπόδιο στην βελτιστοποίηση, που θα πρέπει να ολοκληρωθεί σε ένα περιορισμένο χρόνο, όπως αυτός συνήθως εκφράζεται μέσω του μέγιστου αριθμού επαναλήψεων ή υπολογισμών της στοχαστικής συνάρτησης. Για παράδειγμα, ας υποθέσουμε ένα πρόβλημα προσομοίωσης που απαιτεί περίπου 1.5 λεπτό για κάθε μια προσομοίωση και ένα αλγόριθμο βελτιστοποίησης που απαιτεί 10 000 επαναλήψεις για να προσεγγίσει το ολικό ελάχιστο. Μια τέτοια διαδικασία θα διαρκούσε περισσότερο από 10 μέρες, γεγονός που την καθιστά πρακτικά μη εφικτή.

Σύμφωνα με τους [Razavi et al. \[2010\]](#), οι προσεγγίσεις για την εξάλειψη του υπολογιστικού φόρτου, που επιβάλλεται από χρονοβόρα μοντέλα προσομοίωσης, μπορεί να κατηγοριοποιηθεί σε τέσσερις κύριες κατηγορίες: (1) παράλληλος προγραμματισμός [[e.g., Schutte et al., 2004; Cheng et al., 2005; Vrugt et al., 2006; Feyen et al., 2007; He et al., 2007; Regis and Shoemaker, 2009; Dias et al., 2013](#)], (2) υπολογιστικά αποτελεσματικούς αλγόριθμους βελτιστοποίησης [[e.g., Tolson and Shoemaker, 2007; Kuzmin et al., 2008; Tan et al., 2008; Tolson et al., 2009](#)], (3) στρατηγικές για την αποφυγή υπολογισμών με χρονοβόρα μοντέλα [[e.g., Ostfeld and Salomons, 2005; Razavi et al., 2010; Matott et al., 2012](#)], και (4) υποκατάστατα μοντέλα τα οποία αναφέρονται και ως, μετα-μοντέλα [[Blanning, 1975](#)], μοντέλα επιφάνειας απόκρισης, και μοντέλα εξομοίωσης [[Razavi et al., 2012a](#)], όπου στοχεύουν στην προσέγγιση των αποκρίσεων του πραγματικού μοντέλου προσομοίωσης. Είναι σημαντικό να επισημανθεί, ότι στα πλαίσια του συνδυασμένου σχήματος προσομοίωσης-βελτιστοποίησης, τα υποκατάστατα μοντέλα παίζουν το ρόλο προσεγγίσεων μαύρου-κουτιού που στοχεύουν στην δημιουργία μιας σχέσης εξάρτησης μεταξύ των μεταβλητών ελέγχου του μοντέλου προσομοίωσης (επεξηγηματικές μεταβλητές) και της στοχαστικής συνάρτησης του μοντέλου βελτιστοποίησης (μεταβλητή απόκρισης). Ο παράλληλος προγραμματισμός από την άλλη επιτρέπει την εκτέλεση ανεξάρτητων προσομοιώσεων από πολλαπλούς επεξεργαστές, και αναπόφευκτα απαιτεί σημαντικές επενδύσεις σε υλικοτεχνικό εξοπλισμό που τον καθιστά μη πρακτικό για κοινή χρήση. Αξίζει να σημειωθεί ότι για να μειωθεί ο υπολογιστικός χρόνος τρεις τάξεις μεγέθους – μια λογική απαίτηση για ένα πολύπλοκο πρόβλημα προσομοίωσης – θα πρέπει να χρησιμοποιηθούν 1 000 παράλληλοι επεξεργαστές, κάτι το οποίο απέχει από την πραγματικότητα. Οι δυο επόμενες επιλογές, δηλ., η βελτίωση της αποτελεσματικότητας των ήδη υπάρχοντων αλγορίθμων, καθώς και η διακοπή της διαδικασίας όταν η επίδοση του μοντέλου δείχνει να είναι φτωχή από τα πρώτα βήματα της προσομοίωσης μπορούν να εξοικονομήσουν χρόνο, αλλά όχι όσο απαιτείται. Από την άλλη, τα μοντέλα υποκατάστατων δεν έχουν κάποια συγκεκριμένη απαίτηση σε υπολογιστικούς πόρους και διασφαλίζουν πολύ γρήγορους υπολογισμούς καθώς αντικαθιστούν, σε κάποιο βαθμό, το ακριβά υπολογιστικά μοντέλα προσομοίωσης. Ο βασικός τους στόχος είναι η δημιουργία ενός μοντέλου που είναι ακριβές σε μια συγκεκριμένη περιοχή του χώρου αναζήτησης (συνήθως γύρω από το ολικό βέλτιστο) και επομένως οδηγούν ευφυώς την βελτιστοποίηση [[Couckuyt et al., 2013](#)]. Η σημαντική δυναμική των μοντέλων αυτών παρουσιάζεται στα Κεφάλαια 8 και 9 μέσω της ανάπτυξης νέων μεθοδολογιών και αλγορίθμων που βασίζονται στα υποκατάστατα μοντέλα

για προβλήματα προσομοίωσης-βελτιστοποίησης υδατικών συστημάτων υπό αβεβαιότητα (δηλ., συστήματα που ελέγχονται από στοχαστικές εισόδους).

## ΕΠΙΣΚΟΠΗΣΗ ΚΑΙ ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

Ο κύριος στόχος της παρούσας διδακτορική διατριβής είναι η ανάπτυξη καινοτόμων εργαλείων και μεθοδολογιών για την ρεαλιστική μοντελοποίηση και προσομοίωση των υδρομετεωρολογικών διεργασιών (δηλ., παραγωγή συνθετικών υδρομετεωρολογικών χρονοσειρών με τις επιθυμητές πιθανοτικές και στοχαστικές ιδιότητες), και ταυτόχρονα αντιμετωπίζει τον επιπρόσθετο υπολογιστικό φόρτο που προκύπτει από την χρήση συνθετικών χρονοσειρών μεγάλου μήκους ως εισόδοι σε προβλήματα προσομοίωσης-βελτιστοποίησης. Κατά συνέπεια, εξασφαλίζει την πρακτική υλοποίηση προβλημάτων βελτιστοποίησης υδατικών συστημάτων υπό το καθεστώς αβεβαιότητας.

Πιο συγκεκριμένα, ο βασικός στόχος της παρούσας διδακτορική διατριβής είναι διττός και αφορά:

α) Την ανάπτυξη μη-Γκαουσιανών στοχαστικών μοντέλων προσομοίωσης που είναι ικανά να αναπαράγουν τις ιδιαιτερότητες που συνήθως συναντώνται στις υδρομετεωρολογικές διεργασίες, όπως είναι η διαλείπουσα φύση, η αυτο- και έτερο- συσχέτιση, η περιοδικότητα, καθώς επίσης και η ανά χρονική κλίμακα μεταβαλλόμενη πιθανοτική και στοχαστική συμπεριφορά των μεταβλητών (Κεφάλαια 3 με 7).

β) Την ανάπτυξη κατάλληλων μεθοδολογιών βελτιστοποίησης βασισμένων σε μοντέλα υποκαταστατών που είναι αποτελεσματικές στην αντιμετώπιση προβλημάτων βελτιστοποίησης-προσομοίωσης υδατικών συστημάτων υπό το καθεστώς αβεβαιότητας, δηλ., όταν γίνεται χρήση στοχαστικών εισόδων στην διαδικασία προσομοίωσης-βελτιστοποίησης (Κεφάλαια 8 με 9).

Η παρούσα διατριβή μπορεί να διαβαστεί στο σύνολό της, ή στη βάση μεμονωμένων κεφαλαίων. Κάθε κεφάλαιο βασίζεται πάνω σε δημοσιευμένα ή υπό-κρίση επιστημονικά άρθρα, και κάθε ένα είναι αυτοτελές με δική του εισαγωγή, μεθοδολογία, αποτελέσματα και συμπεράσματα. Το περιεχόμενο (με τυπική γραφή), η βασική συνεισφορά και τα ευρήματα (με πλάγια γραφή) κάθε κεφαλαίου περιγράφονται παρακάτω:

Το **Κεφάλαιο 2** συνοψίζει τα κύρια χαρακτηριστικά των υδρομετεωρολογικών διεργασιών καθώς επίσης και τις επικρατέστερες ως τώρα μεθοδολογίες μοντελοποίησης και προσομοίωσης τους.

[1] Αυτό το κεφάλαιο παρέχει επισκόπηση των πλέον σύγχρονων πρακτικών μοντελοποίησης και προσομοίωσης για την γέννηση συνθετικών χρονοσειρών υδρομετεωρολογικών διεργασιών, και συζητά κατά πόσο επιτυγχάνουν την αναπαραγωγή των βασικών χαρακτηριστικών τους.

Το **Κεφάλαιο 3** διερευνά την καταλληλότητα μιας συγκεκριμένης κατηγορίας στοχαστικών μοντέλων, που χρησιμοποιείται ευρέως για την παραγωγή συνθετικών χρονοσειρών στον τομέα υδρολογίας, και συγκεκριμένα των γραμμικών στοχαστικών μοντέλων με θόρυβο από μη-Γκαουσιανές κατανομές.

[2] Το Κεφάλαιο αυτό αποκαλύπτει ένα σημαντικό ελάττωμα αυτής της κατηγορίας μοντέλων, αποκαλούμενο ως περιβάλλουσα συμπεριφορά (*envelope behavior*), το οποίο παρέμενε κρυμμένο για πάνω από μισό αιώνα. Τα μοντέλα αυτά είναι επιρρεπή στην παραγωγή μη

φυσικών, και κατά συνέπεια ασυνεπών δομών εξάρτησης που δεν παρατηρούνται στις φυσικές διεργασίες. Αυτή η συμπεριφορά αποδίδεται στον μηχανισμό γέννησης τους, που στερείται ρητής υπόθεσης όσον αφορά την από κοινού δομή εξάρτησης (σήμερα μοντελοποιείται μέσω συναρτήσεων *copula*).

Το **Κεφάλαιο 4** εισάγει το επονομαζόμενο από κοινού μοντέλο πιθανότητας Nataf (NDM; Nataf's Joint Distribution Model), μια κεντρική ιδέα της παρούσας διατριβής, που σχετίζεται με την έννοια των (Gaussian) *copulas*, και με τη σειρά του επιτρέπει την μοντελοποίηση και προσομοίωση μη-Γκαουσιανών τυχαίων (δεσμευμένων ή μη) μεταβλητών και στοχαστικών ανελίξεων (διαδικασιών). Το Κεφάλαιο ξεκινάει με μια εισαγωγή της θεωρητικής βάσης του μοντέλου, και την περιγραφή, μέσω του NDM, της από κοινού πολυμεταβλητής κατανομής μη-Γκαουσιανών τυχαίων μεταβλητών. Εν συνέχεια, το μοντέλο NDM επεκτείνεται στις περιπτώσεις δεσμευμένων κατανομών και στοχαστικών ανελίξεων. Το Κεφάλαιο επίσης περιέχει πλήθος παραδειγμάτων προσομοίωσης τυχαίων μεταβλητών και στοχαστικών ανελίξεων, με διακριτές, συνεχείς και μικτού τύπου περιθώριες κατανομές.

- [3] Η ιδέα του NDM και τα συναφή εργαλεία παρέμεναν άγνωστα στην υδρολογική κοινότητα για χρόνια, καθώς δεν υπάρχει άμεση αναφορά σε αυτά. Η παρούσα διατριβή αποτελεί την πρώτη εργασία στον τομέα αυτό, ενώ τυποποιεί την ιδέα και παρέχει μια εκτενή αντιμετώπιση του θέματος.
- [4] Ένα επιπλέον καινοτόμο σημείο της διατριβής αποτελεί η χρήση του NDM για την εξαγωγή της πολυμεταβλητής δεσμευμένης κατανομής.
- [5] Διατύπωση γενικών οδηγιών για την κατασκευή στοχαστικών ανελίξεων και μοντέλων προσομοίωσης βασισμένων στο NDM, πέραν αυτών που περιγράφονται στο παρόν κείμενο (δες επόμενο Κεφάλαιο).
- [6] Τέλος, μια ακόμα συνεισφορά αυτού του Κεφαλαίου είναι η ανάπτυξη μιας απλής και ευέλικτης διαδικασίας Monte Carlo για την αναγνώριση των επονομαζόμενων ισοδύναμων συντελεστών συσχέτισης που έχουν ένα ιδιαίτερα σημαντικό, αλλά συχνά παραμελημένο, ρόλο στην ανάπτυξη μεθοδολογιών βασισμένων στο μοντέλο NDM.

Το **Κεφάλαιο 5** εστιάζει στην μοντελοποίηση και προσομοίωση στάσιμων στοχαστικών ανελίξεων, και συγκεκριμένα αφορά δυο σημαντικά χαρακτηριστικά των υδρομετεωρολογικών διεργασιών που είναι η μη-Γκαουσιανή φύση των κατανομών (περιλαμβανομένης της διαλείπουσας φύσης) και η δομή αυτό-συσχέτισης τους (μακροπρόθεσμη ή βραχυπρόθεσμη).

- [7] Χτίζοντας πάνω στα ευρήματα του Κεφαλαίου 4, αναπτύχθηκαν δυο καινοτόμα μοντέλα βασισμένα στην ιδέα του NDM. Το μοντέλο *Symmetric Moving Average (nearly) To Anything (SMARTA)* και το μοντέλο *Contemporaneous Multivariate Autoregressive (nearly) to Anything (CMARTA)*. Και τα δυο μοντέλα είναι ικανά να προσομοιώσουν στάσιμες μονομεταβλητές ή πολυμεταβλητές (έτερο-συσχετισμένες) ανελίξεις με οποιαδήποτε δομή εξάρτησης (μακροπρόθεσμη ή βραχυπρόθεσμη) και περιθώρια κατανομή.
- [8] Η θεωρητική βάση των παραπάνω μοντέλων και η ευέλικτη φύση τους περιγράφεται μέσα από μια σειρά υποθετικών και πραγματικών σεναρίων προσομοίωσης, καθώς και μέσα από τη σύγκρισή τους με άλλα γνωστά μοντέλα προσομοίωσης.

Το **Κεφάλαιο 6** αφορά στην μοντελοποίηση κυκλο-στάσιμων ανελιξων (π.χ., μηνιαίων) των οποίων η προσομοίωση είναι αναμφισβήτητα μια πρόκληση, δεδομένης της εποχιακά μεταβαλλόμενης δομής συσχέτισης και των περιθωρίων κατανομών.

- [9] Το Κεφάλαιο αυτό εισάγει ένα κυκλ-οστάσιμο NDM μοντέλο (δες επίσης Κεφάλαιο 4), με την ονομασία *Stochastic Periodic AutoRegressive To Anything (SPARTA)*, το οποίο επιτρέπει την προσομοίωση μονομεταβλητών ή πολυμεταβλητών κυκλο-στάσιμων ανελιξων με οποιοσδήποτε, περιοδικά μεταβαλλόμενη, περιθώρια κατανομή.
- [10] Θεωρητικά και πρακτικά οφέλη της προτεινόμενης μεθόδου, συγκρινόμενα με αποτελέσματα από άλλα ευρέως χρησιμοποιούμενα στοχαστικά μοντέλα, παρουσιάζονται μέσα από πραγματικά, καθώς και υποθετικά, παραδείγματα μηνιαίας προσομοίωσης και αφορούν τόσο μονομεταβλητές αλλά και πολυμεταβλητές διαδικασίες.
- [11] Μια χαρακτηριστική συνεισφορά αποτελεί η αναπαραγωγή δομών εξάρτησης που δεν μπορεί να προκύψει με χρήση κλασικών στοχαστικών μοντέλων (ένα ζήτημα που συζητείται επίσης στο Κεφάλαιο 3).

Το **Κεφάλαιο 7** συζητά και αντιμετωπίζει το πρόβλημα της παραγωγής συνθετικών χρονοσειρών που είναι συνεπείς σε πολλαπλά χρονικά επίπεδα. Αυτή η απαίτηση είναι υψίστης σημασίας σε πολλές εργασίες που σχετίζονται με την έννοια της διακινδύνευσης.

- [12] Αυτό το Κεφάλαιο παρουσιάζει μια καινοτόμα προσέγγιση, αποκαλούμενη ως *Nataf-based Disaggregation To Anything (NDA)*, για την σύζευξη, μέσω διαδικασιών επιμερισμού, στοχαστικών μοντέλων βασισμένων στην προσέγγιση *Nataf* (π.χ. Κεφάλαια 4, 5 and 6).
- [13] Η ανεξαρτησία από την χρονική κλίμακα και ο “top-down” χαρακτήρας της μεθόδου *NDA* επιτρέπει την ανάπτυξη μιας ποικιλίας από στοχαστικά σχήματα προσομοίωσης (μέσω της σύζευξης πολλαπλών μοντέλων) για την παραγωγή, συνεπών σε πολλαπλά χρονικά επίπεδα, συνθετικών χρονοσειρών από υδρομετεωρολογικές διεργασίες που έχουν οποιαδήποτε κατανομή και δομή συσχέτισης (περιλαμβανομένων στάσιμων και κυκλο-στάσιμων διαδικασιών). Στην παρούσα διατριβή, αναπτύσσεται ένα ενοποιημένο σχήμα προσομοίωσης τριών χρονικών επιπέδων για την παραγωγή πολυμεταβλητών χρονοσειρών, συνεπών από την ετήσια μέχρι την ημερήσια χρονική κλίμακα, με οποιαδήποτε πιθανοτική κατανομή και δομή συσχέτισης.
- [14] Η επίδοση του σχήματος τριών επιπέδων επιβεβαιώνεται μέσω δυο διαφορετικών παραδειγμάτων, ενός που αφορά την προσομοίωση της ημερήσιας βροχής-απορροής σε μια λεκάνη απορροής, και ενός που αφορά στην παραγωγή διαλειπουσών μη-Γκαουσιανών χρονοσειρών βροχής σε τέσσερις σταθμούς.
- [15] Η ευελιξία της μεθόδου *NDA* να προσομοιώσει διαδικασίες σε ακόμη χαμηλότερες χρονικές κλίμακες αποδεικνύεται περαιτέρω από ένα επιπλέον παράδειγμα που περιλαμβάνει τον επιμερισμό ημερήσιας βροχής σε ωριαία.

Όπως συζητήθηκε στο κεφάλαιο 1.3, η χρήση στοχαστικών εισόδων σε συνδυασμό με μοντέλα προσομοίωσης και/ή μεθόδους βελτιστοποίησης (δες κεφάλαιο 1.1) θέτουν ένα πρακτικό πρόβλημα καθώς αυξάνουν δραματικά τον απαιτούμενο υπολογιστικό φόρτο. Τα επόμενα δυο Κεφάλαια στοχεύουν στην αντιμετώπιση αυτού του ζητήματος.

Το **Κεφάλαιο 8** εξετάζει το πρόβλημα της αντιμετώπισης χρονοβόρων πολυμεταβλητών προβλημάτων βελτιστοποίησης στα πλαίσια περιορισμένου υπολογιστικού χρόνου/φόρτου. Σαν παράδειγμα, στοχεύουμε στην ανάπτυξη επιχειρησιακών κανόνων λειτουργίας για συστήματα πολλαπλών ταμειωτήρων - ένα δύσκολο πρόβλημα, που προκύπτει από τον αριθμό



των μεταβλητών απόφασης, τους αντικρουόμενους στόχους, τη μη γραμμικότητα των δυναμικών του συστήματος και την υδρολογική αβεβαιότητα. Αυτή η αβεβαιότητα μπορεί να αντιμετωπιστεί μέσω της σύζευξης μοντέλων προσομοίωσης με πολυκριτηριακούς αλγορίθμους βελτιστοποίησης και τη χρήση στοχαστικών υδρολογικών χρονοσειρών σαν εισόδους. Ωστόσο, αυτή η προσέγγιση έχει μεγάλο υπολογιστικό φόρτο και θέτει πρακτικά εμπόδια στην αποτελεσματική εξερεύνηση του χώρου λύσεων. Το παρόν Κεφάλαιο, σε μια προσπάθεια να αντιμετωπίσει αυτό το πρόβλημα:

[16] Αναπτύσσει μια πολυκριτηριακή έκδοση του γνωστού φειδωλού πλαισίου *parameterization-simulation-optimization (PSO)*, που επιτρέπει την εισαγωγή της υδρολογικής αβεβαιότητας μέσω της χρήσης στοχαστικών εισόδων και πιθανοτικών στοχικών συναρτήσεων.

[17] Εξερευνά την δυνατότητα των επονομαζόμενων *multi-objective surrogate-based optimization (MOSBO)* αλγορίθμων να αντιμετωπίσουν τον υπολογιστικό φόρτο. Συγκεκριμένα, τρεις αλγόριθμοι τύπου MOSBO συγκρίνονται με δυο πολυκριτηριακούς εξελικτικούς αλγορίθμους. Τα αποτελέσματα υποδεικνύουν ότι οι αλγόριθμοι MOSBO είναι ικανοί να παρέχουν εύρωστους επιχειρησιακούς κανόνες υπό αβεβαιότητα πολύ ταχύτερα, χωρίς έλλειψη της γενικότητας.

Το **Κεφάλαιο 9** αποτελεί μια καινοτόμα συνεισφορά όσον αφορά γενικά τα χρονοβόρα προβλήματα προσομοίωσης-βελτιστοποίησης. Τέτοιες περιπτώσεις προκύπτουν όταν η εκτίμηση της στοχικής συνάρτησης επιβάλλει τη χρήση χρονοβόρων μοντέλων προσομοίωσης. Ο υπερβολικά μεγάλος χρόνος που απαιτείται από την όλη διαδικασία περιορίζει την εφαρμογή τέτοιων μεθόδων ή επιβάλλει τον τερματισμό της διαδικασίας πολύ νωρίτερα. Όπως συζητείται στο κεφάλαιο **1.3** και παρουσιάζεται στον παρόν Κεφάλαιο, μια πολλά υποσχόμενη στρατηγική για την αντιμετώπιση αυτού του μειονεκτήματος είναι η ενσωμάτωση μοντέλων υποκαταστατών σε αλγορίθμους ολικής βελτιστοποίησης. Σε αυτό το πλαίσιο, εισάγεται ο αλγόριθμος *Surrogate-Enhanced Evolutionary Annealing-Simplex (SEEAS)*. Ο SEEAS συνδυάζει την δύναμη των μοντέλων υποκαταστατών με την αποτελεσματικότητα και αποδοτικότητα των εξελικτικών αλγορίθμων. Ο αλγόριθμος ενσωματώνει τρεις διαφορετικές τεχνικές βελτιστοποίησης (εξελικτική αναζήτηση, προσομοιωμένη ανόπτηση και μεθόδους αναζήτησης κατερχόμενου απλόκου), ενώ οι βασικές αποφάσεις καθοδηγούνται ευφυώς από προσεγγίσεις της στοχικής συνάρτησης μέσω μοντέλων υποκαταστατών.

[18] Η επίδοση του προτεινόμενου αλγορίθμου ελέγχεται έναντι άλλων αλγορίθμων βασισμένων σε μοντέλα υποκαταστατών, τόσο σε θεωρητικά (δηλ., 6 μαθηματικές συναρτήσεις, που δημιουργούν 24 μοναδικά προβλήματα βελτιστοποίησης) όσο και πρακτικά προβλήματα (δηλ., ένα που αφορά την βαθμονόμηση υδρολογικών μοντέλων και ένα πρόβλημα πολλαπλών ταμιευτήρων), με περιορισμένο αριθμό επαναλήψεων (λιγότερες από 1 000).

[19] Τα αποτελέσματα φανερώνουν την δυνατότητα του SEEAS να διαχειριστεί και να αντιμετωπίσει δύσκολα προβλήματα βελτιστοποίησης, που εμπεριέχουν χρονοβόρες προσομοιώσεις.

Το **Κεφάλαιο 10** παρουσιάζει μια περίληψη των πιο σημαντικών επιστημονικών αποτελεσμάτων και συζητά τις δυνατότητες για μελλοντική έρευνα.

Το **Παραρτήματα A, B, C and D** παρέχουν επιπλέον συμπληρωματικό υλικό για τα Κεφάλαια **3, 5, 6 and 7**, αντίστοιχα.

## ΕΠΙΛΟΓΟΣ

Στην παρούσα διδακτορική διατριβή, χρησιμοποιώντας τις θεωρίες στατιστικής, πιθανοτήτων και στοχαστικών ανελίξεων, αναπτύσσεται περαιτέρω και μελετάται διεξοδικά μια κατηγορία θεωρητικά συνεπών, μονό-μεταβλητών και πολύ-μεταβλητών μη-Γκαουσιανών στάσιμων και κύκλο-στάσιμων στοχαστικών μοντέλων. Αυτού του τύπου μοντέλα, ήταν μέχρι πρότινος άγνωστα στην υδρολογική κοινότητα, και αυτή η διατριβή είναι μια πρώτη προσπάθεια μελέτης τους και εναρμόνισης τους με τη στοχαστική υδρολογία.

Τα προτεινόμενα μοντέλα είναι σε θέση να προσομοιώσουν όλα τα χαρακτηριστικά των υδρομετεωρολογικών διεργασιών ενώ ταυτόχρονα χαρακτηρίζονται από απλότητα και φειδωλή παραμετροποίηση. Επιπλέον, με βάση τα παραπάνω μοντέλα, και τη βοήθεια επιμεριστικής διαδικασίας, αναπτύσσεται ένα αρθρωτό στοχαστικό πλαίσιο προσομοίωσης το οποίο επιτρέπει την αναπαραγωγή της πιθανοτικής και στοχαστικής συμπεριφοράς των υδρομετεωρολογικών διεργασιών σε πολλαπλές χρονικές κλίμακες (π.χ., από την ετήσια ως και ημερήσια ή και ακόμη μικρότερες κλίμακες). Τα πλεονεκτήματα των παραπάνω μοντέλων, αλλά και του αρθρωτού πλαισίου στοχαστικής προσομοίωσης, παρουσιάζονται και επαληθεύονται μέσα από πληθώρα υποθετικών και πραγματικών περιπτώσεων στοχαστικής προσομοίωσης.

Τέλος, προκειμένου να διασφαλιστεί η αποτελεσματική εκμετάλλευση και ενσωμάτωση των νέων αυτών εξελίξεων, σχετικών με τη στοχαστική προσομοίωση υδρομετεωρολογικών διεργασιών, στο πλαίσιο του βέλτιστου σχεδιασμού και διαχείρισης υδατικών συστημάτων (σε συνδυασμό με στοχαστικά δεδομένα εισόδου), η παρούσα εργασία αναπτύσσει κατάλληλες μεθόδους και αλγορίθμους βελτιστοποίησης, που βασίζονται σε υποκατάστατα μοντέλα, για τον αποτελεσματικό χειρισμό προβλημάτων υδατικών συστημάτων υπό υδρομετεωρολογική αβεβαιότητα, ελαττώνοντας έτσι σημαντικά τον απαιτούμενο υπολογιστικό φόρτο.

This page is intentionally left blank.



# INTRODUCTION

---

*The only certainty is that nothing is certain*

~ Gaius Plinius Secundus (23–79 AD)

## 1.1 SETTING THE SCENE

Hydrological sciences are not exempted from the introductory aphorism. A phrase that despite the tremendous technological advancements and the rise of the era of information technology and computing is undoubtedly still valid. Evidently, the omnipresence of uncertainty poses an intriguing challenge in decision making process regardless the scientific discipline applied to. Decision making under uncertainty remains, and probably will remain, a fruitful scientific area of continuous development and interest.

The need to account for uncertainty within hydrological decision making is highlighted by the relationship that exists between, climate and water-related engineering works and operations, and human life and security. A characteristic example is that of a water-system comprised of several engineering works especially designed and managed to serve multiple purposes, such as flood protection, energy production and water supply for potable and irrigation purposes. The critical nature of these operations, and their apparent connection with the so-called water-energy-food nexus, pose stringent reliability requirements and require the derivation of optimal solutions, especially when considering the potential impacts of changing climatic conditions.

Key ingredients of every hydrological study, and simultaneously one of the main sources of uncertainty, are hydrometeorological inputs. A historical record of such observations will rarely if ever repeat in the future, due to the high variability, randomness and uncertainty that is inherent in the processes. This type of uncertainty is often referred to as hydrometeorological uncertainty and is arguably of utmost importance in related engineering works and studies.

The scientific discipline of stochastic hydrology, attempts to address the challenging task of handling hydrometeorological uncertainty, though the employment of statistical concepts, probability laws and the theory of stochastic processes. The assumption that hydrometeorological time series (i.e., sequences of observations ordered in time) are realizations of stochastic processes allows their analysis, modelling, simulation and forecasting. In this vein, it can be argued that embracing the existence of randomness and stochasticity in such processes is a first step towards their understanding and the development of uncertainty-aware methodologies for water-systems optimization.

It is widely acknowledged (see next section) that stochastic simulation of hydrometeorological processes, which essentially, translates in generating alternative (statistically equivalent) plausible realizations of the processes (i.e., synthetic time series), providing the means to uncertainty-proof the decision-making process of design and operation of water-systems.

Due to the typical size of historical data, which is not sufficient to extract safe conclusions about the long-term performance of a system, the common procedure entails driving a deterministic simulation model that implements the operation of the associated system with stochastic inputs, i.e., realizations of input hydrometeorological processes that statistically resemble the parent information, which are generally (but not solely) derived from historical data. Therefore, one can obtain long time series of simulated realizations of the system's operation that are conditioned to the statistical characteristics of the stochastic inputs. This approach essentially enables the establishment of Monte Carlo experiments, where the intrinsic uncertainty of the inputs is propagated through a deterministic filter (i.e., the simulation model) in order to derive and assess the probabilistic behaviour of the outputs of interest. Further to this, when the objective is the optimization of the deterministic model's control variables (i.e., model's parameters) with respect to some quantity or metric (i.e., objective), the above procedure can (and should) be embedded within an iterative procedure driven by an optimization algorithm (i.e., establishing uncertainty-aware simulation-optimization frameworks). Arguably, the use of stochastic inputs provides a conceptually flexible and operationally effective approach for handling optimization problems of water-systems under uncertainty, but inevitably, their use, significantly increases the required computational effort.

This Thesis focuses on two important aspects of this procedure, namely,

- a) the realistic stochastic modelling and simulation of hydrometeorological processes, and
- b) the effective and efficient implementation of optimization procedures for water-systems problems under uncertainty (i.e., driven by stochastic inputs).

## 1.2 STOCHASTIC MODELLING AND SIMULATION OF HYDROMETEOROLOGICAL PROCESSES

The idea of using synthetic time series within hydrological studies dates back, more than 100 years, to *Hazen* [1914], who in order to create a *synthetic* realization of annual streamflow combined several historical observations into one enhanced time series record. This simple approach was the first of many that followed (see the review of section 2.3), since his idea was greatly valued by the hydrological community, and significantly motivated the birth of the scientific discipline of synthetic (or operational) hydrology.

According to the classical classification by *Matalas* [1975], synthetic hydrology constitutes a sub-branch of stochastic hydrology, and its emergence owes much to the pivotal works conducted by the Harvard water program [*Maass et al.*, 1962] and the associated works of *Thomas and Fiering* [1962], who were probably the first that employed the theory of stochastic processes for the synthesis of monthly streamflow time series. According to *Koutsoyiannis* [2000], other significant developments that forged the establishment of the field (and stochastic hydrology in general) was the remarkable advances in computing in the 1950's accompanied by the emergence and wide-spread adaptation of Monte Carlo methods in several scientific fields (e.g., physics, biology and finance). Another notable contribution that stemmed from a different scientific domain, was the publication of the now-classic textbook in time series analysis by *Box and Jenkins* [1970] that offered a comprehensive treatment on the subject, as well as provided a detailed classification of linear stochastic models including autoregressive (AR), moving average (MA) and their combination, autoregressive moving average (ARMA) models.

Early works that employed the notion of synthetic time series aimed in assessing the performance of reservoir systems in probabilistic terms, i.e., by evaluating their reliability on the basis of simulated water release data [e.g., Hazen, 1914; Sudler, 1927; Barnes, 1954; Thomas and Fiering, 1962; Klemeš, 1981]. Today, synthetic data are used in a variety of studies (with structure similar to the aforementioned Monte Carlo experiments), among them, the optimal planning and management of reservoir systems [e.g., Koutsoyiannis and Economou, 2003; Celeste and Billib, 2009; Giuliani et al., 2014; Tsoukalas and Makropoulos, 2015a, 2015b; Feng et al., 2017], risk assessment of flood [e.g., Wheeler et al., 2005; Haberlandt et al., 2011; Paschalis et al., 2014; Qin and Lu, 2014; Moustakis et al., 2017] and drought events [e.g., Herman et al., 2016], as well as water resources simulation under future climate conditions [e.g., Fowler et al., 2000; Baltas, 2007; Kilsby et al., 2007; Baltas and Karaliolidou, 2008; Fatichi et al., 2011; Nazemi et al., 2013].

A key requirement for extracting consistent statistical outcomes from such Monte Carlo experiments (provided that the simulation model is a faithful representation of the underlying system dynamics) is the concise representation and simulation of the hydrometeorological inputs; which in turn requires stochastic simulation schemes that are able to account for the main peculiarities of hydrometeorological processes, that is, non-Gaussianity, intermittency, auto-dependence (short- or long-range), cross-dependence and periodicity (see section 2.2).

In a more abstract level, the first two characteristics (non-Gaussianity and intermittency) are associated with the marginal properties of the process, and imply the need for a suitable distribution model. On the other hand, auto- and cross-dependencies are associated with the stochastic (joint) properties of the process, both in time and space, and point out the need for stochastic simulation models *per se*. In fact, if the physical processes to simulate were not (auto- or cross-)correlated, the problem would be substantially simpler, as the generation of synthetic data would be made by, generating uniform numbers and then employing probability integral transformations. Finally, periodicity introduces further complexity, since it implies representing the processes as cyclostationary, thus differentiating their marginal and joint properties not only across different temporal scales but also across seasons (or systematically repeated time intervals, in general).

Arguably an appropriate stochastic simulation scheme should be able to reproduce the probabilistic and stochastic behavior (i.e., marginal and joint properties) of a hydrometeorological process, which varies according to the variable type (e.g., rainfall, streamflow or temperature) and the time-scale of study (e.g., annual, monthly, daily or finer).

The need for generic simulation schemes that allow producing synthetic data from multiple distributions primarily originates from the fact that the probabilistic behavior of many of hydrometeorological processes is not satisfactorily captured by classical stochastic models (see the review of section 2.3). Many of these models (i.e., classic linear stochastic models, point process models, resampling models) are not designed to reproduce significant probabilistic aspects of the processes (e.g., maxima and minima, associated with the tails of the distribution), since their standard hypothesis does not lie in the reproduction of a specific distribution, but the resemblance of some low-order statistics (e.g., mean, variance, skewness) and correlations in time and space. Further to this, as shown by Tsoukalas et al. [2018a], and further discussed in Chapter 3, usual modelling strategies can lead to bounded, hence unrealistic, and non-natural dependence patterns, even though the *essential*, low-order statistical characteristics of the parent data may be well-preserved.

Furthermore, the reproduction of the distribution function of a process is considered of paramount importance, as suggested by both theoretical reasoning and empirical evidence. This is also emphatically highlighted by *Klemeš and Borůvka [1974]*, who argue that (our emphasis):

*Simulation of a serially correlated series with a given marginal distribution is one of the important prerequisites of synthetic hydrology and of its applications to analysis of water resource systems.*

It is worth noting that the literature offers alternative approaches for synthetic time series generation, such as the so-called two-part models and the recently emerged copula-based type of models. These type of models are capable of synthesizing realizations with the target marginal distributions, yet they are constrained by narrow type of correlation structures (e.g., two-part models typically neglect temporal dependence, i.e., auto-dependence) and cumbersome generation mechanisms (for further details see section 2.3).

Nevertheless, a common characteristic of most of the existing simulation approaches is their focus on simulating processes at a single time scale and do not explicitly account for the reproduction of the process's properties (either in term of a distribution function or a set of statistical properties) at multiple temporal levels. Highlighting that multi-scale simulation of hydrometeorological processes still remains an open challenge. For a detailed problem description, as well as a potential remedy that combines the new developments in modelling and simulation of hydrometeorological processes of Chapter 4-6, into an integrated scheme, see Chapter 7.

Arguably, the primary difficulty in simulating hydrometeorological processes, originates from the fact that the classical linear stochastic models (see section 2.3.1), which are capable of modelling and simulating, univariate and multivariate, stationary and cyclostationary, processes with a wide range of dependence structures, are unable to reproduce the non-Gaussian and intermittent nature of hydrometeorological processes, since most of them are formally developed for the simulation of Gaussian processes. This inconvenience may also be related with Chester Kisiel's [1967] *pray to the theoretical hydrologist*, which reads [*Klemeš, 1997 p. 288*]:

*Oh, Lord, please keep the world linear and Gaussian.*

### 1.3 OPTIMIZATION OF WATER-SYSTEM PROBLEMS UNDER UNCERTAINTY

Coupling of simulation and optimization methods is a powerful technique that has gained significant attention in water resources science and technology, since it ensures great advantages over the traditional individual implementation of the two approaches [e.g., *Koutsoyiannis and Economou, 2003*]. In this context, a simulation model is used to faithfully represent the dynamics of the system under study in subsequent time steps and next to evaluate its overall performance against one or more user-specified criteria. Provided that these criteria are expressed in terms of objective function, simulation can be driven by an optimization model, which employs systematic search through the parameter (or decision) space to maximize the system performance; at each trial, new values are assigned to the control variables of the simulation model, which runs automatically to update the value of the objective function.

Combined simulation-optimization schemes for water resource systems can be generally classified into two categories [*Tsoukalas et al., 2016*]: (a) Decision-making problems, in which the system properties and associated processes are known a priori, but either some of its design



quantities or its management policy are unknown; and (b) calibration problems (or inverse problems), in which some internal properties of the system, either physical or conceptual, are unknown and have to be inverted by minimizing the departures of the simulated responses against the observed ones. Despite their different rationale, both types of problems suffer from significant uncertainties and complexities, and they are subject to multiple (and often conflicting) criteria as well as numerous constraints.

The need for advanced global optimization tools (e.g., evolutionary algorithms) has been early recognized by the hydrological community, which has significant experience in their use and also remarkable contribution in their development. In the literature are found numerous reviews of optimization approaches in such problems. For instance, in the context of water resources planning and management, we distinguish the works by *Labadie [2004]*, *Fowler et al. [2008]*, *Nicklow et al. [2010]*, *Reed et al. [2013]* (emphasis to multiobjective applications) and *Ahmad et al. [2014]*. The literature for hydrological calibration is even more extended. For convenience, we highlight the recent works by *Duan [2013]* and *Efstratiadis and Koutsyiannis [2010]*, who provide a comprehensive review of global and multiobjective calibration approaches, respectively. It is also worth mentioning the article by *Maier et al. [2014]*, who summarize the current status of evolutionary algorithms and other metaheuristics, and highlight new directions for future research across water resources applications.

In the whole computational procedure, simulation is by far the most time-consuming component. As models become more complex and data-demanding, their requirements in computational time and/or CPU increase substantially [e.g., *Tolson and Shoemaker, 2007; Keating et al., 2010; Razavi et al., 2010; Efstratiadis et al., 2015; Tsoukalas and Makropoulos, 2015b, 2015a, Tsoukalas et al., 2015b, 2015a, 2016*]. A typical example is the case of physically-based hydrological models of fine spatial and temporal resolution, in contrast to lumped conceptual rainfall-runoff models.

In other applications, referred to as stochastic simulation-optimization problems (i.e., Monte Carlo experiments, also involving optimization; see section 1.1), the computational effort increases by orders of magnitude due to the use of synthetic (instead of historical) time series of very large length (e.g., thousands of years), in order to provide estimations for probabilistic quantities (e.g., reliability, risk) with satisfactory accuracy. Depending on the number of parameters and the irregularity of the response surface, the optimization algorithm may need to call the simulation model thousands of times, in order to converge to a good solution. Therefore, the time effort of simulation imposes a practical barrier to optimization, which is necessary to run with significantly restricted *budget*, by means of maximum allowable number of function evaluations. For instance, consider a simulation model that requires approximately 1.5 minutes for a single simulation run and an optimization algorithm that requires 10 000 function evaluations (iterations) to approximate the global minimum. Such a procedure would last more than ten days, which makes it practically infeasible.

According to *Razavi et al. [2010]*, the approaches to alleviate the computational burden imposed by time-consuming simulation models are classified into four main categories: (1) parallel computing [e.g., *Schutte et al., 2004; Cheng et al., 2005; Vrugt et al., 2006; Feyen et al., 2007; He et al., 2007; Regis and Shoemaker, 2009; Dias et al., 2013*]; (2) computationally efficient optimization algorithms [e.g., *Tolson and Shoemaker, 2007; Kuzmin et al., 2008; Tan et al., 2008; Tolson et al., 2009*]; (3) strategies to avoid opportunistically (expensive) model evaluations [e.g., *Ostfeld and Salomons, 2005; Razavi et al., 2010; Matott et al., 2012*]; and (4) surrogate modelling techniques, also referred to as meta-modelling [*Blanning, 1975*], function

approximation, response surface modelling and model emulation [Razavi et al., 2012a], where surrogate approaches are used to approximate the responses of the original simulation model. It is important to remark that in the context of combined simulation-optimization schemes, surrogate models play the role of black-box approaches that aim to establish a data-driven relationship between the control variables of the simulation model (i.e., explanatory variables) and the objective function of the optimization model (i.e., response variable). Parallel computing, on the other hand, allows the execution of independent simulations by multiple processors, and inevitably requires significant investments in hardware infrastructure, which makes it impractical for common use. Note that in order to reduce the time of computations by three orders of magnitude – a reasonable requirement when dealing with complex simulation models – 1 000 parallel processors should be used, which is far from realistic. The other two options, i.e., the improvement of efficiency of existing algorithms, as well as the interruption of the function evaluation procedure, when the model performance seems to be very poor from early steps of simulation, may save some time but not as much as required. Surrogate models do not have any specific requirements in computer resources and also ensure very fast computations, since they replace, to some context, the (expensive) simulation model. Their key objective is to generate models that are accurate in a certain region of the search space (i.e., around a potential optimum) and thus intelligently guide the optimization [Couckuyt et al., 2013]. The significant potential of such methods is also illustrated in Chapter 8 and 9 through the development of new surrogate-based methodologies and algorithms for water-system simulation-optimization problems under uncertainty (i.e., systems driven by stochastic inputs).

#### 1.4 THESIS OVERVIEW AND CONTRIBUTION

The main aim of this Thesis is to provide innovative tools and methodologies for the realistic modelling and simulation of hydrometeorological processes (i.e., the generation of synthetic hydrometeorological time series with the desirable probabilistic and stochastic properties), and simultaneously tackle the additional computational effort, which arises when long synthetic time series are used to represent the input uncertainty in simulation-optimization frameworks. Thereby, eventually ensuring the practical implementation of uncertainty-aware water-system optimization problems.

More specifically, the main objectives of this PhD Thesis are twofold, and regard:

- a) The development of novel non-Gaussian stochastic simulation models, able to account also for the other peculiarities typically encountered in hydrometeorological processes, such as, intermittency, auto- and cross- dependence, periodicity, as well as their scale-varying probabilistic and stochastic behavior (Chapter 4 to 7).
- b) The development of surrogate-based optimization methodologies and algorithms that can efficiently and effectively confront water-system simulation-optimization problems under uncertainty, i.e., when using stochastic inputs to drive the simulation-optimization procedure (Chapter 8 and 9).

The Thesis can be read as a whole, or in a Chapter-wise basis. Each Chapter is built upon published or under-review journal articles, and all of them are self-contained with their own introduction, methodology, results and conclusion sections. The content (regular typeface), as well as the main contributions and findings (*italics typeface*) of each Chapter are described below:

**Chapter 2** summarizes the characteristic properties of hydrometeorological processes, as well as the prevailing modelling and simulation methodologies.

[1] *This Chapter reviews the current state-of-the-art modelling and simulation practices for synthetic times generation of hydrometeorological processes, and discusses whether they can satisfactory resemble key the characteristics of such processes.*

**Chapter 3** explores the applicability of a particular class of stochastic models, extensively used for synthetic time series generation within the hydrological domain, that of linear stochastic models coupled with non-Gaussian white noise.

[2] *This Chapter reveals a major flaw of this type of models, the so-called “envelope behavior” that remained well-hidden for over half a century. These models are prone to the establishment of non-natural, hence physically inconsistent dependence patterns which cannot be observed in natural processes. This behavior is attributed to their generation mechanism which lacks of explicit assumption regarding the joint dependence structure (nowadays modelled using copulas functions) of the process.*

**Chapter 4** introduces the so-called Nataf’s joint distribution model (NDM), a pivotal concept of this Thesis, that is closely related with the notion of (Gaussian) copulas, which in turn allows modelling and simulation (unconditional and conditional) of non-Gaussian random variables and processes. The Chapter begins with an introduction of the theoretical basis of the model, and the establishment, through NDM, of the multivariate joint distribution of non-Gaussian random variables, which was also its original purpose. Beyond this, NDM is progressively extended to conditional distributions and stochastic processes. These are supported by several simulation examples, that include correlated random variables (or processes) with continuous, discrete and mixed-type marginal distributions.

[3] *The concept of NDM and the related constructs have been unknown within hydrological community for years, since there are no direct reference to it. This Thesis is the first work within the domain that formalizes it and provides an extensive treatment on the subject.*

[4] *An additional innovation point of this Thesis is the use of NDM for the derivation of multivariate conditional distributions.*

[5] *Formulation of general guidelines for the development of Nataf-based stochastic processes and simulation models, beyond those developed herein (see the next three Chapters).*

[6] *Finally, an additional contribution of this Chapter is the development of a simple and versatile Monte Carlo procedure for the identification of the so-called equivalent correlation coefficients, which have an important, yet often neglected, role in the establishment of NDM-based constructs.*

**Chapter 5** focuses on modelling and simulation of stationary stochastic processes, and particularly concerns two distinguishing characteristics hydrometeorological processes, that are non-Gaussianity (including intermittency) and auto-dependence, short- or long-range.

[7] *By building upon the developments of Chapter 4, two novel Nataf-based stochastic models, termed Symmetric Moving Average (nearLy) To Anything (SMARTA) and Contemporaneous Multivariate Autoregressive (nearLy) to Anything (CMARTA), are being developed. Both models are capable of simulating stationary univariate and multivariate contemporaneously cross-correlated processes with any-range dependence and arbitrary marginal distributions.*

[8] *The sound theoretical basis, as well as the flexible character of the models are illustrated through a series of hypothetical and real-world simulation studies, as well as with a comparison with a well-established simulation model.*

**Chapter 6** concerns modelling and simulation of cyclostationary process (e.g., monthly), which are arguably challenging to simulate, due to the seasonally-varying correlations and distributions.

[9] *This Chapter introduces a cyclostationary Nataf-based model (see also Chapter 4), termed Stochastic Periodic AutoRegressive To Anything (SPARTA), which holds out the promise of simulating univariate and multivariate cyclostationary processes with arbitrary marginal distributions, which can also be seasonally varying.*

[10] *Theoretical and practical benefits of the proposed method, contrasted to outcomes from widely-used stochastic models, are demonstrated by means of real-world, as well as hypothetical monthly simulation examples involving both univariate and multivariate time series.*

[11] *An incidental contribution is the reproduction of dependence patterns that cannot be captured by classical stochastic simulation models (an issue also highlighted in Chapter 3).*

**Chapter 7** discusses and addresses the problem of generating multi-scale (temporal) consistent synthetic time series. This modelling requirement is of paramount importance in many water-related risk-based studies, and arguably still remains an open challenge.

[12] *This Chapter presents a novel approach, termed Nataf-based Disaggregation To Anything (NDA), for the pairwise coupling of Nataf-based stochastic models (e.g., Chapter 4, 5 and 6) through disaggregation procedures.*

[13] *The scale-free and “top-down” character of NDA enables the development of a variety of stochastic simulation schemes (by coupling multiple Nataf-based models) for the generation of multi-scale consistent realizations of hydrometeorological processes (univariate and multivariate) with any distribution and correlation structure (including both stationary and cyclostationary ones). Herein, an integrated three-level simulation scheme is being developed for the synthesis of multivariate time series with any distribution and correlation structure that are consistent from annual to daily time scale.*

[14] *The performance of the three-level scheme is validated on a multi-scale basis using two particularly distinct case studies, one that concerns the simulation of daily rainfall-runoff series at a single location, and another that involves the synthesis of non-Gaussian, intermittent rainfall time series at four locations.*

[15] *The flexibility, as well as the modularity of NDA to simulate processes at even lower time-scales is demonstrated through an additional study of disaggregation of daily rainfall to hourly.*

As discussed in section 1.3, the use of stochastic inputs in combination with simulation models and/or optimization techniques (see section 1.1) unwillingly pose a practical barrier in their application, since they substantially increase the required computational effort. The following two Chapters, aim to address this issue.

**Chapter 8** considers the problem of handling time *expensive* multi-objective problems using limited computational budget. As an example, we aim at developing operational rules for multi-reservoir systems; a challenging problem, that arises from the number of decision variables and conflicting objectives, the non-linearity of system dynamics and the hydrological uncertainty. This uncertainty can be addressed by coupling simulation models with multi-

objective optimization algorithms driven by stochastically generated hydrological time series but the computational effort required imposes barriers to the exploration of the solution space. This Chapter in an effort to address this problem,

- [16] *Develops a multi-objective version of the well-established and parsimonious parameterization-simulation-optimization (PSO) framework, that allow to embed hydrological uncertainty though the use of stochastic inputs and probabilistic objective functions.*
- [17] *Explores the potential of multi-objective surrogate-based optimization (MOSBO) to alleviate the computational burden. Three MOSBO algorithms are compared against two multi-objective evolutionary algorithms. The results suggest that MOSBOs are indeed able to provide robust, uncertainty-aware operation rules much faster, without significant loss of neither the generality of evolutionary algorithms nor of the knowledge embedded in domain-specific models.*

**Chapter 9** is as a novel contribution towards time *expensive* water resources simulation-optimization problems. Such cases, arise when the evaluation of the objective function entails the use of an *expensive* simulation model. The excessive time required by the overall procedure may limit the applicability of such approaches, or terminate the optimization process much earlier than required. As discussed in section 1.3 and demonstrated in Chapter 8 and this Chapter, a promising strategy to address these shortcomings is the use of surrogate modelling techniques within global optimization algorithms. With this in mind, the Surrogate-Enhanced Evolutionary Annealing-Simplex (SEEAS) algorithm is being introduced. SEEAS couples the strengths of surrogate modelling with the effectiveness and efficiency of the evolutionary annealing-simplex method. The algorithm combines three different optimization approaches (evolutionary search, simulated annealing and the downhill simplex search scheme), in which key decisions are intelligently guided by surrogate-based approximations of the objective function.

- [18] *The performance of the proposed algorithm is benchmarked against other surrogate-assisted algorithms, in both theoretical (i.e., 6 test functions, configured into 24 unique optimization problems) and practical problems (i.e., one that concerns hydrological calibration and another that concerns multi-reservoir problems), within a limited budget of trials (less than 1000).*
- [19] *The results reveal the significant potential SEEAS in handling challenging optimization problems, involving time-consuming simulations.*

**Chapter 10** concludes the Thesis, by presenting a summary of its most significant research outcomes, and discusses opportunities for further work.

**Appendix A, B, C** and **D** provide additional documentation and supplementary material for Chapter 3, 5, 6 and 7 respectively.

## MODELLING AND SIMULATION OF HYDROMETEOROLOGICAL PROCESSES: A REVIEW OF THE STATE-OF-THE-ART

---

### PREAMBLE

This Chapter provides some basic concepts and definitions used throughout this Thesis (section 2.1), as well as discusses the main characteristics of hydrometeorological processes (section 2.2), which are in turn related with the development of appropriate stochastic simulation models for hydrometeorological time series generation. In this vein, section 2.3 aims at providing a brief overview and discussion on the most prominent modelling and simulation practices for this task; also hinting a critical flaw of linear stochastic models coupled with non-Gaussian white noise, extensively discussed in Chapter 3. Finally, section 2.4 summarizes the identified problems and constraints in existing simulation schemes, which have motivated the development of alternative simulation models (Chapter 4 to 7).

## 2.1 BASIC CONCEPTS AND DEFINITIONS

Before describing the main characteristics of hydrometeorological processes and the associated simulation schemes it is considered useful to provide some basic definitions regarding random variables (RVs) and stochastic processes [Yaglom, 1962; Papoulis, 1991; Lindgren, 2013; Koutsoyiannis et al., 2018], that are important in this Thesis. The following definitions concern continuous RVs, since they are easily extended for the case of discrete RVs (using summation operators, instead of integration).

A *random variable*  $\underline{x}$  is defined by its cumulative distribution function (CDF),  $F_{\underline{x}}(x) := P(\underline{x} \leq x)$ , or simply distribution function, which in turn is related with the corresponding probability density function (PDF) by,  $f_{\underline{x}}(x) := dF_{\underline{x}}(x)/dx$ . The inverse relationship is,  $F_{\underline{x}}(x) = \int_{-\infty}^x f_{\underline{x}}(w)dw$ , where  $w$  is a (dummy) variable used for integration. A *realization* of a RV  $\underline{x}$  is denoted by  $x$  and can be obtained by  $x = F_{\underline{x}}^{-1}(u)$ , where  $F_{\underline{x}}^{-1}(\cdot)$  denotes the inverse cumulative distribution function (ICDF, or quantile function) and  $u \in [0,1]$  denotes probability. Important quantitative measures related with distribution functions are raw ( $\mu'_{\underline{x}}(r)$ ) and central moments ( $\mu_{\underline{x}}(r)$ ) of order  $r$ . Also known as, product moments about the origin and the mean respectively. The former are defined by,  $\mu'_{\underline{x}}(r) := E[\underline{x}^r] = \int_{-\infty}^{\infty} \underline{x}^r f_{\underline{x}}(x) dx = \int_0^1 (F_{\underline{x}}^{-1}(u))^r du$ . Note that,  $\mu'_{\underline{x}}(0) = 1$  and  $\mu'_{\underline{x}}(1) = E[\underline{x}] = \int x f_{\underline{x}}(x) dx$ , which is the mean of RV  $\underline{x}$ , also denoted by  $\mu_{\underline{x}}$ . The central moment of order  $r$  is given by,  $\mu_{\underline{x}}(r) := E[(\underline{x} - \mu_{\underline{x}})^r] = \int_{-\infty}^{\infty} (\underline{x} - \mu_{\underline{x}})^r f_{\underline{x}}(x) dx = \int_0^1 (F_{\underline{x}}^{-1}(u) - \mu_{\underline{x}})^r du$ . Note that,  $\mu_{\underline{x}}(1) = 0$  and  $\mu_{\underline{x}}(2)$ , denotes the variance of  $\underline{x}$ . i.e.,  $\text{Var}[\underline{x}] = \sigma_{\underline{x}}^2 = \mu_{\underline{x}}(2)$ . Additional, and commonly used in hydrology, measures of distribution shape, are the skewness ( $C_{s_{\underline{x}}} = E\left[\left(\frac{\underline{x} - \mu_{\underline{x}}}{\sigma_{\underline{x}}}\right)^3\right] = \frac{\mu_{\underline{x}}(3)}{\sigma_{\underline{x}}^3}$ ) and kurtosis ( $C_{k_{\underline{x}}} = E\left[\left(\frac{\underline{x} - \mu_{\underline{x}}}{\sigma_{\underline{x}}}\right)^4\right] = \frac{\mu_{\underline{x}}(4)}{\sigma_{\underline{x}}^4}$ ) coefficients. Similarly, two RVs  $\underline{x}$  and  $\underline{y}$  are defined by their joint CDF,  $F_{\underline{x}\underline{y}}(\underline{x}, \underline{y})$  or PDF,  $f_{\underline{x}\underline{y}}(\underline{x}, \underline{y}) = \frac{\partial^2 F_{\underline{x}\underline{y}}(\underline{x}, \underline{y})}{\partial \underline{x} \partial \underline{y}}$ . The most common measure of association between  $\underline{x}$  and  $\underline{y}$  is the Pearson's product-moment correlation coefficient defined by  $\rho_{\underline{x}\underline{y}} = \text{Corr}[\underline{x}, \underline{y}] = E[\underline{x}\underline{y}] - E[\underline{x}]E[\underline{y}] / \sqrt{\text{Var}[\underline{x}]\text{Var}[\underline{y}]}$ , where  $E[\underline{x}\underline{y}] = \int \int \underline{x}\underline{y} f_{\underline{x}\underline{y}}(\underline{x}, \underline{y}) d\underline{x}d\underline{y}$  is the first order joint moment of  $\underline{x}$  and  $\underline{y}$ . This definition, implies that both their mean and variance have to be finite; a standard assumption in hydrology, also implied throughout this Thesis (see the discussion in section 4.3.9). Extensions of those definitions to multiple RVs are presented in the above-referenced works.

A *stochastic process*  $\{\underline{x}_t\}_{t \in T}$  is a collection or a sequence of (typically infinite and correlated) random variables indexed using an argument  $t \in T$ . This index typically refers to time and may take continuous or discrete values. Specifically, if  $T$ , the so-called index set, refers to time, and it is comprised of continuous values, then the process is called continuous-time, while if it is comprised of discrete values, it is referred as discrete-time process. Similarly, depending on the state space of  $\underline{x}_t$ , the process can be classified as discrete- or continuous state process. In general, a finite-dimensional stochastic process can be completely defined by the joint distribution  $F_{\underline{x}_{t_1}, \underline{x}_{t_2}, \dots, \underline{x}_{t_k}}(\underline{x}_{t_1}, \underline{x}_{t_2}, \dots, \underline{x}_{t_k})$  of  $\underline{x}_{t_1}, \underline{x}_{t_2}, \dots, \underline{x}_{t_k}$ . However, in practice, such level of sophistication or complexity is rarely required, due to the common, simplifying and convenient assumptions of stationarity and cyclostationarity, as well as ergodicity.

*Strict (or full) stationarity* implies that the any-order distribution function of the process remains invariant regardless the absolute value of argument  $t$ .

*Weak (or second order) stationarity* implies that the second order distribution of the process remains invariant regardless the absolute value of argument  $t$ . In the case of Gaussian processes, strict and weak stationarity are equivalent, since such processes are fully described by the mean value and the covariance structure.

*Cyclostationarity* implies a cyclic (and deterministic, in terms of recurrence) fluctuation of the marginal and joint properties of the process according to the value of argument  $t$ .

A *realization* of a process  $\underline{x}_t$  is denoted by  $x_t$  and if it observed at multiple  $t_i; i = 1, 2 \dots$  is termed time series.

## 2.2 CHARACTERISTICS OF HYDROMETEOROLOGICAL PROCESSES

Depending on the type and time-scale of study, hydrometeorological variables exhibit a variety of different characteristics, that need to be reproduced by a *good* stochastic model. According to [Moran \[1970\]](#), [Salas et al. \[1980\]](#), as well as [Koutsoyiannis \[2005b\]](#) the most prominent are the following:

**Non-Gaussianity:** It is widely acknowledged, and also supported by theoretical and empirical findings [e.g., [Kroll and Vogel, 2002](#); [Koutsoyiannis, 2005c](#); [McMahon et al., 2007](#); [Bowers et al., 2012](#); [Papalexiou and Koutsoyiannis, 2012, 2016](#); [Blum et al., 2017](#)], that hydrometeorological processes are characterized by non-Gaussian distribution functions, which is partially attributed to their non-negative nature (common for all processes) and the intermittent behaviour of some processes (see below). This peculiarity, usually quantified in terms of third-order moments, is amplified as the time scale becomes finer. In fact, it is argued that non-Gaussianity and intermittency (next paragraph) are the origin of most of the theoretical and computational challenges encountered in stochastic hydrology.

**Intermittency:** This phenomenon, denoting the realization of a sequence of zero values interposed between non-zero ones, is a dominant characteristic of finely-resolved (i.e. sub-monthly) processes, such as rainfall and runoff, in arid and semiarid regions [e.g., [Koutsoyiannis, 2006](#)]. The same behaviour is also evident across spatially-distributed processes, e.g., point rainfall simulated at different locations [e.g., [Bardossy and Plate, 1992](#); [Wilks, 1998](#)].

**Auto-dependence** (also referred to as intra-dependence, temporal dependence, memory, or persistence): This is a property of all hydrometeorological processes at all temporal scales, whereby the current value of a process depends on its previous ones. Usually, we distinguish short- and long-range dependence, i.e., SRD and LRD. SRD refers to a stochastic process with a weak autocorrelation structure (e.g., exponential) that decays rapidly. On the other hand, LRD implies the exact opposite. In this case, the autocorrelation structure is a slowly decreasing function (typically power-type) of the time lag. LRD processes are omnipresent in geophysics, hydrology, climate and other scientific disciplines [[Beran, 1994](#); [Koutsoyiannis, 2002](#); [Beran et al., 2013](#); [O'Connell et al., 2016](#)]. LRD dependence is associated with the widely studied Hurst phenomenon [[Hurst, 1951](#)] and fractional Gaussian noise process [[Mandelbrot and Wallis, 1969a, 1969b, 1969c](#)]; which are special cases of LRD, implying a simple scaling behavior. Recently, the term Hurst-Kolmogorov (HK) dynamics was introduced [[Koutsoyiannis and Montanari, 2007](#); [Koutsoyiannis, 2011a](#)] to give credit to the early mathematical work by [Kolmogorov \[1940\]](#).



**Cross-dependence** (also referred to as interdependence): Hydrometeorological processes exhibit statistical interdependencies attributed either to cause-effect relationships (e.g., rainfall-runoff) or to spatial proximity and thus climatic homogeneity [Efstratiadis et al., 2014a]. In several cases, water-system models are driven by more than one inputs, thus highlighting the need for multivariate stochastic simulation schemes, able to represent multiple correlated processes simultaneously.

**Periodicity:** This characteristic implies a kind of non-stationary behavior that cyclically alternates the marginal and joint properties of the process. Within the context of rainfall and streamflow processes, its presence is typically encountered in monthly time scale, which is often the time scale of interest in many water resources management studies. Depending on the type of the process (e.g., wind or solar radiation), periodicity can be detected at finer time scales (e.g., hourly). The effect of periodicity is generally handled either by standardizing the data or by employing explicit cyclostationary schemes with seasonally varying marginal and joint characteristics. We remark that the classic standardization has notable drawbacks since it fails to seasonally vary higher than second order marginal properties (e.g., skewness) and season-to-season correlations coefficients (due to the underlying assumption of stationarity) [Tiao and Grupe, 1980; Bras and Rodríguez-Iturbe, 1985 p. 118]. For a review on the topic (that spans beyond hydrology), see the work of Gardner et al. [2006].

## 2.3 SIMULATION SCHEMES

Arguably, a *good* stochastic model should be able to provide synthetic realizations that resemble the above characteristics. This has led to intriguing challenges, which have motivated a significant amount of research during the last decades. In this vein, a plethora of simulation schemes have been developed (typically for rainfall processes), thus leading to a divert literature landscape. The following paragraphs aim at the ambitious task of providing a critical overview of some of the most widely used simulation schemes. The simulation schemes are organized according to their generation mechanism in a similar manner to the recent classification of rainfall stochastic models by Haberlandt et al. [2011]. Specifically, the models are categorized to: 1) Linear stochastic models, 2) point process models, 3) two-part models, 4) resampling models, and 5) copula-based models. Special attention is given to the class of linear stochastic models, both because they have been for years the main tool for stochastic simulation of hydrometeorological processes, but also because they are the main building blocks of the stochastic simulation models proposed in this Thesis (see Chapters 4, 5, 6 and 7).

### 2.3.1 Linear stochastic models

Early attempts to generate synthetic time series were based on the theory of stochastic processes and the use of linear stochastic models. Almost all these models have been originally developed for the simulation of stationary Gaussian processes, that are either ARMA-type, hence short-range dependent [e.g., Fiering, 1964; Matalas, 1967; Matalas and Wallis, 1971, 1976; Pegram and James, 1972; Camacho et al., 1985, 1987] or long-range dependent [e.g., Mandelbrot and Wallis, 1969a; Ditlevsen, 1971; Mandelbrot, 1971; Mejia et al., 1972; Granger and Joyeux, 1980; Hosking, 1984; Koutsoyiannis, 2000, 2002]. In addition, the hydrological literature also offers several models for cyclostationary Gaussian processes [e.g., Thomas and Fiering, 1962; Salas and Pegram, 1977; Troutman, 1979; Salas et al., 1980, 1982; Tiao and Grupe, 1980; Vecchia, 1985; Bartolini et al., 1988; Salas and Abdelmohsen, 1993; Rasmussen et al., 1996; Shao and Lund, 2004]. These early days developments, as well as detailed descriptions of the associated models can be found in the classic textbooks of stochastic hydrology [Kottegoda, 1980; Salas et

*al.*, 1980; *Bras and Rodríguez-Iturbe*, 1985; *Salas*, 1993; *Hipel and McLeod*, 1994; *Reddy*, 1997]. Beyond hydrology, an in-depth treatment on the subject can be found in textbooks that discuss stochastic processes in general [e.g., *Yaglom*, 1962; *Papoulis*, 1991; *Lindgren*, 2013] or time series modelling and analysis [e.g., *Box and Jenkins*, 1970; *Brockwell and Davis*, 2006; *Cryer and Chan*, 2008; *Tsay*, 2013].

To account for the non-Gaussian and skewed character of hydrometeorological process, these models had to be modified accordingly. The need for simulation schemes able to account for non-Gaussian distributions, was early recognized by many researchers [e.g., *Thomas and Burden*, 1963; *Matalas*, 1967; *Fiering and Jackson*, 1971; *Klemeš and Borůvka*, 1974; *Matalas and Wallis*, 1976; *Lawrance and Kottegoda*, 1977] and motivated the introduction of appropriate adjustments and modifications, which are briefly summarized below.

The standard hypothesis for synthetic time series generation via linear stochastic models does not lie in the reproduction of a specific distribution, but on the resemblance of the *essential* statistical characteristics of the parent historical time series. These are usually expressed in terms of low-order statistics (e.g., mean, variance, skewness) and correlations (that express dependence) in time and space [*Matalas and Wallis*, 1976; *Salas*, 1993]. For a given set of low-order statistics multiple distribution functions may be represented [cf. *Matalas and Wallis*, 1976 p. 66], thus making the simulation problem only partially defined.

The standard approaches to handle skewness within linear stochastic models can be further classified in three categories [*Tsoukalas et al.*, 2018e]: a) Explicit methods, b) transformation-based methods, and c) implicit methods, that treat skewness via employing non-Gaussian white noise for the innovation term.

**Explicit** methods are typically designed, and hence constrained, to generate realizations from a specific distribution family [e.g., *Matalas*, 1967; *Klemeš and Borůvka*, 1974; *Lawrance and Lewis*, 1981a; *Lombardo et al.*, 2012, 2017]. Common approaches include the stationary multivariate lag-1 model with Log-Normal distribution, proposed by *Matalas* [1967], and the univariate first order gamma-autoregressive (GAR) model of *Lawrance and Lewis* [1981a], as well as its periodic extension [*Fernandez and Salas*, 1986] (see also the model of *Klemeš and Borůvka* [1974] with similar capabilities). Recent literature also offers alternative schemes, such as the univariate Log-Normal model of *Lombardo et al.* [2012, 2017] which was specifically designed for the simulation of processes with HK autocorrelation structure. As such, these schemes are either limited in simulating a narrow type of autocorrelation functions or restricted to specific non-Gaussian distributions (typically Gamma and Log-Normal). Furthermore, they are typically able to simulate only univariate processes (with the exception of *Matalas* [1967]), which is a major limitation, since in most water resources applications multiple processes have to be represented simultaneously.

**Transformation-based** approaches initially aim to *normalize* the non-Gaussian historical data through a transformation function, which in turn allows modelling a plethora of dependence structures (due to the well-developed theory of Gaussian processes); next, parameter estimation and simulation are performed on the normalized data and the final product, i.e., the synthetic data, are obtained via the inverse transformation [*Salas et al.*, 1985]. See for example the stochastic simulation software packages SPIGOT [*Grygier and Stedinger*, 1990], SAMS-2003 [*Salas et al.*, 2006] and SAMS-2007 [*Sveinsson et al.*, 2007]. The key component of such schemes is the transformation function. Early attempts used relatively simple conversions, such as Box-Cox, logarithmic, and alternatives, which cannot always ensure a satisfactory normalization (e.g., when the original data are too asymmetric or contain many zero values).

For this reason, for the case of hydrometeorological data, exhibiting significant skewness, more complex schemes have been proposed, involving however several unknown parameters and also require the use of optimization [e.g., *Koutsoyiannis et al., 2008; Papalexiou et al., 2011*]. In fact, the increase of complexity inevitably raises several questions about the transformation function, such as,

- How many parameters should be used?
- How does the sample size affect their estimation?
- In the case of multivariate and cyclostationary simulations, should we use the same transformation function for all processes and seasons?

Nevertheless, even an accurate normalization procedure does not ensure that the inverse transformation (i.e., the normalization – simulation – de-normalization scheme) will preserve both the statistical characteristics (let alone the marginal distribution) and the correlation structure of the original variables [*Salas et al., 1980 p. 73; Bras and Rodríguez-Iturbe, 1985; Lall and Sharma, 1996; Sharma et al., 1997*]. Actually, it is argued that a general method for normalizing all types of data does not exist [*Papalexiou et al., 2011*]. We could also argue that neither an optimal transformation for each specific process exists (particularly in the multivariate case), since the selection and the parameters of the transformation model are prone to subjectivity and indefiniteness. To avoid such ill-transformations, the practice has leaned towards incorporating skewness within the generation mechanism of the stochastic model itself (see below).

**Implicit** schemes embed non-Gaussian white noise within the innovation term, and are arguably the most popular approaches for synthetic time series generation [e.g., *Matalas, 1967; Matalas and Wallis, 1971, 1976; McMahan and Miller, 1971; O’Connell, 1974; Lettenmaier and Burges, 1977; Lawrance and Kottegoda, 1977; Todini, 1980; Vogel and Stedinger, 1988; Koutsoyiannis and Manetas, 1996; Koutsoyiannis, 1999, 2000; Koutsoyiannis et al., 2003b; Unal et al., 2004; Kim et al., 2008; Jothiprakash and Shanthy, 2009; Efstratiadis et al., 2014a; Adeloje et al., 2015; Detzel and Mine, 2017; Montaseri et al., 2017*]. The first attempts are attributed to *Thomas and Fiering [1963]* (presented in the book of *Thomas and Burden [1963]*) and *Fiering and Jackson [1971]* who proposed a univariate simulation scheme for skewed and periodic streamflow data. Their key assumption is the preservation of the desirable statistical characteristics through the generation of white noise from a given distribution, usually the three-parametric Gamma (i.e., Pearson type-III). It is remarked that such approaches generate *explicitly* gamma-distributed variables for the white noise, while the strict *explicitness* is lost when these are synthesized to provide the variables of interest [cf. *Matalas and Wallis, 1976 p. 66; Moschopoulos, 1985*]. Hence, the desirable distribution is only approximately preserved through the reproduction of the process’s moments [*Fiering and Jackson, 1971 pp. 53-57; Lettenmaier and Burges, 1977; Koutsoyiannis and Manetas, 1996*]. Implicit approaches, that employ skewed white noise, have been developed for several other linear stochastic models. We distinguish those of the first order AR model [*Thomas and Fiering, 1963; Matalas, 1967*], the low order univariate stationary ARMA model, [*O’Connell, 1974; Lettenmaier and Burges, 1977*], the modification of *Lettenmaier and Burges [1977]* to the fast fractional Gaussian noise (ffGn) model [*Mandelbrot, 1971*]), the univariate and multivariate broken line model [*Ditlevsen, 1971; Mejia et al., 1972; Bras and Rodríguez-Iturbe, 1985 pp. 266-280*], the first order multivariate periodic autoregressive (PAR) model [see, *Matalas and Wallis, 1976; Koutsoyiannis, 1999*], the univariate and multivariate symmetric moving average (SMA) model [*Koutsoyiannis, 2000, 2002*], which has been recently extended [*Koutsoyiannis et al., 2018 and references therein*] for the reproduction of moments higher than skewness (e.g., kurtosis) by

the inclusion of additional model parameters, as well as the disaggregation-based approach implemented in Castalia software [Efstratiadis et al., 2014a; Tsoukalas et al., 2018c].

Despite the approximation of the marginal distribution, this class of implicit schemes, exhibits a series of other constraints and limitations that are thoroughly discussed in Chapter 3, as well as demonstrated by means of simulation studies in Chapters 5 and 6. Briefly, they are prone to the generation of negative values, encounter difficulties when modelling highly skewed (univariate or multivariate) processes [Todini, 1980; Koutsoyiannis, 1999], while, only few schemes, such as the SMA model, are able to describe a variety of temporal correlation structures. Furthermore, Chapter 3, as well as Tsoukalas et al. [2018a], reveal almost half a century after their introduction, an important and well-hidden, physical inconsistency of the implicit approach related with the reproduction of dependencies through such schemes. Particularly, it shown, both empirically and theoretically that this type of approach may lead to bounded, thus unrealistic and non-natural dependence patterns, that do not agree with observations.

Finally, it is noted that all above categories of linear stochastic models are typically employed for the simulation of hydrometeorological processes at annual and monthly time scales. This is due to the weaknesses discussed above, that limit their capability to handle intermittency without the use of additional modelling *tricks*, such as, truncation of negative values to zero, power-transformation functions or latent Gaussian processes [e.g., Bell, 1987; Bardossy and Plate, 1992; Rasmussen, 2013].

### 2.3.2 Point process models

Point process models, such as the Newman-Scott and Barlett-Lewis rectangular pulse models, are alternative simulation schemes specifically designed for the simulation of fine time scale processes, typically rainfall. This class of models builds upon the theory of continuous-type point processes and have been originally introduced in hydrology by Rodriguez-Iturbe et al. [1987]. Since then numerous simulation schemes have been developed that either improve or extend these schemes [e.g., Rodriguez-Iturbe et al., 1988; Cowpertwait, 1991; Bo et al., 1994; Onof and Wheeler, 1994a, 1995; Onof et al., 2000; Koutsoyiannis and Onof, 2001; Smithers et al., 2002; Koutsoyiannis et al., 2003b; Kilsby et al., 2007; Burton et al., 2008; Evin and Favre, 2008; Kaczmarek et al., 2014; Kossieris et al., 2015, 2016]. Their main advantages are the physical interpretation of the model's parameters and their *signature* feature regards their potential to reproduce the statistical characteristics of rainfall at multiple time scales. However, it is remarked, that similar to the classic linear stochastic schemes, this type of models aim at the resemblance of low order marginal statistics (typically up to skewness coefficient) and not to the reproduction of the marginal distribution of the process. Other notable weaknesses are related with difficulties in simulating multivariate processes and cyclostationary correlation structures, the reproduction of probability dry and extreme events [e.g., Rodriguez-Iturbe et al., 1988; Onof and Wheeler, 1994b], as well as parameter identification through optimization techniques [e.g., Wheeler et al., 2005]. Detailed reviews are given by Onof et al. [2000], Wheeler et al. [2005] and Kossieris et al. [2016].

### 2.3.3 Two-part models

Two-part models (also referred as product models or chain-dependent processes), are fine time scale simulation schemes for intermittent processes (e.g., daily, or sub-daily). This class of models has been introduced by Todorovic and Woolhiser [1975] and later formalized by Katz [1977] for the simulation of univariate intermittent daily rainfall processes. It builds upon the

idea of representing the intermittent rainfall process as the product of two distinct processes: the *occurrence process*, which is utilized to express the realization or not of a certain event (e.g., rain or no rain; wet or dry state), and the *amount* (or *intensity*) process, which in turn is employed to assign an amount value in the case of event occurrence (e.g., rainfall amount, given a realization of a rain event, i.e., wet state). This dichotomy is an attempt to cope with the discrete-continuous nature typically encountered in several hydrometeorological processes such as rainfall. The following discussion is centered around rainfall simulation schemes, which was the main driver for the development of two-part models.

The prevailing approach to model the occurrence process lies in the use of Markov chain models, which provide realizations of the occurrence process that typically alternate between wet and dry states through the use of the so-called transition probability matrix. Most commonly, first-order, two-state models are employed, which are parsimonious and easy to simulate [e.g., *Gabriel and Neumann, 1962; Todorovic and Woolhiser, 1975; Katz, 1977; Stern and Coe, 1984*]. Higher order Markov chain models have also been proposed to better describe the occurrence process, yet at the cost of additional model parameters and complexity [e.g., *Pattison, 1965; Gates and Tong, 1976; Chin, 1977; Wilks, 1999; Srikanthan and Pegram, 2009; Mammias and Lekkas, 2018*]. Beyond Markov chains, an alternative, albeit not so frequently employed, approach lies in the use on alternate renewal processes [e.g., *Buishand, 1978; Foufoula-Georgiou and Lettenmaier, 1987; Acreman, 1990; Wilby et al., 1998*]. A somewhat dated, yet thorough, comparison of Markov chains and alternate renewal processes for modelling daily rainfall occurrence processes is given by *Roldan and Woolhiser [1982]*.

Regarding the amount process, the typical modelling approach, employed by such simulation schemes, uses sequences of independent, identically, distributed (i.i.d.) random variables, which may be cross-correlated in the case of multivariate models, from a variety of distributions functions. The explicit use of distribution models is one of their key advantages, yet the assumption of independence implies that two-part models ignore the serial correlation of amounts (with the exception of few univariate [e.g., *Katz and Parlange, 1995; Lee, 2016; Lombardo et al., 2017*] and multivariate [e.g., *Breinl et al., 2013, 2015*] models that only account for short-range dependence). This is considered as a major limitation for the simulation of processes at sub-daily time scales, where temporal dependence extends for many time lags. In the course of time, a plethora of distribution functions have been employed, that span from classic continuous-type distributions, such as Exponential [e.g., *Todorovic and Woolhiser, 1975; Richardson, 1981; Wilby, 1994*], Gamma [*Katz, 1977; Richardson, 1981; Richardson and Wright, 1984; Stern and Coe, 1984; Srikanthan and Pegram, 2009; Lee, 2017*], Weibull [e.g., *Breinl et al., 2013*] and log-Normal [e.g., *Katz and Parlange, 1995; Lombardo et al., 2017*] to more complex, mixed-type distributions, that arguably better describe the behavior of extremes [e.g., *Foufoula-Georgiou and Lettenmaier, 1987; Wilks, 1998; Furrer and Katz, 2008; Neykov et al., 2014; Breinl et al., 2015; Evin et al., 2018*]. Another common assumption is that the probability distribution of the amount process is conditionally independent of the previous states of the occurrence process. This means that the distribution function at every time step is the same, regardless of the state of occurrence (e.g., wet or dry) at previous time steps. This modelling approach, in combination with the fact that the final process is obtained as the product of occurrence and amount processes, may lead to sudden and sharp transitions (from heavy to no rain at all) among consecutive time steps. As remarked by *Bárdossy and Plate [1992]* and *Wilks [1998]* the same issue is also apparent in spatial scale. However, there are some rare exceptions (limited in univariate and rather complex models) that explicitly model

this behavior [e.g., *Chin, 1977; Katz, 1977; Wilks, 1999*] through the use of additional parameters and conditional distributions.

An important step towards the widespread adaptation and further development of two-part models was the pivotal contribution of *Wilks [1998]* who proposed the first multivariate two-part simulation scheme for daily rainfall processes. The model combines a multivariate first-order, two-state Markov chain model with multivariate distribution sampling of contemporaneously correlated (yet serially independent) random variables. Its basis lies in the establishment of an empirical link between an auxiliary multivariate Gaussian distribution and its mapping to the real-domain via combining the notion of probability integral transformation and the target marginal distributions. Interestingly, as noted by *Tsoukalas et al. [2018e, 2018d]*, and further discussed in section 4.6, the foundations of Wilks' empirical approach can be retrospectively attributed to the theoretical background of the so-called Nataf joint distribution model [*Nataf, 1962*] (and the associated Gaussian copula), which is also a key concept of this Thesis.

The work of Wilks' has motivated the development of numerous multivariate daily rainfall stochastic simulation schemes, that either employ parametric distributions [e.g., *Brissette et al., 2007; Khalili et al., 2009; Srikanthan and Pegram, 2009; Baigorria and Jones, 2010; Mhanna and Bauwens, 2012; Breinl et al., 2013; Lee, 2017*] or use non-parametric ones, in combination with resampling schemes [e.g., *Beersma and Buishand, 2003; Mehrotra, 2005; Mehrotra et al., 2006; Breinl et al., 2013*]. For a short discussion on non-parametric distribution and the associated resampling scheme, see section 2.3.4. Further details and discussion on this type of rainfall models can be found in the review of works *Srikanthan and McMahon [2001]* and *Haberlandt et al. [2011]*.

Beyond the realm of rainfall simulation, such models are key components of many (univariate and multivariate) *weather generator models*, a term probably firstly used by *Richardson and Wright [1984]* and popularized by the review work of *Wilks and Wilby [1999]*. Weather generators facilitate the simulation of additional weather variables (e.g., solar radiation, minimum and maximum temperature), by conditioning them on the state of rainfall process [*Richardson, 1981; Richardson and Wright, 1984*]. However, these schemes are not truly multivariate, in the sense that they consist of two distinct sub-models, one to simulate the intermittent rainfall process and another, typically a low-order Gaussian ARMA model (that links upon the former rainfall model), to simulate the other weather processes. Further to these early day schemes, the literature offers a variety of weather generation models, that build upon, extend or improve these original schemes [e.g., *Semenov and Barrow, 1997; Semenov et al., 1998; Buishand and Brandsma, 2001; Qian et al., 2002; Apipattanavis et al., 2007; Kilsby et al., 2007; Khalili et al., 2009; Flecher et al., 2010; Chen et al., 2014; Breinl et al., 2015, 2017*], as well as detailed review works [*Wilks and Wilby, 1999; Ailliot et al., 2015*].

#### 2.3.4 Resampling models

A well-known alternative simulation scheme is offered by the so-called *non-parametric* approaches, which aim to reproduce the empirical distributions of the observed processes, typically through resampling of historical data (most often using the well-known *k*-nearest neighbor algorithm). This class of models has been pioneered in hydrology by *Lall and Sharma [1996]* and *Sharma et al. [1997]* for the simulation of monthly streamflow processes. Since then, numerous resampling schemes have been developed for the simulation of several hydrometeorological processes at time scales that span from annual [e.g., *Lee and Salas, 2011*],

to monthly [e.g., *Lall and Sharma, 1996; Sharma et al., 1997; Prairie et al., 2007; Lee et al., 2010; Salas and Lee, 2010*], and daily [e.g., *Brandsma and Buishand, 1998; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001; Clark et al., 2004; Mehrotra, 2005; Mehrotra et al., 2006; Apipattanavis et al., 2007; Mehrotra and Sharma, 2007*] or even finer time scales [e.g., *Wójcik and Buishand, 2003; Lee and Jeong, 2014*]. Such approaches have gained particular attention due to their ability to empirically establish marginal distributions that exhibit bi- or multimodality; a characteristic that typically arises in processes driven by multiple (often anthropogenic) generation mechanisms [*Lall and Sharma, 1996; Sharma et al., 1997*]. However, the use of the empirical, *non-parametric*, distributions (instead of fitting a theoretical model) prohibits the extrapolation out of the observed data ranges and the synthesis of unobserved values, which eventually limits their capability to simulate extreme events (low or high). Further to this, the lack of theoretical basis makes it difficult to reproduce long-range dependence and cross-correlations among multiple variables. As *Serinaldi and Kilsby [2014]* critically argue (within the context of rainfall simulation), *Resampling models do not model rainfall but sample the observed values according to suitable rules that preserve the spatiotemporal statistical properties of the rainfall measurements*. Heuristic solutions to the above limitations, such as the optimization-based approach of *Bárdossy [1998]* and the recent scheme of *Borgomeo et al. [2015]*, do not necessarily mitigate these weaknesses. Such schemes are also subject to extremely high computational effort (due to their trial-and-error nature), and they are prone to inherent inefficiencies of optimization algorithms.

### 2.3.5 Copula-based models

Another relatively new option is offered by copula-based simulation schemes. These, build upon the notion of copulas [*Sklar, 1959, 1973*], which provide them with the ability to explicitly model a wide range of distribution functions and dependence structures. For a general discussion on copulas, see for instance, the works of *Embrechts et al. [2003]*, *Nelsen [2007]* or *Joe [2014]*. Copulas have been initially developed for modelling and simulation of random variables, not stochastic processes. Nevertheless, nowadays, the hydrological literature offers several copula-based schemes, typically able to model only short-range dependence structures, specifically designed for the simulation of hydrometeorological processes. Among them, the works of *Bárdossy and Pegram [2009]* and *Serinaldi [2009a]* that proposed simulation schemes for multivariate daily rainfall processes. These schemes are probably the first works that formally use copulas for this purpose. In a similar vein, *Lee and Salas [2011]*, proposed a univariate copula-based model for the simulation of annual streamflow processes, while similar schemes for periodic (typically monthly) univariate [e.g., *Hao and Singh, 2011; Jeong and Lee, 2015*] and multivariate streamflow processes [e.g., *Hao and Singh, 2013; Chen et al., 2015*] have also emerged recently. However, it is argued [*Mikosch, 2005*], that copulas are not directly compatible with the theory of stochastic processes and the associated linear stochastic models, which rely on Pearson's correlation coefficient, since copulas typically employ rank-based correlation statistics (e.g., Spearman's  $\tau_s$  or Kendall's  $\tau$ ) to describe the dependencies among the variables. Furthermore, they are considered more sensitive against sampling uncertainty than classical stochastic schemes, in their attempt to describe complex (i.e., nonlinear) dependencies on the basis of usually limited hydrological data. However, as many researchers argue (see [*Hao and Singh, 2013; Chen et al., 2015*]), they rely on quite complicated and computationally demanding generation schemes, especially in high-dimensional spaces, a fact which may also be related with the emphasis, on the development of (only) short-range dependent simulation schemes.

## 2.4 SUMMARY

Most of the available simulation schemes emphasize on the reproduction of summary statistical characteristics, up to third order (i.e., skewness coefficient), which arguably cannot provide the full behavior of a random variable (or processes), including its tails. Particularly, Chapter 3, focuses on a critical flaw of linear stochastic models with non-Gaussian white noise, which remained well-hidden for over half a century [Tsoukalas et al., 2018a].

Currently, only few stochastic simulation schemes (i.e., two-part and copula-based models) are able to fully and explicitly account for non-Gaussian distributions, yet they are mainly focused on narrow-type of correlation structures (e.g., two-part models often completely neglect temporal dependence, while copula-based schemes often model only few time lags) and involve high-dimensional complex generation mechanisms (e.g., copula-based models). In this respect, to address the associated simulation challenge, Chapter 4 provides the theoretical basis of a new type of models, the so-called Nataf-based models [Tsoukalas et al., 2017a, 2018e, 2018d, 2018b], that can cope with this challenging task. Particularly, by building upon their theoretical background, Chapter 5 and Chapter 6 focus on modelling and simulation of non-Gaussian, stationary and cyclostationary processes respectively.

Furthermore, it is observed that most of available simulation schemes are specifically designed for the simulation of specific type of processes (e.g., rainfall) at a specific time scale (e.g., daily). However, it is argued, that a *good* stochastic model should be able to provide synthetic realizations that resemble the probabilistic behavior and structure of the process across multiple time scales [e.g., Klemeš et al., 1981; Koutsoyiannis, 2005a].

It is well-known that the resemblance of marginal and stochastic properties at a certain time scale (e.g., daily) does not necessarily implies the resemblance of the process's properties at multiple, higher time scales (e.g., annual). This fact imposes the requirement of multi-scale consistency, which is also related with so-called issue of low-frequency variability or over-depression, that is often encountered in many *weather-generation models* [e.g., Wilks and Wilby, 1999].

Depending on the type of study, different aspects of the process may be of interest. For instance, in the case of water resources management studies (e.g., in water supply and/or hydropower reliability studies) that are typically conducted at a monthly basis, it is considered important to simulate both the over-annual correlation structure, the periodic structure at the monthly scale and the marginal distributions. Multi-scale consistency is an important operational requirement, since it can significantly affect the outcome of a Monte Carlo experiment, and hence the probabilistic behavior of the output of interest, and eventually, affect the design and operation of the engineering works.

In this vein, Chapter 7 moves beyond the previously discussed, single-scale, simulation methods, and considers the simulation challenge through the prism of disaggregation/downscaling methods, which in principal aim at the generation of multi-scale consistent realizations via the transfer of information among different time scales.



## ON THE REPRODUCTION OF DEPENDENCIES THROUGH LINEAR STOCHASTIC MODELS WITH NON-GAUSSIAN WHITE NOISE \*

---

*Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.*

~ George Box and Norman Draper [1987 p. 74]

### PREAMBLE

Since the early days of stochastic hydrology back in 1960's, autoregressive (AR) and moving average (MA) models (as well as their extensions) have been widely used to simulate hydrometeorological processes. Initially, AR(1) or Markovian models with Gaussian noise prevailed due to their conceptual and mathematical simplicity. However, the ubiquitous skewed behavior of most hydrometeorological processes, particularly at fine time scales, necessitated the generation of synthetic time series to also reproduce higher-order moments. In this respect, the former schemes were enhanced to preserve skewness through the use of non-Gaussian white noise— a modification attributed to Thomas and Fiering (TF). Although preserving higher-order moments to approximate a distribution is a limited and potentially risky solution, the TF approach has become a common choice in operational practice. In this study, almost half a century after its introduction, we reveal an important flaw that spans over all popular linear stochastic models that employ non-Gaussian white noise. Focusing on the Markovian case, we prove mathematically that this generating scheme provides bounded dependence patterns, which are both unrealistic and inconsistent with the observed data. This so-called *envelope behavior* is amplified as the skewness and correlation increases, as demonstrated on the basis of real-world and hypothetical simulation examples.

The Chapter is structured as follows: section 3.1 provides the historical background of the TF approach; section 3.2 details the issue of *envelope behavior*. Section 3.3 demonstrates this problem through a real-world case study. Section 3.4 and 3.5 discuss and conclude this chapter respectively.

### 3.1 A GLIMPSE OF HISTORY

The celebrated Harvard water program and the development of the so-called Thomas-Fiering (TF) model in the early 60s [Maass et al., 1962; Thomas and Fiering, 1962; Fiering, 1967; Fiering and Jackson, 1971] played a historically crucial role in definition and advancement of the

---

\*Based on:

Tsoukalas, I., S. Papalexiou, A. Efstratiadis, and C. Makropoulos (2018a), A cautionary note on the reproduction of dependencies through linear stochastic models with non-Gaussian white noise, *Water*, 10(6), 771, doi:10.3390/w10060771.

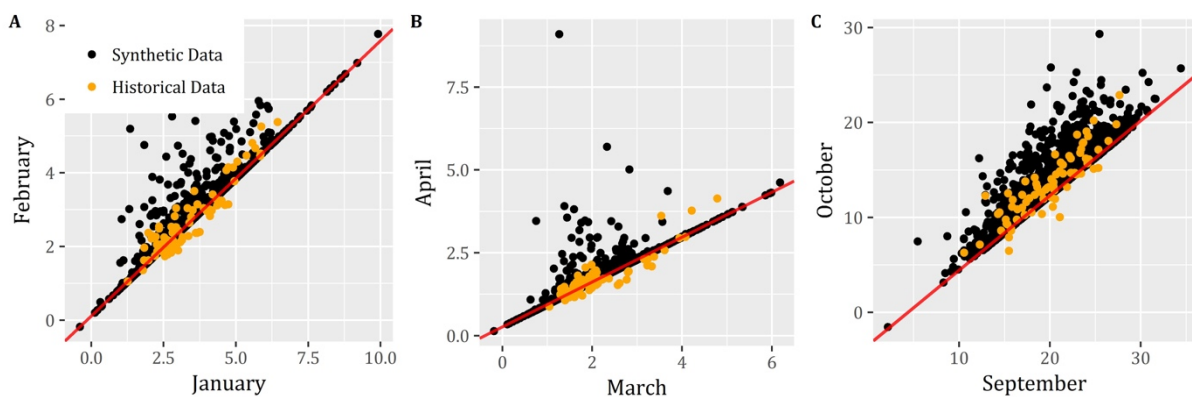
scientific discipline of stochastic hydrology—more specifically, of synthetic hydrology. The emergence of this field was mainly motivated by the need to generate synthetic streamflow data, to be used in water resources planning and management models [Matalas, 1967; Jackson, 1975; Hirsch, 1979]. The use of synthetic streamflow generators (or more generally weather generators) allowed for representing the operation of complex hydrosystems and deriving risk-related quantities that could not be obtained through classical statistics. Among the many different alternative models (see references below, as well as section 2.3.1), the TF model prevailed for many years and still remains a popular choice. To date, the original Thomas-Fiering paper [1962] and the related works of the Harvard water program [Maass et al., 1962; Thomas and Fiering, 1962; Fiering, 1967; Fiering and Jackson, 1971] have been cited in the literature almost 2000 times, a fact highlighting its vast popularity and reasonably justifying its denomination as the *Ford’s Model T* of stochastic hydrology [Klemeš, 1997]. Additionally, more than half a century since its conception, the TF model, its variants, and the associated approach to handle skewness (see below) are standard educational material in most stochastic-hydrology courses and are disclosed in prominent positions in many classic and contemporary textbooks [Kottegoda, 1980; Bras and Rodríguez-Iturbe, 1985; Salas, 1993; Hipel and McLeod, 1994; Reddy, 1997; Loucks and van Beek, 2017]. The wide acceptance of the model is also acknowledged by Salas and Pielke [2003], who asserted that, *the PAR(1) model (also known as the Thomas-Fiering model) is likely one of the most widely used models in hydrology.*

The original TF model is essentially a cyclostationary version of the classic stationary linear autoregressive model of order 1 (AR(1)), also formulated as a periodic autoregressive of order 1 (PAR(1)), in order to account for systematic changes and non-stationarities of statistical characteristics across seasons. The fact that the marginal distributions of many hydrometeorological processes are not Gaussian, motivated Thomas and Fiering [1963] to propose the replacement of the Gaussian white noise with Gamma ( $\mathcal{G}$ ) or Pearson type-III ( $\mathcal{P}$ III) distributed white noise [Fiering and Jackson, 1971, pp. 53-57] in order to account for the skewness coefficient (to our knowledge, this modification first appears in the book of Thomas and Burden [1963]). Note that the  $\mathcal{P}$ III distribution is a simple extension of the  $\mathcal{G}$  distribution, which introduces an additional location parameter.

This approach was subsequently adopted by many other researchers [e.g., Matalas, 1967; McMahon and Miller, 1971; Fiering and Jackson, 1971; O’Connell, 1974; Lawrance and Kottegoda, 1977; Vogel and Stedinger, 1988; Koutsoyiannis and Manetas, 1996; Koutsoyiannis, 1999, 2000; Koutsoyiannis et al., 2003b; Unal et al., 2004; Kim et al., 2008; Jothiprakash and Shanthi, 2009; Efstratiadis et al., 2014a; Adeloje et al., 2015; Montaseri et al., 2017], and can be classified as an implicit one, since it aims to approximate the distribution of the target process via the introduction of non-Gaussian white noise [Tsoukalas et al., 2018e]. Hereafter, we refer to the use of non-Gaussian white noise in linear stochastic models (e.g., AR(1)) as the TF approach.

Nevertheless, herein, we mainly focus on AR models with non-Gaussian white noise, which have been widely adopted in hydrology, and briefly discuss three alternative schemes, two of which are based on moving average (MA) models and one based on an autoregressive moving average model (ARMA). Specifically, we investigate the effect on the established dependence patterns that arise from the use of  $\mathcal{P}$ III white noise within stationary univariate and multivariate linear stochastic models for generating synthetic hydrological data via stochastic simulation. Based on theoretical reasoning and empirical evidence, it is shown that the use of the implicit TF approach results in bounded and thus unrealistic dependence patterns, highlighting this approach’s limitations in simulating skewed hydrometeorological processes.

Our motivation stems from an observation of *Tsoukalas et al. [2018e]*, who noticed this dependence pattern flaw while simulating 2000 years of monthly streamflow data at Aswan dam through the TF approach (i.e., PAR(1) with skewed white noise), hereafter called *envelope behavior*. A characteristic sample of this work is shown **Figure 3.1**, where we depict the scatter plots of historical and synthetic data for three pairs of consecutive months (January–February, March–April, and September–October). It is observed that the synthetically-derived scatter is bounded by a linear threshold, while the historical data clearly extend below this limiting line. It is remarkable that the model reproduces almost perfectly the (often regarded as essential) statistical characteristics of historical data, i.e., the mean, variance, and skewness, as well as the month-to-month linear correlations (Pearson’s), which is the typical measure of statistical dependence that is encountered in all linear stochastic schemes. However, it seems that the preservation of these statistical characteristics does not ensure the generation of fully consistent dependence patterns.



**Figure 3.1** | Comparison of the (A) January–February, (B) March–April, and (C) September–October dependence patterns between historical and synthetic monthly runoff data ( $10^9 \text{ m}^3$ ) of the Nile, at Aswan dam. Synthetic time series were generated by the cyclostationary Thomas-Fiering (TF) approach (adapted by *Tsoukalas et al. [2018e]*); the simulated negative values were not truncated to zero in order to avoid distortion of the dependence pattern). The red line (—) depicts the envelope equation of the TF model (when combined with  $\mathcal{P}$ III white noise. See also Appendix A).

### 3.2 THE ENVELOPE BEHAVIOR OF LINEAR STOCHASTIC MODELS WITH NON-GAUSSIAN WHITE NOISE

#### 3.2.1 The Thomas-Fiering approach

The basic idea of the TF approach lies in using non-Gaussian, skewed, white noise within linear stochastic models in order to resemble the target marginal statistics, i.e., sample mean, variance, and skewness. Note that the derivation of a theoretical formula for the white noise skewness in AR( $p$ ) models of a higher order ( $p \geq 2$ ) aiming to reproduce skewness is theoretically possible but practically of no use, as it involves high-order joint statistics (that are difficult to estimate and are also subject to significant sample uncertainties [*Lombardo et al., 2014*]). Thus, application is possible only based on sample estimates of these joint statistics. This is the major reason why the TF modification was originally restricted in AR(1) models, and thus similarly we also concentrate our main analysis in stationary univariate and multivariate AR(1) models with skewed white noise, while we briefly explore the cases of some other linear stochastic models (i.e., an ARMA and two variants of MA models).

Apparently, the selection of the underlying model determines the stochastic characteristics of the resulting simulation scheme. For example, when an AR(1) model is employed, the overall scheme will reproduce only Markovian autocorrelation structures, while if a more flexible MA-based scheme is used, the simulation scheme will be able to resemble a wider range of correlation structures.

However, regardless of the choice of the underlying model, such schemes exhibit a number of shortcomings and limitations, which are briefly summarized here [Tsoukalas et al., 2018e]: (1) They provide just an approximation of the marginal distribution, as reproducing statistics generally is not equivalent to reproducing the distribution. Furthermore, the resulting distribution is not known a priori (e.g., in general the sum of Gamma distributed variables is not Gamma; see also, Moschopoulos [1985]). We remark that this was acknowledged by the authors [Fiering and Jackson, 1971, pp. 53-57], as well as later remarked by other researchers [Matalas and Wallis, 1976, p. 66; Lettenmaier and Burges, 1977; Koutsoyiannis and Manetas, 1996]; (2) In order to reproduce the skewness of the underlying process it is required (due to central limit theorem) to use white noise with higher skewness [Lettenmaier and Burges, 1977; Kottegoda, 1980; Todini, 1980; Koutsoyiannis, 1999], which can cause, in some cases, failure of the random number generator itself; (3) This simulation scheme generates time series that can have negative values, which is not consistent with many physical processes (e.g., rainfall, wind, streamflow, etc.). This is attributed to the fact that the lower bound of the white noise distribution may be negative in order to match the target statistics (as estimated from observed time series); (4) Finally, we prove and demonstrate in the next sections that this scheme leads to bounded and thus unrealistic dependence patterns that are not observed in natural processes (such as those depicted in Figure 3.1).

### 3.2.2 The envelope behavior in the classical univariate AR(1) model

Let us assume we wish to simulate a continuous-state (not necessarily Gaussian), discrete-time, stationary AR(1) process (also referred to as the Markov process)  $\underline{x}_t, t \in \mathbb{Z}$ , where  $t$  is the time index. The main equation of the model reads:

$$\underline{x}_t = a_1 \underline{x}_{t-1} + \underline{\varepsilon}_t \quad (3.1)$$

where  $a_1 = \rho_1 := \text{Corr}[\underline{x}_t, \underline{x}_{t-1}]$  is a model parameter and  $\underline{\varepsilon}_t$  denotes an i.i.d. random variable (RV) known as white noise or the innovation term. The theoretical autocorrelation function (ACF) of the AR(1) model is  $\rho_\tau := \text{Corr}[\underline{x}_t, \underline{x}_{t-\tau}] = a_1^{|\tau|}$ , where  $\tau$  stands for the time lag. The mean  $\mu_{\underline{x}_t} := E[\underline{x}_t]$  and variance  $\sigma_{\underline{x}_t}^2 := \text{Var}[\underline{x}_t] = E[(\underline{x}_t - \mu_{\underline{x}})^2]$  of  $\underline{x}_t$  are related with those of  $\underline{\varepsilon}_t$  via the following equations (hereafter, due to stationarity, the index  $t$  will be omitted when possible):

$$\mu_{\underline{\varepsilon}} = \mu_{\underline{x}}(1 - a_1) \quad (3.2)$$

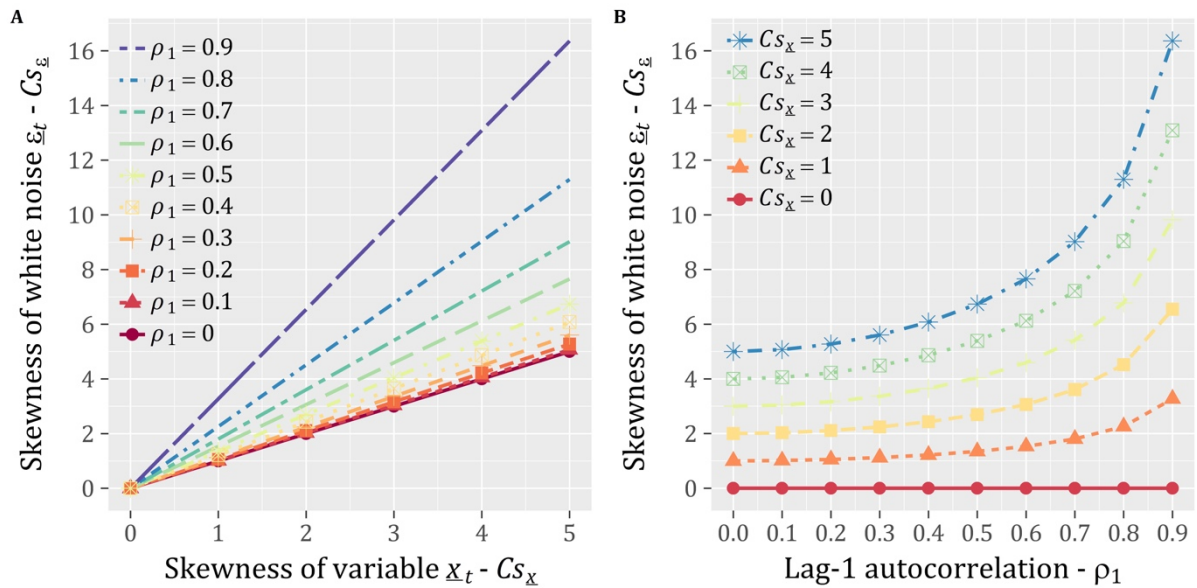
$$\sigma_{\underline{\varepsilon}}^2 = \sigma_{\underline{x}}^2 (1 - a_1^2) \quad (3.3)$$

Apparently, if the process of interest is Gaussian (or well-approximated by it), Equations (2) and (3) in combination with Gaussian white noise would be sufficient and *exact*, since a linear combination of Gaussian RVs is also Gaussian. However, this is not the case for most hydrometeorological processes. In this context, the TF approach attempts to approximate the non-Gaussian behavior of  $\underline{x}_t$  by employing non-Gaussian white noise for  $\underline{\varepsilon}_t$ , where the

skewness coefficient  $C_{S_{\underline{x}}} := E \left[ \left( \frac{\underline{x} - \mu_{\underline{x}}}{\sigma_{\underline{x}}} \right)^3 \right]$  of  $\underline{x}_t$  is related with that of  $\underline{\varepsilon}_t$  by [Fiering and Jackson, 1971; Kottegoda, 1980; Bras and Rodríguez-Iturbe, 1985; Reddy, 1997; Loucks and van Beek, 2017]:

$$C_{S_{\underline{\varepsilon}}} = C_{S_{\underline{x}}} \frac{(1 - a_1^3)}{(1 - a_1^2)^{3/2}} \quad (3.4)$$

Hence, in order to capture the first three marginal moments of  $\underline{x}_t$ , one has to generate non-Gaussian white noise with certain statistical characteristics, which are all functions of the lag-1 autocorrelation coefficient of the process  $\underline{x}_t$ , given that  $a_1 = \rho_1$ . In Figure 3.2 we provide two alternative views of Eq. (3.4), both depicting the variability of the required skewness  $C_{S_{\underline{\varepsilon}}}$  of the white noise against the skewness  $C_{S_{\underline{x}}}$  and the lag-1 autocorrelation  $\rho_1$  of the target process  $\underline{x}_t$ . Particularly, in Figure 3.2A we fix  $\rho_1$  to specific values, ranging from 0 to 0.9, and with  $C_{S_{\underline{x}}}$  varying from 0 to 5, while in Figure 3.2B we set  $C_{S_{\underline{x}}} \in \{1, 2, 3, 4, 5\}$  and vary  $\rho_1$  from 0 to 0.9. Considering a high  $\rho_1 = 0.9$  and aiming to reproduce a moderate skewness, e.g.,  $\approx 1$ , results in a white noise skewness  $\approx 3.5$ , while for a highly skewed variable the deviation becomes much larger (related of course to  $\rho_1$ ). For example, for a process with  $\rho_1 = 0.9$  and  $C_{S_{\underline{x}}} = 4$ , the required white noise skewness is  $C_{S_{\underline{\varepsilon}}} \approx 12.5$ , i.e., more than three times higher than the target value.



**Figure 3.2 |** Relationship between (A) the target skewness coefficient of process  $\underline{x}_t$  and the required skewness for white noise term  $\underline{\varepsilon}_t$  for a given lag-1 autocorrelation coefficient  $\rho_1$ ; and (B) the lag-1 autocorrelation coefficient  $\rho_1$  and the required skewness coefficient of white noise term  $\underline{\varepsilon}_t$  to attain the target skewness coefficient of process  $\underline{x}_t$ .

Within non-Gaussian simulations, the selection of the underlying statistical model of the white noise and the associated random number generation procedure is a pivotal step. Thomas and Fiering proposed the use of Pearson type-III ( $\mathcal{P}_{III}$ ) distribution, which is also one of the most commonly used distributions in hydrology. The probability density function (PDF) of  $\mathcal{P}_{III}$  is given by:

$$= \frac{1}{|b| \Gamma(a)} \left( \frac{\xi - c}{b} \right)^{a-1} \exp \left( -\frac{\xi - c}{b} \right), \begin{cases} \text{if } b > 0 & c \leq \xi < \infty \\ \text{if } b < 0 & -\infty < \xi \leq c \end{cases} \quad (3.5)$$

where  $a > 0$ ,  $b \neq 0$ , and  $c \in \mathbb{R}$  are shape, scale, and location parameters, respectively (if  $c = 0$ , then  $\mathcal{P}_{III}$  reduces to the Gamma distribution). The mean  $\mu_{\underline{\xi}}$ , variance  $\sigma_{\underline{\xi}}^2$  and skewness  $C_{S_{\underline{\xi}}}$  of the random variable  $\underline{\xi}$  are given by:

$$\mu_{\underline{\xi}} = c + ab, \quad \sigma_{\underline{\xi}}^2 = ab^2, \quad C_{S_{\underline{\xi}}} = \frac{2b}{|b|\sqrt{a}} \quad (3.6)$$

Of course, as [Matalas and Wallis \[1976\]](#) noted, choosing the white noise distribution is a matter of convenience (see also discussion in [Tsoukalas et al. \[2018e\]](#)) and simplicity in generating random numbers, given of course that the selected distribution can reproduce the desired statistics of white noise, i.e.,  $\mu_{\underline{\varepsilon}}$ ,  $\sigma_{\underline{\varepsilon}}$ , and  $C_{S_{\underline{\varepsilon}}}$ .

The non-Gaussian formulation of the AR(1) model through the TF approach results in the so-called envelope behavior issue, which is associated with the distribution of the white noise. Let us write Eq. (3.1) in the equivalent form:

$$\underline{x}_t = a_1 \underline{x}_{t-1} + F_{\underline{\varepsilon}}^{-1}(\underline{u}) \quad (3.7)$$

where  $F_{\underline{\varepsilon}}^{-1}$  denotes the inverse cumulative density function (ICDF) of the white noise  $\underline{\varepsilon}_t$  and  $\underline{u}$  expresses a uniform ( $\mathcal{U}$ ) RV in  $[0, 1]$  (probability), i.e.,  $\underline{u} \sim \mathcal{U}(0,1)$ . In the above formulation, we see that in the Gaussian case, where  $\underline{\varepsilon}_t \in (-\infty, \infty)$ , the random variable  $\underline{x}_t$  takes any value in  $(-\infty, \infty)$ . However, when the distribution of  $\underline{\varepsilon}_t$  has a finite left support, as in  $\mathcal{P}_{III}$  or Gamma ( $\mathcal{G}$ ) cases, then  $\lim_{u \rightarrow 0} F_{\underline{\varepsilon}}^{-1}(u) = \ell_{\underline{\varepsilon}}$ , where  $\ell_{\underline{\varepsilon}}$  stands for the lower bound of  $\underline{\varepsilon}_t$ . Hence, for given  $a_1$  (e.g., specified from the data) and  $x_{t-1}$ , we can estimate at any step of the generation procedure the lower bound of  $\underline{x}_t$  by:

$$x_t \geq a_1 x_{t-1} + \ell_{\underline{\varepsilon}} \quad (3.8)$$

and thus calculate the theoretical lower bound of the synthetic data. Similarly, when the distribution of  $\underline{\varepsilon}_t$  is bounded from above (as in the  $\mathcal{P}_{III}$  case adjusted for negative skewness), then  $\lim_{u \rightarrow 1} F_{\underline{\varepsilon}}^{-1}(u) = \nu_{\underline{\varepsilon}}$ , where  $\nu_{\underline{\varepsilon}}$  is the upper bound of the distribution of  $\underline{\varepsilon}_t$ . In this case the generation mechanism is bounded from above, i.e.:

$$x_t \leq a_1 x_{t-1} + \nu_{\underline{\varepsilon}} \quad (3.9)$$

This limitation is especially important since hydrometeorological variables, such as river discharge, cannot be considered unbounded from above, even when the sample statistics erroneously indicate negative skewness. To the best of our knowledge, despite the popularity of the TF model and the associated approach of coupling it with skewed white noise, this shortcoming has never been reported in literature, regardless of its straightforward and intuitive theoretical derivation. This limitation also holds for the univariate cyclostationary TF model (i.e., PAR(1) with  $\mathcal{P}_{III}$  white noise), for which we provide the corresponding relationships in Appendix A.

Apart from the above relationships, based on the previous formulation it can be shown that a simple recursive procedure facilitates the estimation of the theoretical minimum (or maximum) value of the TF approach. Without the loss of generality, assuming  $x_0 := \mu_{\underline{x}}$ , and by sequentially applying Eq. (3.7) for  $q$  steps with  $\varepsilon_t = F_{\underline{\varepsilon}}^{-1}(0) = \ell_{\underline{\varepsilon}}$ , we can obtain the model's theoretical minimum, which can differ from  $\ell_{\underline{\varepsilon}}$  (they are identical when  $\ell_{\underline{\varepsilon}} = 0$ ). The recursive procedure can be written as follows:

$$\begin{aligned}
 x_0 &:= \mu_{\underline{x}} \\
 x_1 &= a_1 x_0 + F_{\underline{\varepsilon}}^{-1}(0) \\
 x_2 &= a_1 x_1 + F_{\underline{\varepsilon}}^{-1}(0) \\
 &\vdots \\
 x_q &= a_1 x_{q-1} + F_{\underline{\varepsilon}}^{-1}(0)
 \end{aligned} \tag{3.10}$$

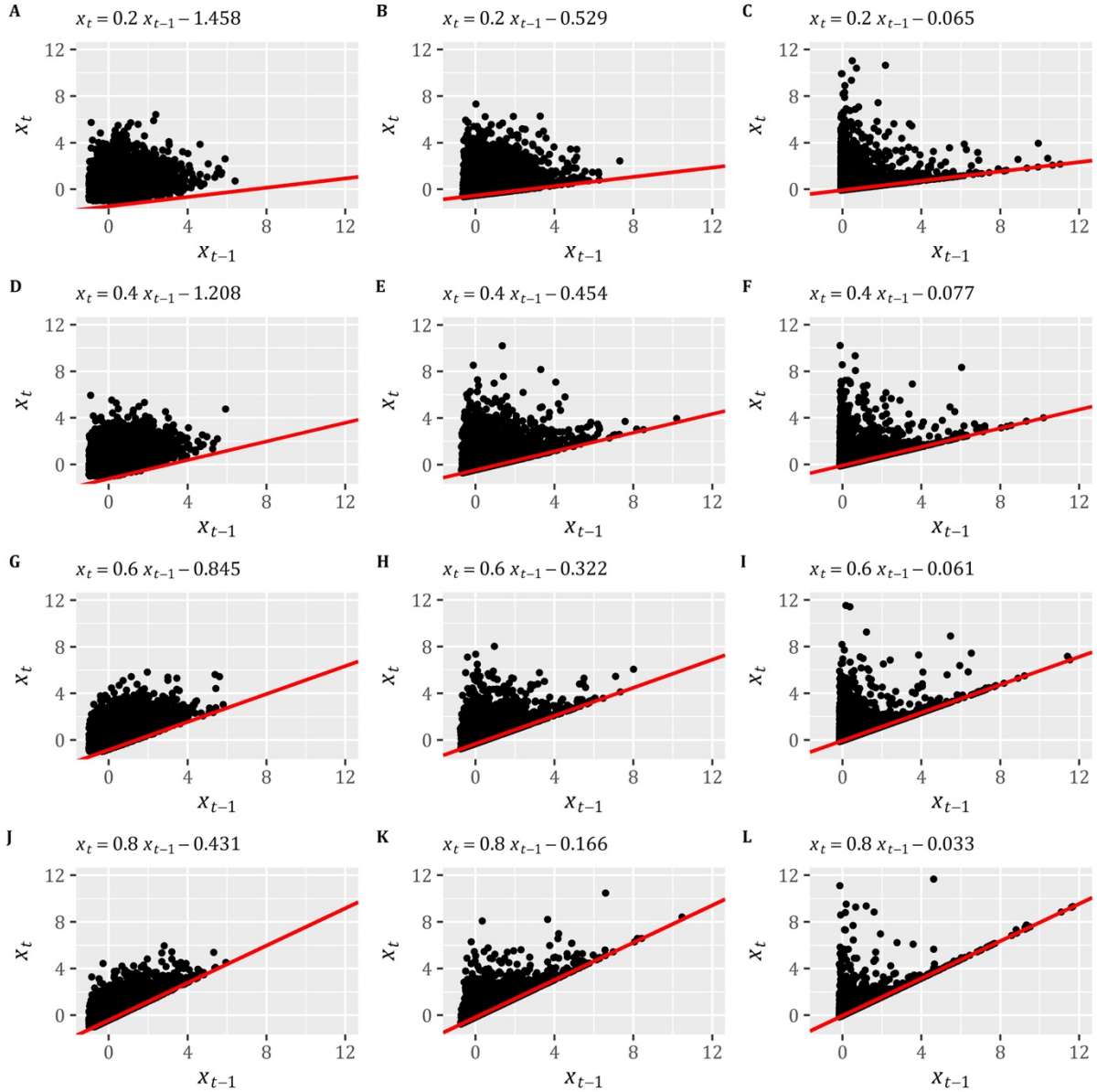
Alternatively, and more vigorously (depending on the support of  $\varepsilon_t$ ), the theoretical minimum and maximum are given, respectively, by  $\min(x_t) = \ell_{\underline{\varepsilon}}/(1 - a_1)$  and  $\max(x_t) = v_{\underline{\varepsilon}}/(1 - a_1)$ .

In order to better demonstrate the envelope behavior, we apply the AR(1) model coupled with PIII white noise (termed AR(1)-PIII) to 12 hypothetical scenarios by simulating 5 000 time steps for each. For all scenarios we fix  $\mu_{\underline{x}} = 0.5$  and  $\sigma_{\underline{x}}^2 = 1$  combined with  $C_{s_{\underline{x}}} \in \{1, 2, 4\}$  and  $\rho_1 \in \{0.2, 0.4, 0.6, 0.8\}$  (see Table 3-1 for a summary). Since the PIII is used for generating white noise and  $C_{s_{\underline{x}}} > 0$ , in all cases a lower bound is anticipated.

**Table 3-1** | Summary of target statistics for all scenarios (in all cases,  $\mu_{\underline{x}} = 0.5$  and  $\sigma_{\underline{x}}^2 = 1$ ).

| Scenario                             | A        | B     | C     | D     | E     | F     | G     | H     | I     | J     | K     | L     |       |
|--------------------------------------|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $C_{s_{\underline{x}}}$              | 1        | 2     | 4     | 1     | 2     | 4     | 1     | 2     | 4     | 1     | 2     | 4     |       |
| $\rho_1 = a_1$                       |          | 0.2   |       |       | 0.4   |       |       | 0.6   |       |       | 0.8   |       |       |
| $\mu_{\underline{\varepsilon}}$      |          | 0.4   |       |       | 0.3   |       |       | 0.2   |       |       | 0.1   |       |       |
| $\sigma_{\underline{\varepsilon}}^2$ |          | 0.96  |       |       | 0.84  |       |       | 0.64  |       |       | 0.36  |       |       |
| $C_{s_{\underline{\varepsilon}}}$    | 1.05     | 2.11  | 4.22  | 1.22  | 2.43  | 4.86  | 1.53  | 3.06  | 6.13  | 2.26  | 4.52  | 9.04  |       |
| PIII distributio<br>n                | <i>a</i> | 3.596 | 0.899 | 0.225 | 2.706 | 0.677 | 0.169 | 1.706 | 0.426 | 0.107 | 0.784 | 0.196 | 0.049 |
|                                      | <i>b</i> | 0.517 | 1.033 | 2.067 | 0.557 | 1.114 | 2.229 | 0.613 | 1.225 | 2.450 | 0.678 | 1.356 | 2.711 |
|                                      | <i>c</i> | -1.45 | -0.52 | -0.06 | -1.20 | -0.45 | -0.07 | -0.84 | -0.32 | -0.06 | -0.43 | -0.16 | -0.03 |
|                                      | 8        | 9     | 5     | 8     | 4     | 7     | 5     | 2     | 1     | 1     | 6     | 3     |       |

As theoretically expected, the model reproduces the target ACF and target statistics for all scenarios with high accuracy (see Figure A.1 and Table A-1 of Appendix A). However, the envelope behavior of the dependence pattern is apparent and indicates its limitation, a fact demonstrated by the scatter plots (Figure 3.3) corresponding to the 12 simulation scenarios. The theoretically-derived Eq. (3.8), defining the lower bound of the feasible space of the  $(x_{t-1}, x_t)$  points, is depicted by a red line (Figure 3.3). Note that labels in each subplot follow the scenarios' naming convention in Table 3-1 (e.g., panel C corresponds to scenario C of Figure 3.3). Apparently, in every case, regardless of the  $C_{s_{\underline{x}}}$  and  $\rho_1$  values, the model generates bounded dependence patterns enveloped by Eq. (3.8). This behavior appears even for low combinations of  $C_{s_{\underline{x}}}$  and  $\rho_1$  (e.g., scenario A).



**Figure 3.3** | Scatter plots depicting the simulated (using the TF model, i.e., the autoregressive model of order 1 (AR(1))-PIII) lag-1 dependence pattern among consecutive time steps (i.e., pair values (•) of the previous and current time steps). The labels of each plot resemble the corresponding scenarios of **Table 3-1**. The red line (—) depicts the envelope equation shown in the title of each plot.

### 3.2.3 From the univariate to the multivariate AR(1) model

It is reasonable to expect that the envelope behavior will also be observed in the multivariate case, i.e., when the multivariate autoregressive process of order 1 is used (MAR(1)) in combination with non-Gaussian white noise. Let us assume that we wish to generate an  $m$ -dimensional vector  $\underline{x}_t = [x_t^1, \dots, x_t^i, \dots, x_t^m]^T$  of  $m$  cross-correlated AR(1) processes, indexed by  $i$ . The generation mechanism of the model is:

$$\underline{x}_t = \mathbf{A}_1 \underline{x}_{t-1} + \underline{\varepsilon}_t \quad (3.11)$$

where  $\mathbf{A}_1$  is an  $m \times m$  matrix and  $\underline{\varepsilon}_t$  is an  $m$ -dimensional column vector of cross-correlated yet serially independent RVs with covariance  $\Sigma_{\varepsilon} \in \mathbb{R}^{m,m}$ . The model is often called the



*multivariate lag-1 model* when a full  $\mathbf{A}_1$  matrix is employed, while there exists a variation that assumes a diagonal  $\mathbf{A}_1$  matrix, often called *multivariate Markov model* or *contemporaneous multivariate autoregressive model of order 1* (i.e., CMAR(1)). Both formulations explicitly account for the lag-0 cross-correlations of the variables while their major difference is that the former is able to explicitly account for the lag-1 cross-correlations [Pegram and James, 1972; Matalas and Wallis, 1976; Kottegoda, 1980]. On the other hand, the use of diagonal  $\mathbf{A}_1$  ensures that each individual process is a Markov process and significantly simplifies the parameter estimation procedure, since the lag-1 cross-correlations are not explicitly modeled. Its use is often advocated by the literature, since several authors suggest that lag-1 cross-correlations can be neglected [e.g., Pegram and James, 1972; Camacho et al., 1985; Salas, 1993; Koutsoyiannis and Manetas, 1996; Efstratiadis et al., 2014a; Tsoukalas et al., 2018e]. Yet it is noted that while this simplification could be valid for processes considered at a coarse time scale (e.g., monthly or annual), it should be used with caution in cases of fine time scale processes (e.g., hourly) or for modelling phenomena characterized by cause-effect relationships (e.g., rainfall-runoff). Nevertheless, here we focus on the so-called multivariate Markov model (i.e., CMAR(1)). Regarding its parameter estimation and assuming that  $\mathbf{A}_1$  is a diagonal matrix of the form:

$$\mathbf{A}_1 = \begin{bmatrix} a_{1[1,1]} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & a_{1[m,m]} \end{bmatrix} = [\mathbf{A}_1]_{i,j} \quad (3.12)$$

where  $a_{1[i,i]} = \text{Cov}[\underline{x}_t^i, \underline{x}_{t-1}^i] / \text{Var}[\underline{x}_{t-1}^i] = \text{Corr}[\underline{x}_t^i, \underline{x}_{t-1}^i] = \rho_1^i$ , the following holds true:

$$\boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}} = \mathbf{M}_0 - \mathbf{A}_1 \mathbf{M}_0 \mathbf{A}_1^T \quad (3.13)$$

where  $\mathbf{M}_0 = \text{Cov}[\underline{x}_t, \underline{x}_t] \in \mathbb{R}^{m,m}$  is the lag-0 cross-covariance matrix. For instance, its  $i^{\text{th}}$ ,  $j^{\text{th}}$  element is  $[\mathbf{M}_0]_{i,j} = \text{Cov}[\underline{x}_t^i, \underline{x}_t^j]$ . The theoretical cross-covariance matrices  $\mathbf{M}_\tau = \text{Cov}[\underline{x}_t, \underline{x}_{t-\tau}]$  can be obtained for any time lag  $\tau$  by recursively applying the equation:

$$\mathbf{M}_\tau = \mathbf{A}_1 \mathbf{M}_{\tau-1}, \tau > 0 \quad (3.14)$$

Meanwhile, the theoretical and cross-correlation matrices  $\mathbf{R}_\tau = \text{Corr}[\underline{x}_t, \underline{x}_{t-\tau}]$  are obtained by  $\mathbf{R}_\tau = (\text{diag}(\mathbf{M}_\tau))^{-1/2} \mathbf{M}_\tau (\text{diag}(\mathbf{M}_\tau))^{-1/2}$ . Furthermore, the covariance matrix  $\boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}}$  can be expressed as:

$$\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}} \quad (3.15)$$

where  $\mathbf{B}$  is an  $m \times m$ , typically lower triangular, matrix (also known as the square root of  $\boldsymbol{\Sigma}_{\underline{\boldsymbol{\varepsilon}}}$ ) obtained by standard decomposition techniques (e.g., the Cholesky technique) or optimization approaches [Koutsoyiannis, 1999; Higham, 2002]. The latter methods are usually employed when  $\mathbf{B}$  is non-positive definite. Typically, such problems arise when the sample estimates of the required statistics are extracted from historical time series of different and/or limited lengths [Kottegoda, 1980]. Nonetheless, given that  $\mathbf{A}_1$  is diagonal and assuming that  $\underline{\boldsymbol{\varepsilon}}_t = \mathbf{B}\underline{\boldsymbol{\xi}}_t$ , where  $\underline{\boldsymbol{\xi}}_t$  is an  $m$ -dimensional column-vector of i.i.d. RVs, the decomposition of Eq. (3.11) can be given as follows:

$$\underline{x}_t^i = a_{1[i,i]} \underline{x}_{t-1}^i + \sum_{j=1}^m b_{[i,j]} \underline{\xi}_t^j \quad (3.16)$$

Additionally, the moments of  $\underline{\xi}_t$  and  $\underline{x}_t$  are interrelated through (index  $t$  is omitted due to stationarity):

$$\mu_{\underline{\xi}} = E[\underline{\xi}] = \mathbf{B}^{-1}\{E[\underline{x}] - \mathbf{A}_1 E[\underline{x}]\} \quad (3.17)$$

$$\sigma_{\underline{\xi}}^2 = \text{Var}[\underline{\xi}] = [1, \dots, 1]^T \quad (3.18)$$

$$C_{S_{\underline{\xi}}} = \mu_3[\underline{\xi}] = (\mathbf{B}^{(3)})^{-1} \{\mu_3[\underline{x}] - \mathbf{A}_1^{(3)} \mu_3[\underline{x}]\} \quad (3.19)$$

where  $\mu_3[\underline{\xi}]$  and  $\mu_3[\underline{x}]$  denote column-vectors that contain the third central moments of  $\underline{\xi}$  and  $\underline{x}$ , respectively; we remark that  $\underline{\xi}$  coincides with the skewness coefficient, since the model assumes unit variance for  $\underline{\xi}$ . Similar to the univariate case, the white noise term is typically generated using the  $\mathcal{P}$ III distribution (Eq. (3.5)). To illustrate the envelope behavior of the model, we rewrite the Eq. (3.11) similarly to Eq. (3.7), i.e.:

$$\underline{x}_t^i = a_{1[i,i]} \underline{x}_{t-1}^i + \sum_{j=1}^m b_{[i,j]} F_{\underline{\xi}^j}^{-1}(u^j) \quad (3.20)$$

where  $F_{\underline{\xi}^j}^{-1}(u^j)$  denotes the quantile function of  $\underline{\xi}^j$  for a given probability  $u^j$ . If the distribution of  $\underline{\xi}^j$  is bounded below or above by  $\ell_{\underline{\xi}^j}$  or  $\nu_{\underline{\xi}^j}$ , respectively, then  $\lim_{u^j \rightarrow 0} F_{\underline{\xi}^j}^{-1}(u^j) = \ell_{\underline{\xi}^j}$ , and  $\lim_{u^j \rightarrow 1} F_{\underline{\xi}^j}^{-1}(u^j) = \nu_{\underline{\xi}^j}$ . Therefore, we obtain:

$$\underline{x}_t^i \geq a_{1[i,i]} \underline{x}_{t-1}^i + \sum_{j=1}^m b_{[i,j]} \ell_{\underline{\xi}^j} \quad (3.21)$$

$$\underline{x}_t^i \leq a_{1[i,i]} \underline{x}_{t-1}^i + \sum_{j=1}^m b_{[i,j]} \nu_{\underline{\xi}^j} \quad (3.22)$$

or lower (left)- and above (right)-bounded cases, respectively.

However, in the multivariate case, and since  $\underline{x}_t^i$  depends on multiple values of  $\underline{\xi}^j$ , the limiting behavior (assuming that all RVs are left-bounded) is identified by setting  $\mathbf{u} = [u^1, \dots, u^i, \dots, u^m] \rightarrow \mathbf{0}$ . Of course, the envelope behavior diminishes if the white noise term  $\underline{\xi}_t$  is normally distributed (or more generally if  $\underline{\xi}_t \in (-\infty, \infty)$ ), yet in this case skewness cannot be preserved. Without the loss of generality, we examine the bivariate case of  $\underline{x}_t = [\underline{x}_t^1, \underline{x}_t^2]^T$  where both processes exhibit zero autocorrelation but their lag-0 cross-correlation is equal to 0.8. For  $E[\underline{x}] = [0.5, 0.5]^T$ ,  $\text{Var}[\underline{x}] = [1, 1]^T$ , and  $\mu_3[\underline{x}] = C_{S_{\underline{x}}} = [2, 2.5]^T$  we find:

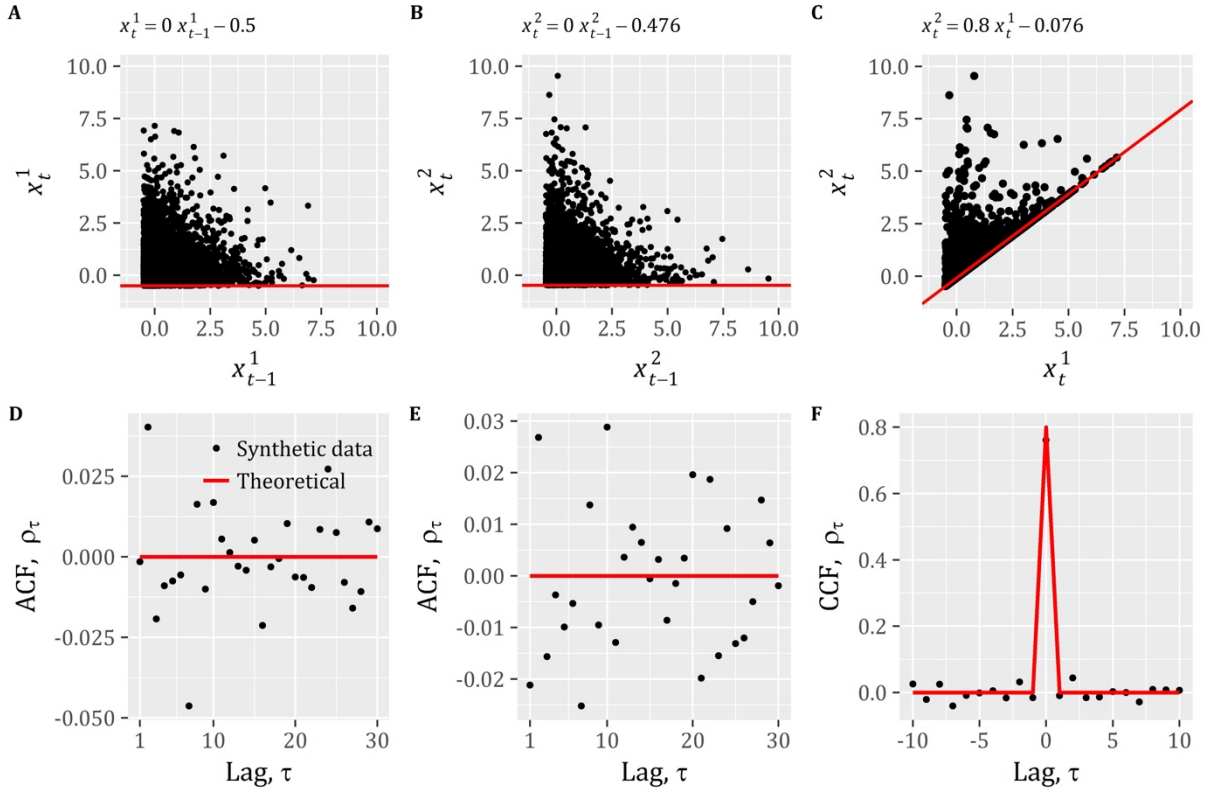
$$\mathbf{A}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0.8 & 0.6 \end{bmatrix} \quad (3.23)$$

where  $\mathbf{B}$  is obtained by the Cholesky decomposition), while the generating equation (Eq. (3.11)) becomes:

$$\begin{bmatrix} \underline{x}_t^1 \\ \underline{x}_t^2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \underline{x}_{t-1}^1 \\ \underline{x}_{t-1}^2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0.8 & 0.6 \end{bmatrix} \begin{bmatrix} \underline{\xi}_t^1 \\ \underline{\xi}_t^2 \end{bmatrix} \quad (3.24)$$

Given the target moments of  $\underline{x}$ , the statistics of the white noise term are calculated as  $E[\underline{\xi}] = [0.50, 0.16]^T$ ,  $\text{Var}[\underline{\xi}] = [1, 1]^T$ , and  $\mu_3[\underline{\xi}] = C_{s_\xi} = [2.00, 6.83]^T$ . Using the PIII for white noise generation we obtain the lower bound vector  $\underline{\boldsymbol{\rho}}_\xi = [-0.500, -0.126]^T$ . Thus, from Eq. (3.22) the limiting envelope equations are  $\underline{x}_t^1 = 0 \underline{x}_{t-1}^1 - 0.500$  and  $\underline{x}_t^2 = 0 \underline{x}_{t-1}^2 - 0.475$ . In this case, it is also possible to estimate the envelope relationship a priori between  $\underline{x}_t^1$  and  $\underline{x}_t^2$  as  $\mathbf{A}_1$  is a zero matrix. Particularly, since  $\underline{x}_t^1 = \underline{\xi}_t^1$  and  $\underline{x}_t^2 = 0.8\underline{\xi}_t^1 + 0.6\underline{\xi}_t^2$ , and substituting the former into the latter, we get  $\underline{x}_t^2 = 0.8\underline{x}_t^1 - 0.6\underline{\xi}_t^2$ , and by setting  $\underline{\xi}_t^2 = \underline{\ell}_{\xi^2}$  the envelope line  $\underline{x}_t^2 = 0.8\underline{x}_t^1 - 0.076$  is obtained.

In order to demonstrate the envelope behavior in the multivariate case, we employ the above model and synthesized a time series of 5 000 time steps. **Figure 3.4**A–C depicts the established dependence patterns of each individual process for lag-1 (panels A and B), while panel C shows the pattern among the two processes for lag-0. Also, at each panel we display the corresponding envelope equation. We remark that the model was able to accurately reproduce the theoretical stochastic structure, expressed by the autocorrelation (ACF) and cross-correlation functions (CCF) shown in **Figure 3.4**D–F, as well as, to approximate very well the target moments (**Table A-2**).



**Figure 3.4** | Scatter plots depicting the simulated (using the contemporaneous multivariate autoregressive model of order 1 (CMAR(1) model) with  $\mathcal{P}$ III white noise) for (A) and (B) lag-1 dependence patterns of the zero-autocorrelated processes  $\underline{x}_t^1$  and  $\underline{x}_t^2$ , respectively, for consecutive time steps (i.e., pair values ( $\bullet$ ) of the previous and current time steps). Panel (C) depicts the contemporaneous dependence (lag-0) of  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . The red line ( $-$ ) depicts the envelope equation shown in the title of each plot. Panel (D) compares the simulated and theoretical autocorrelation function (ACF) of  $\underline{x}_t^1$  while panel (E) compares that of  $\underline{x}_t^2$ . Finally, panel (F) compares the simulated and theoretical cross-correlation function (CCF) of  $\underline{x}_t^1$  and  $\underline{x}_t^2$ .

In order to extend our working examples, we simulate another vector of bivariate time series (5 000 time steps) using the same marginal moments as before, but this time with a different autocorrelation structure. Specifically, we assumed  $\text{Corr}[\underline{x}_t^1, \underline{x}_{t-1}^1] = \rho_1^1 = 0.7$  and  $\text{Corr}[\underline{x}_t^2, \underline{x}_{t-1}^2] = \rho_1^2 = 0.5$ . Thus, we get:

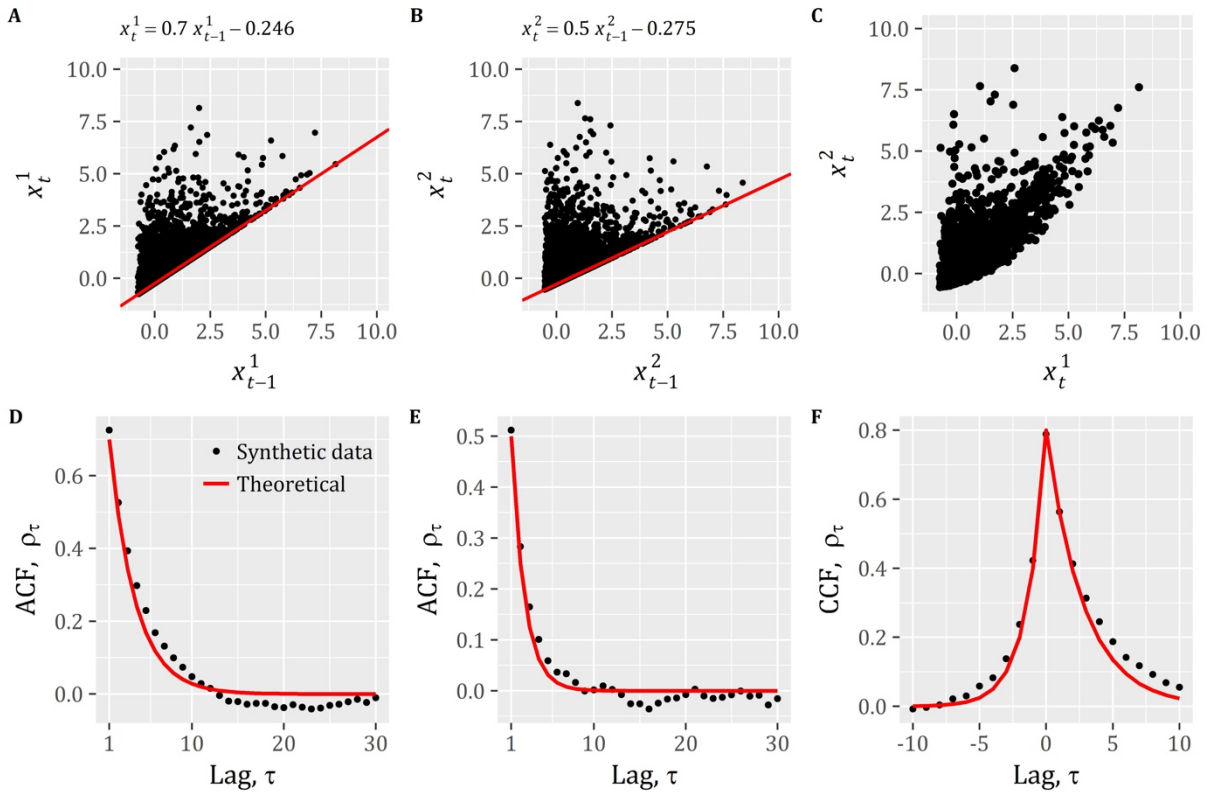
$$\mathbf{A}_1 = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.5 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0.714 & 0 \\ 0.728 & 0.468 \end{bmatrix} \quad (3.25)$$

and the generating formula (i.e., Eq. (3.11)):

$$\begin{bmatrix} \underline{x}_t^1 \\ \underline{x}_t^2 \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} \underline{x}_{t-1}^1 \\ \underline{x}_{t-1}^2 \end{bmatrix} + \begin{bmatrix} 0.714 & 0 \\ 0.728 & 0.468 \end{bmatrix} \begin{bmatrix} \xi_t^1 \\ \xi_t^2 \end{bmatrix} \quad (3.26)$$

Similar to the previous analysis, **Figure 3.5**A–C depicts the established lag-1 and lag-0 dependence patterns, while the envelope equation of each process is displayed in the title of each panel. It is apparent that at each simulated step, the model poses significant constraints in the range of subsequent plausible values, which is far from reality. We remark that in this case the contemporaneous lag-0 relationship cannot be displayed in a two-dimensional (2D) plot since it involves the lag-1 values of  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . Nevertheless, the model successfully reproduced

the target stochastic structure (**Figure 3.5D–F**) and the marginal moments (see **Table A-3**), at the cost, however, of unrealistically bounded dependence patterns.



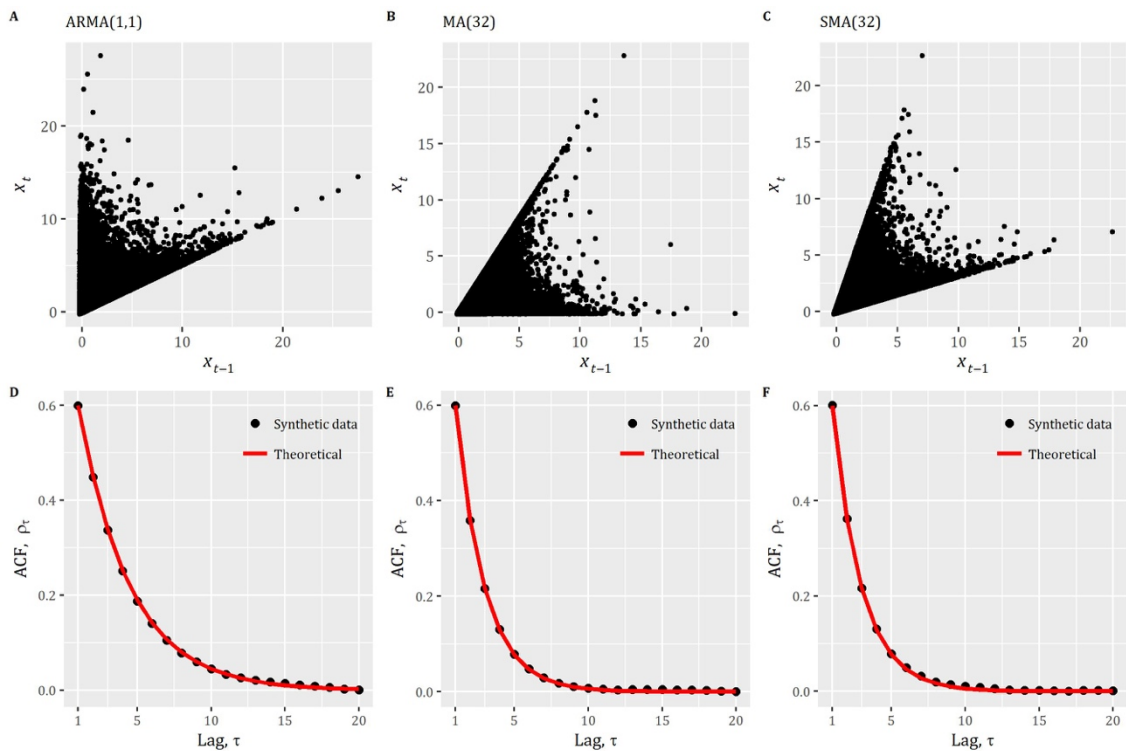
**Figure 3.5** | Scatter plots depicting the simulated (using the CMAR(1) model with  $\mathcal{P}$ III white noise) for (A) and (B) lag-1 dependence pattern of the autocorrelated processes  $x_t^1$  and  $x_t^2$ , respectively, for consecutive time steps (i.e., pair values ( $\bullet$ ) of the previous and current time steps), while panel (C) depicts the contemporaneous dependence (lag-0) of  $x_t^1$  and  $x_t^2$ . The red line ( $-$ ) depicts the envelope equation shown in the title of each plot. Panel (D) compares the simulated and theoretical ACF of  $x_t^1$  while panel (E) compares that of  $x_t^2$ . Lastly, panel (F) compares the simulated and theoretical CCF of  $x_t^1$  and  $x_t^2$ .

### 3.2.4 The envelope behavior beyond AR models

To demonstrate the impact of employing non-Gaussian white noise in combination with other linear stochastic models, we employed (1) a low-order autoregressive moving average model  $ARMA(p,q)$ ; (2) a simple moving average model  $MA(q)$ ; and (3) its symmetric variant, termed  $SMA(q)$ . The parameters  $p$  and  $q$  determine the order of the models. As shown by *O’Connell [1974]* and later discussed by *Lettenmaier and Burges [1977]*, it is possible for the case of  $ARMA(1,1)$  to derive an analytical relationship that links the skewness of the white noise with the skewness of the target process. Furthermore, it has been shown [*Koutsoyiannis, 2000*] that similar relationships can be established for the two moving average schemes regardless of the order  $q$  (i.e.,  $MA(q)$  and  $SMA(q)$ ).

In this demonstration we utilized the aforementioned relationships for the simulation of a univariate stationary process with the characteristics of the hypothetical Scenario I of **Table 3-1**, which refers to the Markovian process with  $\rho_\tau = 0.6^{|\tau|}$  and  $C_{S_x} = 4$ . Regarding the  $ARMA(1,1)$  process, it is noted that its autocorrelation structure is somewhat different from the Markovian one, hence we carefully choose its parameters in order to have  $\rho_1 = 0.6$ . On the

other hand, both  $MA(q)$  and  $SMA(q)$  are able to resemble the Markovian autocorrelation structure with satisfactory accuracy by setting  $q = 32$ . Nonetheless, in all cases we used  $\mathcal{P}III$  distribution for the white noise, hence the models are termed  $ARMA(1,1)-\mathcal{P}III$ ,  $MA(32)-\mathcal{P}III$ , and  $SMA(32)-\mathcal{P}III$ . However, due to a lack of analytical solution for the envelope function, and in order to derive a clear picture of the established dependence patterns, we generated very long realizations, each one consisting of 500 000 time steps. **Figure 3.6** shows the lag-1 dependence pattern obtained by the three schemes as well as a comparison of the simulated and theoretical ACF, which is very well reproduced by all models. Despite the accurate reproduction of the target marginal statistics (mean, variance, and skewness) by all models, they establish peculiarly-shaped and always bounded dependence forms. However, it should be noted that the  $MA(q)$  and  $SMA(q)$  schemes are typically employed for the simulation of annual processes, which are typically well approximated by the Gaussian distribution, and thus it is reasonable to expect a minimization of this issue.

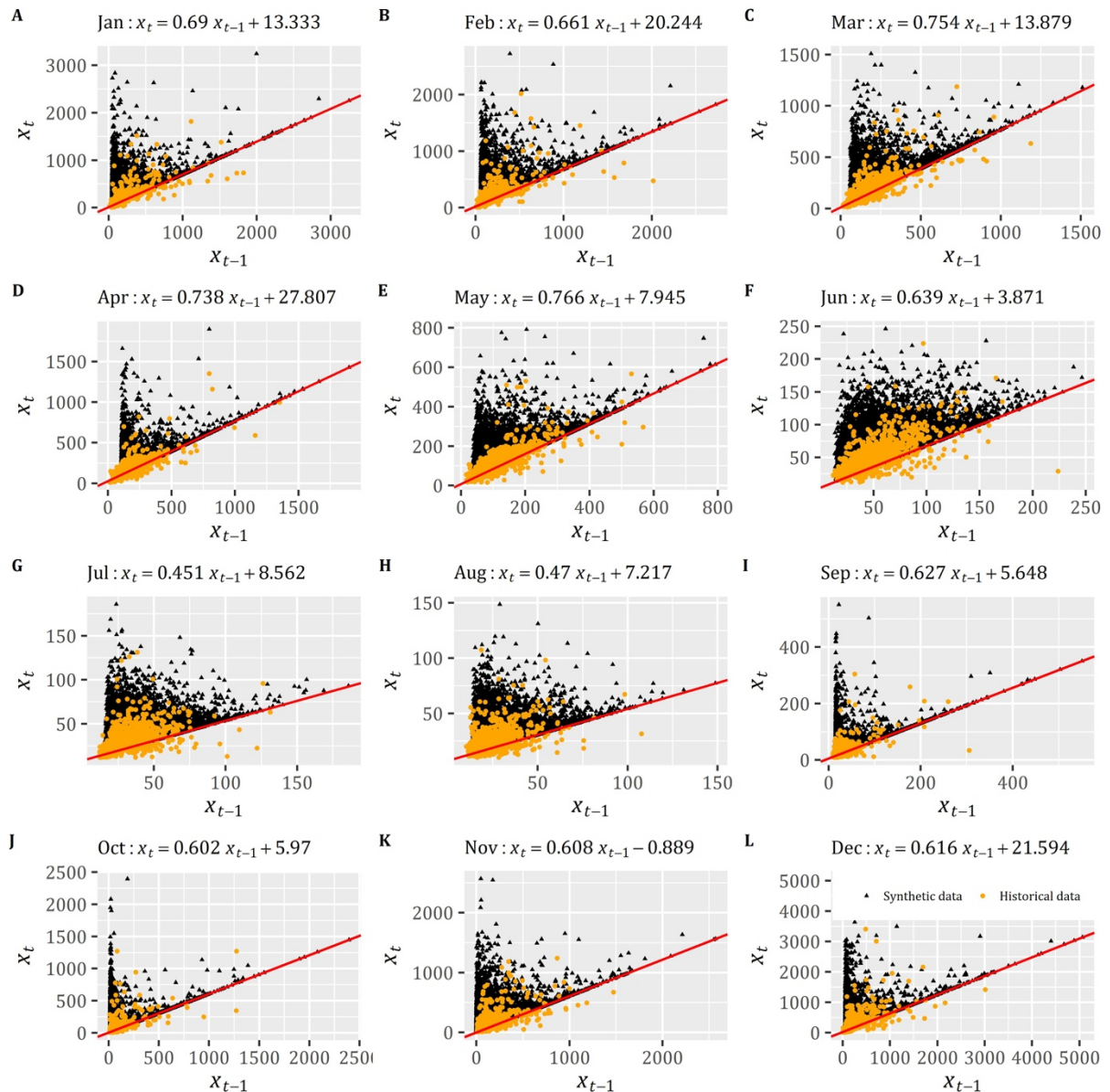


**Figure 3.6** | Scatter plots depicting the simulated lag-1 dependence pattern among consecutive time steps (i.e., pair values (•) of the previous and current time steps) obtained by: (A)  $ARMA(1,1)-\mathcal{P}III$ ; (B)  $MA(32)-\mathcal{P}III$ ; and (C)  $SMA(32)-\mathcal{P}III$  models. Comparison of synthetic and theoretical autocorrelation function (ACF) obtained by: (D)  $ARMA(1,1)-\mathcal{P}III$ ; (E)  $MA(32)-\mathcal{P}III$ ; and (F)  $SMA(32)-\mathcal{P}III$  models.

### 3.3 REAL-WORLD CASE STUDY

In this section we demonstrate the envelope behavior of the TF approach using a real-world and long dataset (1 January 1970 to 31 December 2008) of daily streamflow data ( $m^3/s$ ) of river Achelous at Kremasta dam in Western Greece. It is assumed that the autocorrelation structure of the daily streamflow of each month can be described by a stationary  $AR(1)$  model. The historical monthly and daily time series are characterized by non-Gaussian distributions and skewness coefficients ranging from 1.6 (June) up to 6.7 (October). Specifically, we generate daily synthetic time series with a length of 1000 years, using for each month a different  $AR(1)$  model with  $\mathcal{P}III$  white noise (i.e.,  $AR(1)-\mathcal{P}III$ ). The model very satisfactorily reproduced the

target historical marginal statistics of each month (Table A-4), as well as the theoretical Markovian autocorrelation structure (see Figure A.2 for a comparison among the empirical, synthetic, and theoretical ACFs), which however deviates from the empirical ACF for some months, showing a more persistent behavior. Yet a comparison of the lag-1 dependence patterns between the synthetic and the historical data, using scatter plots for each month (Figure 3.7), reveals the omnipresence of the envelope behavior. As shown, the model generates unrealistic dependence patterns that are far from the historical ones. The synthetic pairs of values are bounded by the theoretical envelope function (red line; embedded in each plot), while the historical pairs clearly extend beyond that line. In an effort to provide a quantitative metric, we calculate the empirical probability of a historical pair to lie below the envelope function. The overall mean value of this metric estimated from all months is 27%, while it ranges from 14% (in November) to 42% (in April).



**Figure 3.7** | Scatter plots showing the lag-1 dependence pattern of the daily streamflow ( $\text{m}^3/\text{s}$ ) of the Achelous river at the Kremasta dam, Greece (orange dots;  $\bullet$ ) and of a synthetic time series generated using an  $\text{AR}(1)$ - $\mathcal{P}$ III model (black dots;  $\bullet$ ). The red line ( $-$ ) depicts the envelope equation embedded each plot.

### 3.4 DISCUSSION

Historically, most of the questions raised regarding the TF approach have concerned the case of the AR(1) model and the range of attainable skewness coefficients [McMahon and Miller, 1971; Lettenmaier and Burges, 1977; Obeysekera and Yevjevich, 1985]. This was mainly due to the use of Wilson-Hilferty transformation which was used for generating Gamma or Pearson type-III RVs [Kirby, 1972]. Nowadays, this technical issue is out of interest, since such RVs can be easily generated with high accuracy by modern random number generators which are available in almost every programming language (e.g., R, MATLAB, etc.). Additionally, we note that McMahon and Miller [1971] reported that Thomas and Burden [1963] and Fiering [1967] tested their approach for skewness values ranging in  $(-0.5, 1.0)$ .

This work focused on the effect of using Pearson type-III white noise in AR(1) models and we show that this approach leads to unrealistic dependence patterns. Furthermore, preliminary investigations have also shown that this issue extends over other popular linear stochastic models when coupled with non-Gaussian white noise. Particularly, we demonstrated three such cases using PIII white noise in combination with (1) a classical ARMA(1,1); (2) a simple MA( $q$ ); and (3) its symmetrical variant SMA( $q$ ) [Koutsoyiannis, 2000]. In all cases the resulting dependence patterns exhibited a peculiar and unrealistic bounded shape which can be bounded from both directions.

Nevertheless, it is noteworthy that Song *et al.* [1996] and Jeong and Lee [2015] also observed this issue independently while studying AR(1) with exponential white noise [Gaver and Lewis, 1980; Lawrance and Lewis, 1981b, 1981a] and periodic Gamma autoregressive (PGAR) processes [Fernandez and Salas, 1986], respectively. However, to the best of our knowledge, these works, or any other, have not revealed the envelope limitation, neither provided a theoretical proof and a justification for this behavior, which probably arises from the lack of explicit assumption regarding the joint dependence structure of the process. Particularly, the joint moment of order  $k + n$  of two continuous RVs,  $\underline{x}$  and  $\underline{y}$ , is given by:

$$E[\underline{x}^k \underline{y}^n] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \underline{x}^k \underline{y}^n f_{\underline{xy}}(x, y) dx dy \quad (3.27)$$

where  $f_{\underline{xy}}$  denotes the joint probability distribution function (PDF) of  $\underline{x}$  and  $\underline{y}$ . The first cross product joint moment is embedded in the definition of covariance, as well as in the Pearson's correlation, i.e.:

$$\rho_{\underline{xy}} = \frac{E[\underline{x} \underline{y}] - E[\underline{x}] E[\underline{y}]}{\sqrt{\text{Var}[\underline{x}] \text{Var}[\underline{y}]}} \quad (3.28)$$

Hence, this points to the requirement for an assumption regarding  $f_{\underline{xy}}$ . When both  $\underline{x}$  and  $\underline{y}$  are Gaussian, and simulated through a typical decomposition scheme (e.g., the Cholesky technique) which applies linear operations on them, the joint PDF of  $\underline{x}$  and  $\underline{y}$  is also Gaussian (due to the affine transformation property of Gaussian RVs). When  $\underline{x}$  and  $\underline{y}$  are not Gaussian, this convenient property does not hold. By analogy, the joint moment of order  $k + n$  of a continuous-state, discrete-time stochastic process  $\underline{x}_t$  can be expressed as:



$$E[\underline{x}_t^k \underline{x}_{t-\tau}^n] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \underline{x}_t^k \underline{x}_{t-\tau}^n f_{\underline{x}_t, \underline{x}_{t-\tau}}(x_t, x_{t-\tau}) dx_t dx_{t-\tau} \quad (3.29)$$

If  $\underline{x}_t$  is Gaussian and modeled using a linear stochastic process (e.g., AR or MA-type) with Gaussian white noise, then it is well known that the joint PDF  $f_{\underline{x}_t, \underline{x}_{t-\tau}}$  is also Gaussian. This implies linear operations on Gaussian RVs. On the other hand, this does not hold for the TF approach, which uses non-Gaussian white noise and thus the form of  $f_{\underline{x}_t, \underline{x}_{t-\tau}}$  is unclear.

We remind that summary statistics such as mean, variance, skewness, and correlation are nothing more than some useful measures of location, dispersion, asymmetry, and dependence, and do not involve in their estimation the actual joint distribution. A classic example is provided by the so-called Anscombe's Quartet [Anscombe, 1973] and recently by Matejka and Fitzmaurice [2017]. Both works stress the importance of data science's first principle: *Visualize your data*. They demonstrate this issue by devising several examples of datasets that have the same summary statistics but completely different dependence patterns. Apparently, as shown in this work, the aforementioned simple principle also applies in synthetic hydrology.

Nowadays, multivariate random variables are typically modeled by copula functions [Sklar, 1959, 1973], which despite the early-days skepticism [Sklar, 1973] have found wide acceptance and practical use. In hydrology, copulas have been popularized by the studies of De Michele and Salvadori [2003] and Favre et al. [2004], and since then have been widely applied for the description of correlated yet time-independent variables [e.g., De Michele and Salvadori, 2003; Favre et al., 2004; Salvadori and De Michele, 2004, 2007; Zhang et al., 2006; Genest and Favre, 2007; Zhang and Singh, 2007; Hao and Singh, 2016; Wang et al., 2017], while only lately they have been adapted and modified accordingly to account for time-dependence, which led to the development of copula-based schemes for the simulation of hydrometeorological processes [e.g., Bárdossy and Pegram, 2009; Serinaldi, 2009a; Gyasi-Agyei, 2011; Lee and Salas, 2011; Hao and Singh, 2013; Chen et al., 2015; Lee, 2017].

A conceptually related, yet until recently unknown to the hydrological community, approach relies on the so-called Nataf joint distribution model [Nataf, 1962], which is associated to the well-known Gaussian copula [Sklar, 1973; Lebrun and Dutfoy, 2009]. Since their inception, Nataf-based models have been developed and applied for the simulation of univariate or multivariate autoregressive processes with arbitrary marginal distribution mainly within operations research, [e.g., Cario and Nelson, 1996; Biller and Nelson, 2003], and probabilistic engineering mechanics [e.g., Grigoriu, 1998; Deodatis and Micaletti, 2001], while recently they have been aligned with hydrological stochastics [Serinaldi and Lombardo, 2017; Tsoukalas et al., 2017a, 2018e, 2018d, 2018b; Papalexiou, 2018] in order to account for non-Gaussian processes, both univariate and multivariate, exhibiting intermittency, periodicity, and any-range dependence.

Apparently, both Nataf- and copula-based approaches can provide a remedy to the limitations of the TF approach, as well as explicitly account for non-Gaussianity, which is omnipresent within hydrometeorological processes [e.g., Kroll and Vogel, 2002; Koutsoyiannis, 2005c; McMahan et al., 2007; Bowers et al., 2012; Papalexiou and Koutsoyiannis, 2012, 2016; Blum et al., 2017]. We deem that Nataf-based models provide a convenient and more precise alternative given that they utilize (in an auxiliary or parent role) existing and well-known stochastic models which provide the basis for a straightforward and operational efficient generation scheme. It is also noted that the celebrated Log-Normal model of Matalas [1967], which incidentally can be

classified as a Nataf-based approach [Tsoukalas et al., 2018e, 2018d], does not exhibit the TF approach limitation and thus can provide a rather easy and consistent option for practitioners.

### 3.5 SUMMARY

To conclude, we bring back the aphorism and the question set by Box and Draper. Paraphrasing, we could say that indeed *since all models are wrong and TF is not an exception, the question is how wrong the TF approach has to be to not be useful*. A way to answer this question is through impact assessments of the envelope behavior in real-world applications, e.g., in important engineering studies (reservoir design and sizing, hydropower assessment, reliability-based studies, etc.), and of its effect on decision-making related to water resources management. Another question arising here is why should one use a model with known limitations and flaws (irrespective of whether these flaws have minor or major impacts on real-world applications) which reproduces unrealistic rainfall or streamflow patterns?

We recognize that the TF model and the associated approach was a major contribution that shaped stochastic hydrology, yet in practice linear stochastic models should be used cautiously when combined with non-Gaussian white noise, given the limitations shown in this Chapter. This approach preserves important summary statistics (i.e., mean, variance, and skewness) and correlations, yet for processes showing medium to high skewness values and/or correlations it will inevitably reproduce bounded and unrealistic dependence patterns that are then used in simulations.

In this context, after half a century of use of this model and approach, we would suggest that it is time to move to alternative methods which are consistent in generating realistic dependence structures as well as the marginal distribution itself. The theoretical background of such an approach will be discussed in the next Chapter 4, while Chapters 5 to 7, are build upon it, and, propose a series of novel stochastic simulation schemes, suitable for a variety of stochastic simulation problems (i.e., univariate or multivariate, stationary or cyclostationary, as well as multi-scale).

## NON-GAUSSIAN MODELS FOR UNCONDITIONAL, CONDITIONAL AND STOCHASTIC SIMULATION OF RANDOM VARIABLES AND PROCESSES \*

---

### PREAMBLE

This Chapter provides an overview of the theoretical foundations of the Nataf's joint distribution model, a construct closely related with the Gaussian copula, which in turn allows modelling, and simulation (unconditional and conditional) of non-Gaussian random variables and processes. The description and background of the model is initially introduced for the establishment of the multivariate joint distribution (section 4.1) of non-Gaussian random variables, while progressively is extended to conditional distributions (section 4.2) and stochastic processes (section 4.3). Furthermore, section 4.5, presents a simple and efficient algorithm, based on a hybrid Monte Carlo procedure that is used to approximate the so-called equivalent correlation coefficients, an important concept of any Nataf-based method. Section 4.6 discusses and highlights the similarity between the rationale of Nataf-based methods with other commonly used in hydrology approaches; which interestingly can be retrospectively classified as Nataf-based methods. Finally, section 4.7 summarizes the key points and findings of the Chapter.

---

\* Partially based on:

Tsoukalas, I., C. Makropoulos, and D. Koutsoyiannis (2018d), Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions, *Water Resour. Res.*, doi:10.1029/2017WR022462.

Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2018e), Stochastic Periodic Autoregressive to Anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions, *Water Resour. Res.*, 54(1), 161–185, doi:10.1002/2017WR021394.

Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2018b), Building a puzzle to solve a riddle: a new approach to multi-temporal stochastic simulation, *J. Hydrol.*, doi:(in review).

Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2017a), Stochastic simulation of periodic processes with arbitrary marginal distributions, in 15th International Conference on Environmental Science and Technology. CEST 2017., Rhodes, Greece.

## 4.1 ON THE NATAF JOINT DISTRIBUTION MODEL

### 4.1.1 Introduction and historical background

The problem of obtaining the joint distribution of random variables has long been discussed within the statistical community. The formal introduction of copulas [Sklar, 1959, 1973] was arguably a hallmark development with broad impact. Copulas have been developed to describe multivariate distribution functions using simpler mathematical objects (i.e., using univariate distributions and the so-called copula functions) [e.g., Fréchet, 1951; Féron, 1956; Dall’Aglio, 1959; Nataf, 1962; Mardia, 1970]. For a general discussion on copulas, see for instance, the works of Embrechts *et al.* [2003], Nelsen [2007] or Joe [2014].

Among them, Nataf [1962] proposed a quite simple, yet general solution by mapping a multivariate Gaussian distribution with a given correlation matrix to multivariate uniform variables, which in turn are mapped to the desired distributions via the corresponding inverse cumulative functions, hereafter called Nataf joint distribution model (NDM). The key challenge of NDM is to identify the *equivalent correlations* to be applied within the generation of random variables in the normal (Gaussian) domain, in order to attain the desired correlation in the actual (or else, real) domain. In their classical work, Liu and Der Kiureghian [1986] showed that the NDM is suitable for describing a wide range of correlation values. Later, Cario and Nelson [1997] developed a generalized procedure based on NDM and referred to as NORTA (NORmal To Anything), for the generation of correlated random vectors with arbitrary marginal distributions, including also, combinations of continuous and discrete random variables. NDM may be considered as a specific case of copulas [Sklar, 1973], and more specifically the Gaussian one. In fact, it is argued, that linear stochastics are *naturally* compatible with this copula, since both use the Pearson’s linear correlation as measure of dependence. Lebrun and Dutfoy [2009], in view of copula theory, provide an extensive and insightful discussion on the relation of NDM with the Gaussian copula, as well as provide an alternative formulation of the former in terms of Spearman’s  $r_s$  and Kendall’s  $\tau$  (under some, rather strict, assumptions – see also the discussion below).

Admittedly Cario and Nelson [1997] argued that the generality of their approach came at the cost of computational efficiency (i.e., computational time), since the estimation of equivalent correlations presupposed solving numerically a double integral in the infinite domain. However, continuous advances in computing make this issue less and less relevant.

### 4.1.2 Theoretical background

Let  $\underline{x} = [x_1, \dots, x_m]^T$  denote a vector of  $m$  cross-correlated (yet, time-independent) random variables (RVs), indexed using  $i$ , each one characterized by an arbitrarily specified marginal distribution function  $F_{x_i}(x) := P(x_i \leq x)$ , with finite variance; also referred to as cumulative distribution function (CDF). Let also  $f_{x_i}(x) := dF_{x_i}(x)/dx$  denote the corresponding univariate probability density function (PDF). Furthermore, let  $\mathbf{R} := \text{Corr}[\underline{x}, \underline{x}^T]$  denote their (target) correlation matrix ( $m \times m$ ).

Let also,  $\underline{z} = [z_1, \dots, z_m]^T$  be a vector characterized by a  $m$ -dimensional multivariate standard normal distribution, i.e.,  $\underline{z} \sim \mathcal{N}_m(\underline{\mu}, \tilde{\Sigma})$ , where  $\underline{\mu} := E[\underline{z}] = \mathbf{0}^T$  is the mean vector ( $m \times 1$ ) and  $\tilde{\Sigma} := \text{Cov}[\underline{z}, \underline{z}^T] = E[(\underline{z} - \underline{\mu})(\underline{z} - \underline{\mu})^T] = E[\underline{z} \underline{z}^T] - \underline{\mu} \underline{\mu}^T$  is the covariance matrix ( $m \times m$ ), which has to be positive semi-definite and in the case of multivariate standard normal distribution is synonymous with its correlation matrix,  $\tilde{\mathbf{R}} := \text{Corr}[\underline{z}, \underline{z}^T] = \tilde{\Sigma}$ . The multivariate

standard normal CDF,  $\mathcal{N}_m$  is denoted for simplicity as  $\Phi_m(\mathbf{z}; \tilde{\mathbf{R}})$ , while its multivariate PDF as  $\varphi_m(\mathbf{z}; \tilde{\mathbf{R}})$ . Notice that the mean, has been omitted for brevity. Apparently, each element of  $\underline{\mathbf{z}}$  is also characterized by standard normal distribution,  $\Phi(\cdot)$  with density  $\varphi(\cdot)$ , i.e.,  $\underline{z}_i \sim \mathcal{N}(0,1)$ .

The main idea of NDM lies into establishing the multivariate joint distribution  $F_{\underline{\mathbf{x}}}(\mathbf{x}) = F_{\underline{\mathbf{x}}}(x_1, \dots, x_m) = P(\underline{x}_1 \leq x_1, \dots, \underline{x}_m \leq x_m)$  of  $\underline{\mathbf{x}}$  through the joint CDF of  $\underline{\mathbf{z}}$ . Particularly, by expressing each element of  $\underline{\mathbf{z}}$  as,

$$\underline{z}_i = \Phi^{-1}\left(F_{\underline{x}_i}(x_i)\right) \quad (4.1)$$

where  $\Phi^{-1}(\cdot)$  denotes the quantile function, else known as inverse cumulative density function (ICDF), of the univariate standard normal distribution. It is straightforward to see that by employing the probability integral transformation to each marginal CDF we obtain  $\underline{u}_i := F_{\underline{x}_i}(x_i)$  which is a uniformly distributed RV in  $[0, 1]$  that denotes probability. See also, [Papoulis \[1991 p. 101\]](#). Nevertheless, through the rules of probability transformation, the joint distribution (CDF) of  $\underline{\mathbf{x}}$  can be written as,

$$F_{\underline{\mathbf{x}}}(x_1, \dots, x_m) = \Phi_m\left(\Phi^{-1}\left(F_{\underline{x}_1}(x_1)\right), \dots, \Phi^{-1}\left(F_{\underline{x}_m}(x_m)\right); \tilde{\mathbf{R}}\right) \quad (4.2)$$

It is interesting to note that Eq. (4.2) is identical with the Gaussian copula. In brief copulas, denoted with  $C(\cdot)$ , are  $m$ -dimensional distribution functions on  $[0, 1]^m$  with uniform marginal distributions. [Sklar \[1959\]](#), established the theory of copulas and provided their general properties. Among them, it has been shown that any multivariate joint distribution can be regarded as a copula function. Particularly, Sklar's theorem states that a multivariate distribution  $F_{\underline{\mathbf{x}}}(\mathbf{x}) = F_{\underline{\mathbf{x}}}(x_1, \dots, x_m)$  with marginal CDFs  $F_{\underline{x}_1}, \dots, F_{\underline{x}_m}$ , assuming that they are with continuous and differentiable, can be written as,

$$F_{\underline{\mathbf{x}}}(x_1, \dots, x_m) = C\left(F_{\underline{x}_1}(x_1), \dots, F_{\underline{x}_m}(x_m)\right) \quad (4.3)$$

In this work we are interested in the Gaussian copula  $C^G(\cdot)$  which is defined as multivariate standard normal distribution with correlation matrix  $\tilde{\mathbf{R}}$  [e.g., [Embrechts et al., 2003](#)],

$$C^G(\mathbf{u}) = C(u_1, \dots, u_m) = \Phi_m(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_m); \tilde{\mathbf{R}}) \quad (4.4)$$

which apparently, after some substitutions can be transformed in in Eq. (4.2), i.e., NDM's main equation. Generally, according to copula theory, assuming that both  $F_{\underline{x}_i}$  and  $C(\cdot)$  are differentiable, the joint PDF of  $\underline{\mathbf{x}}$  can be written as,

$$f_{\underline{\mathbf{x}}}(x_1, \dots, x_m) = c\left(F_{\underline{x}_1}(x_1), \dots, F_{\underline{x}_m}(x_m)\right) \cdot \prod_{i=1}^m f_{\underline{x}_i}(x_i) \quad (4.5)$$

where  $c(\cdot)$  denotes the joint PDF (referred also as copula density) of copula  $C(\cdot)$  and it is given by,

$$c\left(F_{\underline{x}_1}(x_1), \dots, F_{\underline{x}_m}(x_m)\right) = c(u_1, \dots, u_m) = \frac{\partial^m C(u_1, \dots, u_m)}{\partial u_1 \cdots \partial u_m} \quad (4.6)$$

Interestingly, by rearranging Eq. (4.5) as follows,

$$c\left(F_{\underline{x}_1}(x_1), \dots, F_{\underline{x}_m}(x_m)\right) = c(u_1, \dots, u_m) = \frac{f_{\underline{x}}(x_1, \dots, x_m)}{\prod_{i=1}^m f_{\underline{x}_i}(x_i)} \quad (4.7)$$

It can be observed that the copula density  $c(\cdot)$  denotes the ratio of multivariate joint PDF to the case of independence, which can be translated as the necessary *correction* to transition from independence to dependence. It can be shown that in the case of Gaussian copula its joint density  $c^G(\cdot)$  can be written as:

$$c^G(u_1, \dots, u_m) = \frac{\varphi_m(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_m); \tilde{\mathbf{R}})}{\prod_{i=1}^m \varphi(\Phi^{-1}(u_i))} \quad (4.8)$$

Thus in the case of NDM and Gaussian copula the joint PDF of  $\underline{x}$  is given by substituting Eq. (4.8) to Eq. (4.5) [cf. Liu and Der Kiureghian, 1986],

$$f_{\underline{x}}(x_1, \dots, x_m) = \frac{\varphi_m\left(\Phi^{-1}\left(F_{\underline{x}_1}(x_1)\right), \dots, \Phi^{-1}\left(F_{\underline{x}_m}(x_m)\right); \tilde{\mathbf{R}}\right)}{\prod_{i=1}^m \varphi\left(\Phi^{-1}\left(F_{\underline{x}_i}(x_i)\right)\right)} \cdot \prod_{i=1}^m f_{\underline{x}_i}(x_i) \quad (4.9)$$

From these equations it is clear that, copula theory, in general, as well as NDM specifically, allow us to describe complex multivariate distributions using as individual components the marginal distributions  $F_{\underline{x}_1}, \dots, F_{\underline{x}_m}$  and the copula  $C(\cdot)$ , which eventually allow the formulation of the joint distribution.

Nevertheless, it is shown from Eq. (4.2) and Eq. (4.9), that in the case of NDM (i.e., Gaussian copula) the joint distribution of  $\underline{x}$  depends on the correlation matrix  $\tilde{\mathbf{R}}$  of  $\underline{z}$  and not directly on  $\mathbf{R}$  of  $\underline{x}$ . To elaborate, let us consider the inverse case where  $\underline{x}$  is obtained through  $\underline{z}$  via the following mapping equation:

$$\underline{x}_i = F_{\underline{x}_i}^{-1}\left(\Phi(\underline{z}_i)\right) \quad (4.10)$$

where  $F_{\underline{x}_i}^{-1}$  is the ICDF of variable  $\underline{x}_i$ . It is noted that similar to the previous case (i.e., Eq. (4.1))  $\underline{u}_i := \Phi(\underline{z}_i)$  is also a RV uniformly distributed in  $[0, 1]$  that denotes probability. A direct outcome of Eq. (4.10) is that for two variables  $\underline{x}_i$  and  $\underline{x}_j$  their correlation is given by:

$$\rho_{i,j} := \text{Corr}[\underline{x}_i, \underline{x}_j] = \text{Corr}\left[F_{\underline{x}_i}^{-1}\left(\Phi(\underline{z}_i)\right), F_{\underline{x}_j}^{-1}\left(\Phi(\underline{z}_j)\right)\right] \quad (4.11)$$

thus the target correlations  $\rho_{i,j}$  of  $\mathbf{R}$  are associated with the corresponding elements  $\tilde{\rho}_{i,j}$  of  $\tilde{\mathbf{R}}$ . An apparent approach could be setting  $\tilde{\mathbf{R}} \equiv \mathbf{R}$ , however, both theoretical and empirical evidence have indicated that this assumption will result in misspecification of the underlying model (i.e., NDM) and lead to systematically underestimating correlations within the generated

data. The theoretical justification of this behavior stems from the Pearson correlation coefficient itself, since it is not invariant under non-linear monotonic transformations, such as those imposed by the ICDFs [Embrechts *et al.*, 1999 p. 8]. More specifically, the largest the departure of the actual distribution,  $F_{\underline{x}_\xi}$ , from the normal one, the largest will be the underestimation. Therefore, and except the trivial normal case, in order to eliminate biases, it necessary to *a priori* identify the values of  $\tilde{\rho}_{i,j}$ .

NDM and its theoretical background can provide a theoretical solution to the above problem by means of specifying an appropriate (i.e., equivalent) correlation matrix  $\tilde{\mathbf{R}}$  that leads to the target correlation matrix  $\mathbf{R}$ . As highlighted by Liu and Der Kiureghian [1986], in order to employ NDM it is essential to ensure 1) one to one mapping of Eq. (4.10), and 2) positive definite correlation matrix  $\tilde{\mathbf{R}}$ . The first requirement is by definition valid in typical cases of distributions used in hydrology, while the second is also usually satisfied, since the distances  $\varepsilon_{i,j} := |\rho_{i,j} - \tilde{\rho}_{i,j}|$  are often small (provided, of course, that the target matrix  $\mathbf{R}$  is positive definite). The following procedure is applied to each specific pair of variables  $\underline{x}_i$  and  $\underline{x}_j$  of  $\underline{x}$  (i.e.,  $m(m-1)/2$  times). Given the definition of Pearson's correlation coefficient, i.e.,

$$\rho_{i,j} = \text{Corr}[\underline{x}_i, \underline{x}_j] = \frac{E[\underline{x}_i \underline{x}_j] - E[\underline{x}_i] E[\underline{x}_j]}{\sqrt{\text{Var}[\underline{x}_i] \text{Var}[\underline{x}_j]}} \quad (4.12)$$

where  $E[\underline{x}_i]$ ,  $E[\underline{x}_j]$  and  $\text{Var}[\underline{x}_i]$ ,  $\text{Var}[\underline{x}_j]$  are the mean and variance of  $\underline{x}_i$  and  $\underline{x}_j$  respectively, which are known since the associated marginal distributions are already specified (and have finite variance, otherwise the Pearson correlation coefficient cannot be defined). Thereby, the computational procedure is limited to identifying  $E[\underline{x}_i \underline{x}_j]$ . Since the corresponding variables to be mapped,  $\underline{z}_i$  and  $\underline{z}_j$ , respectively, are by definition jointly normally distributed, with correlation  $\text{Corr}[\underline{z}_i, \underline{z}_j] = \tilde{\rho}_{i,j}$ , then, using Eq. (4.10), the fundamental theorem of expectation (also known as the law of unconscious statistician) and the first cross-product moment of  $\underline{x}_i$  and  $\underline{x}_j$  we get:

$$\begin{aligned} E[\underline{x}_i \underline{x}_j] &= E\left[F_{\underline{x}_i}^{-1}\left(\Phi(\underline{z}_i)\right) F_{\underline{x}_j}^{-1}\left(\Phi(\underline{z}_j)\right)\right] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}_i}^{-1}\left(\Phi(z_i)\right) F_{\underline{x}_j}^{-1}\left(\Phi(z_j)\right) \varphi_2(z_i, z_j; \tilde{\rho}_{i,j}) dz_i dz_j \end{aligned} \quad (4.13)$$

where  $\varphi_2(z_i, z_j; \tilde{\rho}_{i,j})$  is the bivariate standard normal PDF. By substituting Eq. (4.13) to Eq. (4.12) we obtain,

$$\rho_{i,j} = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}_i}^{-1}\left(\Phi(z_i)\right) F_{\underline{x}_j}^{-1}\left(\Phi(z_j)\right) \varphi_2(z_i, z_j; \tilde{\rho}_{i,j}) dz_i dz_j - E[\underline{x}_i] E[\underline{x}_j]}{\sqrt{\text{Var}[\underline{x}_i] \text{Var}[\underline{x}_j]}} \quad (4.14)$$

For simplicity, let us rewrite this relationship as,

$$\rho_{i,j} = \mathcal{F}\left(\tilde{\rho}_{i,j} \middle| F_{\underline{x}_i}, F_{\underline{x}_j}\right) \quad (4.15)$$

where  $\mathcal{F}(\cdot)$  denotes an arbitrary function, which has the meaning that each target  $\rho_{i,j}$  is a function of  $\tilde{\rho}_{i,j}$ , that is embedded in  $\varphi_2(z_i, z_j; \tilde{\rho}_{i,j})$ , and the given marginal distributions  $F_{\underline{x}_i}$  and  $F_{\underline{x}_j}$ . Eq. (4.15) have to be inverted in order to identify the values of  $\tilde{\rho}_{i,j}$  that result in the target values  $\rho_{i,j}$ . i.e.,

$$\tilde{\rho}_{i,j} = \mathcal{F}^{-1} \left( \rho_{i,j} \left| F_{\underline{x}_i}, F_{\underline{x}_j} \right. \right) \quad (4.16)$$

Unfortunately, Eq. (4.15), and thus Eq. (4.16), does not have a general closed-form solution, with the exception of few special cases [Li and Hammond, 1975; Cario and Nelson, 1997; Crouse and Baraniuk, 1999; Xiao, 2014]. Among them the Log-Normal case [Mostafa and Mahmoud, 1964] which is of particular interest in hydrology (see section 4.5.2). The aforementioned researchers, as well as Liu and Der Kiureghian [1986], provided several Lemmas that can be useful in order to approximate Eq. (4.15). Among them,

**Lemma 1:**  $\rho_{i,j}$  is a strictly increasing function of  $\tilde{\rho}_{i,j}$ .

**Lemma 2:**  $\tilde{\rho}_{i,j} = 0$  for  $\rho_{i,j} = 0$  as well as,  $\tilde{\rho}_{i,j} \geq (\leq) 0$  if  $\rho_{i,j} \geq (\leq) 0$ .

**Lemma 3:**  $|\rho_{i,j}| \leq |\tilde{\rho}_{i,j}|$ .

Note that in Lemma 3, the equality sign is valid when  $\rho_{i,j} = 0$  or when both marginal distributions are normal. Furthermore, the minimum and maximum attainable value of  $\rho_{i,j}$  are in accordance with the Fréchet-Hoeffding bounds [Fréchet, 1957; Hoeffding, 1994] and are given for  $\tilde{\rho}_{i,j} = -1$  and  $\tilde{\rho}_{i,j} = 1$ , respectively. Particularly, the following holds true,  $-1 \leq \mathcal{F} \left( -1 \left| F_{\underline{x}_i}, F_{\underline{x}_j} \right. \right) \leq \rho_{i,j} \leq \mathcal{F} \left( 1 \left| F_{\underline{x}_i}, F_{\underline{x}_j} \right. \right) \leq 1$ . See also the work of Whitt [1976] for a comprehensive discussion on the topic.

In this Thesis, unless stated otherwise, in order to establish the relationship  $\mathcal{F}(\cdot)$  we employ the simple, yet efficient method proposed by Tsoukalas et al. [2018e] which in a nutshell, is based on the evaluation of few pairs of  $\rho_{i,j}$  and  $\tilde{\rho}_{i,j}$  using Monte Carlo simulation and subsequently, the establishment of the  $\mathcal{F}(\cdot)$  relationship through polynomial interpolation. For further details regarding the identification of equivalent correlation coefficients, as well as the above algorithm see section 4.5. Hereafter it is assumed that the equivalent correlation  $\tilde{\rho}_{i,j}$  have been properly identified.

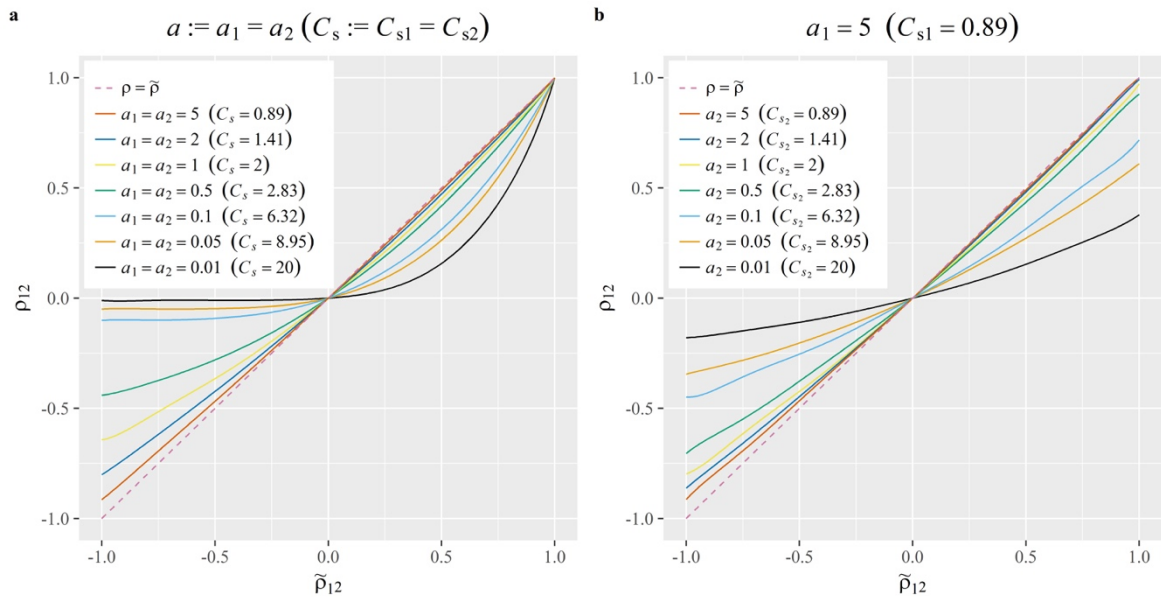
Nevertheless, in order to shed some light on the functional form of  $\mathcal{F}(\cdot)$  (i.e., Eq. (4.15)) let us consider the case of two variables  $\underline{x}_1$  and  $\underline{x}_2$  are described by the two-parameter Gamma distribution ( $\mathcal{G}$ ). Its PDF is given by:

$$f_{\mathcal{G}}(x; a, b) = \frac{1}{|b|\Gamma(a)} \left( \frac{x}{b} \right)^{a-1} \exp \left( -\frac{x}{b} \right), \quad x > 0 \quad (4.17)$$

where  $a > 0$  and  $b \neq 0$  are the shape and scale parameters respectively, while  $\Gamma(\cdot)$  denotes the gamma function. Figure 4.1a depicts the relationship among  $\tilde{\rho}_{1,2}$  and  $\rho_{1,2}$  (i.e.,  $\mathcal{F}(\cdot)$ , computed via numerical integration) for various values of distribution parameters. Specifically, we assumed  $a := a_1 = a_2$  and constant  $b := b_1 = b_2 = 1$ . We remind that the theoretical skewness coefficient of a Gamma distributed variable is given by  $C_{s_{\underline{x}}} = 2/\sqrt{a}$ . From this figure we also observe that the non-linearity of  $\mathcal{F}(\cdot)$  increases with low values of  $a$  (i.e., high

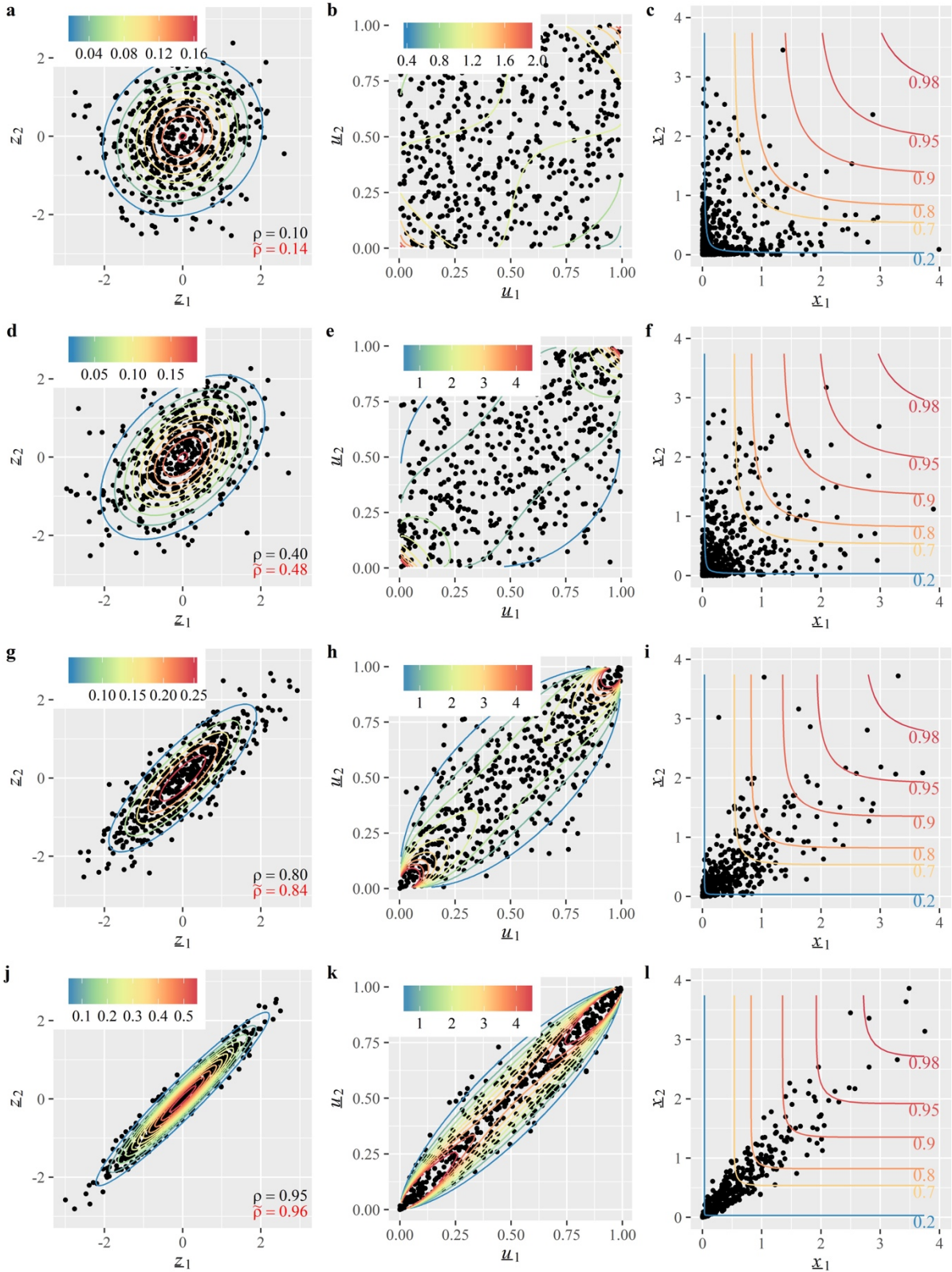


skewness), and that the maximum attainable value of  $\rho_{\xi,\psi}$  is equal to 1, which is due to the fact that  $F_{\underline{x}_2} \equiv F_{\underline{x}_1}$ . In addition, one may observe that the  $a$  is also related with the minimum attainable value of  $\rho_{1,2}$ . For example, when  $a = 0.01$ , the latter value is practically restricted to zero, something that may be considered a reasonable behavior that is attributed to the very high value of positive skewness which does not allow for negative correlations. In a similar vein, in **Figure 4.1b** we set  $a_1 = 5$  and vary the parameter  $a_2$  from 5 to 0.01 (assuming again that  $b := b_1 = b_2 = 1$ ). In this case, both the minimum and maximum attainable values of  $\rho_{1,2}$  are affected. We observe that when  $a_1$  and  $a_2$  exhibit significant differences, the range of feasible values  $\rho_{1,2}$  is getting narrower. This implies that two variables with considerable different shape (expressed through parameter  $a$ ) cannot be highly correlated. From an engineering point of view, and similar to the previous case (i.e., when  $a := a_1 = a_2$ ), this is barely considered a limitation of the NDM approach, since such behavior is rarely encountered in practice. For instance, it is not expected, or rational, two variables (or in general processes), one with skewness  $\sim 0.9$  and one with 20 to be highly (positively or negatively) correlated. In any case, we stress the importance of checking the range of attainable correlation coefficients when employing the concepts of NDM [see, [Demirtas and Hedeker, 2011](#); [Leonov and Qaqish, 2017](#)], especially within the context of stochastic process simulation (see section 4.3). For instance, given the non-linear and asymmetric nature of  $\mathcal{F}(\cdot)$ , for some combinations of marginal distributions, a target correlation coefficient may be inadmissible (this also applies in any Nataf-based construct, including stochastic processes; see section 4.3 as well as, Chapter 5, 6 and 7). However, in the examples (both hypothetical and real-world) employed in this Thesis such problems did not occur, a fact which by no means overrules the aforementioned need for compatibility verification.



**Figure 4.1** | Graphical illustration of function  $\mathcal{F}(\cdot)$  (i.e., Eq. (4.15)) that expresses the relationship between the equivalent,  $\tilde{\rho}_{1,2}$  and target  $\rho_{1,2}$  correlation coefficients assuming that both  $\underline{x}_1$  and  $\underline{x}_2$  are described by the two-parameter Gamma distribution (assuming that  $b := b_1 = b_2 = 1$ ) with a) equal shape parameters (i.e.,  $a := a_1 = a_2$ ) and b) different shape parameters by setting  $a_1 = 5$  and varying  $a_2$  from 5 to 0.01.

Additionally, in order to investigate the form of the established joint distribution functions, let us setup a hypothetic example where both  $\underline{x}_1$  and  $\underline{x}_2$  have identical marginal distribution,  $f_G(x; 0.5, 1)$  and  $\rho_{1,2} \in \{0.1, 0.4, 0.8, 0.95\}$ . Each row of subplots of **Figure 4.2**, corresponds to a specific value of  $\rho_{1,2}$  (the equivalent value of  $\tilde{\rho}_{1,2}$  is also shown with red color – see also the corresponding curve in **Figure 4.1a**) and depicts, from left to right, the joint PDF in the Gaussian and uniform domain (copula density) as well as the joint CDF in the *real* domain. It is remarked that isolines of these plots were drawn using the theoretical equations provided by the mathematical background of NDM while, some random samples were generated (black dots) for visualization purposes. The plot depicts, in a step-by-step manner, the mapping procedure imposed by Eq. (4.10), as it illustrates the transition, from the Gaussian to uniform (copula) and finally the actual domain. Besides the various forms of the attained joint PDF it is also interesting to observe how NDM transforms a homoscedastic input (i.e., Gaussian) to a heteroscedastic one.



**Figure 4.2** | Hypothetical example of two RVs,  $x_1, x_2 \sim \mathcal{G}(a = 0.5, b = 1)$  with  $\rho_{1,2} \in \{0.1, 0.4, 0.8, 0.95\}$ . Each row of subplots corresponds to a specific value of  $\rho_{1,2}$  and each column of subplots, from left to right, depicts the joint PDF in the Gaussian and uniform domain as well the joint CDF in the actual domain.

### 4.1.3 Unconditional Monte Carlo simulation

Unconditional Monte Carlo simulation is implemented in many applications in science and real-world practice [e.g., *Robert and Casella, 2010; Kroese et al., 2011, 2014*]. Characteristic fields are those of physics, biology, finance and engineering. Often, its purpose is to propagate uncertainties related with input variables to the outputs of interest. For example, in finance, it is used to identify optimal portfolios of financial securities. In this case, each security (the input variables) is assumed to be a RV (often time-independent), and thus described by a distribution function. Since the overall portfolio (i.e., its performance; the output of interest) is composed by a linear combination of them, the ultimate goal is to estimate its efficiency and risk, i.e., the uncertainty of the overall portfolio (typically expressed via measures of dispersion or through its distribution function) via propagating the uncertainty of each individual security. This task can be accomplished with the use of Monte Carlo simulation, which, loosely speaking, holds out the promise of generating correlated inputs with the desired marginal distributions. In a similar vein, in engineering, one often has at his disposal a deterministic model that under some inputs returns the output variable of interest. In many cases we are particularly interested in quantifying (and eventually accounting for) the input variables uncertainty<sup>2</sup>. Analogous to the previous example, the central idea is to feed the deterministic model with multiple realizations of the input variables, run the model multiple times, and finally, derive the distribution function of the output variable of interest. Interestingly, both examples, and many other applications [e.g., *Makropoulos et al., 2017; Tsoukalas et al., 2017b; Psarrou et al., 2018*] of Monte Carlo simulation, can be viewed as derived distribution problems, since, regardless of the case, we are interested in deriving the distribution function of the output variable(s). This kind of problems are becoming particularly challenging when the input variables are non-Gaussian and cross-correlated. A potential remedy in such cases relies on the use of copulas. As such, the theoretical background of NDM [*Nataf, 1962*] and the NORTA procedure [*Cario and Nelson, 1997*], can provide a well-justified and easy to implement solution.

Specifically, the problem of generating a  $m$ -dimensional correlated random vector  $\underline{x} = [x_1, \dots, x_j, \dots, x_m]^T$  with *a priori* specified target marginal distributions  $F_{x_1}, \dots, F_{x_j}, \dots, F_{x_m}$  and target correlation matrix  $\mathbf{R} \in [-1, 1]^{m \times m}$  reduces to generating and subsequently map using  $x_j = F_{x_j}^{-1}(\Phi(z_j))$  (i.e., Eq. (4.10)), a  $m$ -dimensional correlated random vector  $\underline{z} = [z_1, \dots, z_j, \dots, z_m]^T$  with multivariate standard normal distribution and equivalent correlation matrix  $\tilde{\mathbf{R}} \in [-1, 1]^{m \times m}$ . Assuming that the elements of  $\tilde{\mathbf{R}}$  has already been specified (e.g., using the algorithm of section 4.5.1), it is straightforward to generate a random vector  $\underline{z}$  through the following linear transformation,

$$\underline{z} = \tilde{\mathbf{B}}\underline{w} \quad (4.18)$$

of an uncorrelated standard normal vector  $\underline{w} = [w_1, \dots, w_j, \dots, w_m]^T$ , where  $\tilde{\mathbf{B}}$  is a  $m \times m$  matrix obtained by,

$$\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T = \tilde{\mathbf{R}} \quad (4.19)$$

---

<sup>2</sup> It could also be of interest the quantification of other sources of uncertainty, such as parameter, structural, etc. Regardless of the case, Monte Carlo simulation is the most frequently employed approach for this purpose.

In order to obtain the matrix  $\tilde{\mathbf{B}}$ , it is essential to solve a decomposition problem, also expressed as finding the square root of  $\tilde{\mathbf{R}}$ . This can be achieved with the use of typical numerical techniques, such as Cholesky or singular value decomposition [e.g., *Johnson, 1987*]. We remark that when  $\tilde{\mathbf{R}}$  is positive definite, it has infinite number of feasible solutions, such as the solutions provided by the aforementioned numerical methods. On the other hand, if  $\tilde{\mathbf{R}}$  is non-positive definite the problem does not have a feasible solution, thus requiring the detection (e.g., through optimization [*Koutsoyiannis, 1999; Higham, 2002*]) of a parameter matrix  $\tilde{\mathbf{B}}^*$  which results to a feasible and near-to-optimum matrix  $\tilde{\mathbf{R}}^* := \tilde{\mathbf{B}}^* \tilde{\mathbf{B}}^{*\text{T}}$  which is as closest (typically quantified in terms of some distance measure; e.g., Euclidean norm) as possible to the original matrix  $\tilde{\mathbf{R}}$ , thereby, ensuring an approximation of the given  $\tilde{\mathbf{R}}$ .

Some applications involve the derivation of the distribution function of some linear combination of  $\underline{\mathbf{x}}$ , expressed using the column vector  $\boldsymbol{\omega} = [\omega_1, \dots, \omega_m]^{\text{T}}$ , which can be viewed also as weight coefficients. The first two moments of the derived distribution can be easily calculated analytically, using the equations below, however higher order moments are hard to estimate analytically (if not impossible) and the derivation of the complete distribution requires the resolution of the so-called convolution integral (in the case of independence). Unconditional Monte Carlo procedures provides the means to approximate the derived distribution by means of simulation, even in cases of non-linear combinations of  $\underline{\mathbf{x}}$ . Regarding the case of linear combination of  $\boldsymbol{\omega}^{\text{T}}\underline{\mathbf{x}}$ , the mean,  $E[\boldsymbol{\omega}^{\text{T}}\underline{\mathbf{x}}]$ , and variance,  $\text{Var}[\boldsymbol{\omega}^{\text{T}}\underline{\mathbf{x}}]$ , can be calculated by [*Lindgren, 2013*],

$$E[\boldsymbol{\omega}^{\text{T}}\underline{\mathbf{x}}] = E\left[\sum_{i=1}^m \omega_i x_i\right] = \boldsymbol{\omega}^{\text{T}}E[\underline{\mathbf{x}}] = \sum_{i=1}^m \omega_i E[x_i] \quad (4.20)$$

$$\text{Var}[\boldsymbol{\omega}^{\text{T}}\underline{\mathbf{x}}] = \text{Var}\left[\sum_{i=1}^m \omega_i x_i\right] = \boldsymbol{\omega}^{\text{T}}\boldsymbol{\Sigma}\boldsymbol{\omega} = \sum_{i=1}^m \sum_{j=1}^m \omega_i \omega_j \text{Cov}[x_i, x_j] \quad (4.21)$$

where  $\boldsymbol{\Sigma} := \text{Cov}[\underline{\mathbf{x}}, \underline{\mathbf{x}}^{\text{T}}] = \mathbf{D}\mathbf{R}\mathbf{D}^{\text{T}}$  is the covariance matrix ( $m \times m$ ) of  $\underline{\mathbf{x}}$  and  $\mathbf{D} := \text{diag}[\sqrt{\text{Var}[x_1]}, \dots, \sqrt{\text{Var}[x_m]}]$  is a diagonal ( $m \times m$ ) matrix which contains the square root of the variables variances (i.e., standard deviations; apparently  $\mathbf{D} = \mathbf{D}^{\text{T}}$ ), which are known for all  $x_i$ , since the marginal distributions are known (or specified). The inverse operation is simply,  $\mathbf{R} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$ . As a side note, it is worth mentioning that the above relationship can be generalized for the estimation of the covariance of linear combinations of two random vectors  $\underline{\mathbf{x}} = [x_1, \dots, x_i, \dots, x_m]^{\text{T}}$  and  $\underline{\mathbf{y}} = [y_1, \dots, y_i, \dots, y_n]^{\text{T}}$  with  $\boldsymbol{\omega} = [\omega_1, \dots, \omega_m]^{\text{T}}$  and  $\boldsymbol{\beta} = [\beta_1, \dots, \beta_n]^{\text{T}}$  expressing the weight coefficients of  $\underline{\mathbf{x}}$  and  $\underline{\mathbf{y}}$  respectively. The covariance  $\text{Cov}[\boldsymbol{\omega}^{\text{T}}\underline{\mathbf{x}}, \boldsymbol{\beta}^{\text{T}}\underline{\mathbf{y}}]$  can be analytically estimated by,

$$\text{Cov}[\boldsymbol{\omega}^{\text{T}}\underline{\mathbf{x}}, \boldsymbol{\beta}^{\text{T}}\underline{\mathbf{y}}] = \text{Cov}\left[\sum_{i=1}^m \omega_i x_i, \sum_{j=1}^n \beta_j y_j\right] = \sum_{i=1}^m \sum_{j=1}^n \omega_i \beta_j \text{Cov}[x_i, y_j] \quad (4.22)$$

#### 4.1.4 Numerical examples

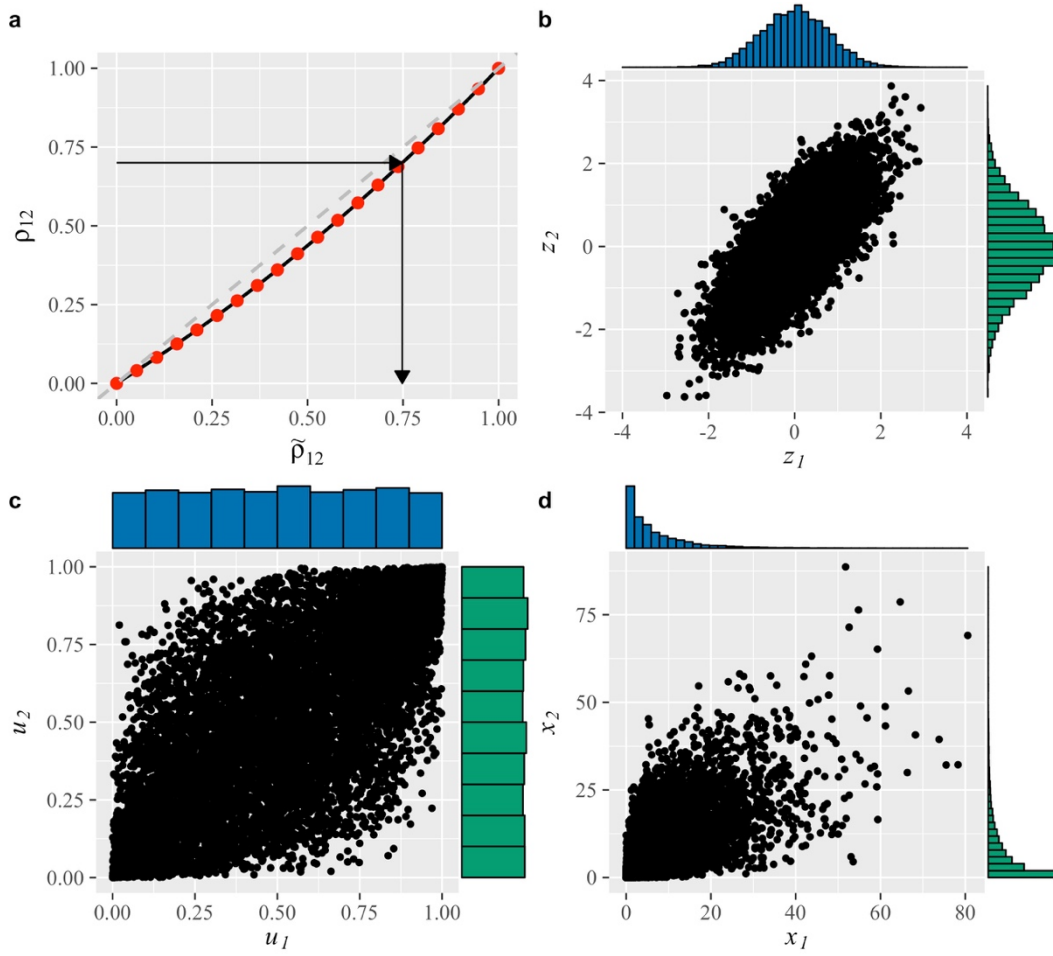
To elaborate more on the practical side of NDM, the unconditional Monte Carlo simulation technique, as well as provide few a *hands-on* examples on the identification of equivalent correlation coefficients and the mapping procedure of Eq. (4.10), let us consider the following simple bivariate cases of generating correlated non-Gaussian RVs. Throughout these examples the algorithms of sections 4.1.3 and 4.5.1 are employed.

##### 4.1.4.1 Continuous-type marginal distributions

Let assume that we wish to generate two correlated variables  $x_1$  and  $x_2$ , both from the same distribution (i.e.,  $F_{x_1} \equiv F_{x_2}$ ), the two-parameter Gamma distribution (Eq. (4.17)). Additionally, let assume that the parameters of both distributions are  $a := a_1 = a_2 = 0.7$ ,  $b := b_1 = b_2 = 10$  and the target correlation among them is  $\rho_{12} = 0.7$ . After employing the algorithm of section 4.5.1 (using  $\tilde{\rho}_{\min} = 0$  and  $\tilde{\rho}_{\max} = 1$ ,  $N = 50\,000$ ,  $\Omega = 20$  and  $p = 2$ ) the following polynomial relationship is established between the equivalent and target correlation coefficients (the indices were omitted for simplicity).

$$\rho = \mathcal{F}(\tilde{\rho} | F_{x_1}, F_{x_2}) \approx 0.2694\tilde{\rho}^2 + 0.7234\tilde{\rho} + 0.0061 \quad (4.23)$$

This relationship is depicted graphically in Figure 4.3a, which highlights its non-linearity. It is apparent (from Figure 4.3a and Eq. (4.23)) that in order to attain the target correlation (= 0.7) between  $x_1$  and  $x_2$  it is suggested to generate standard normal variables ( $z_1$  and  $z_2$ ) with correlation equal to  $\tilde{\rho}_{12} \cong 0.75$ . Therefore, we simulated 10 000 data ( $z_1$  and  $z_2$ ) from a bivariate normal distribution with correlation equal to 0.75 which are first mapped to the uniform domain, i.e.,  $u_1 = \Phi(z_1)$  and  $u_2 = \Phi(z_2)$ , and then mapped to the actual (i.e., real) domain using their ICDF, i.e.,  $x_1 = F_{x_1}^{-1}(u_1)$  and  $x_2 = F_{x_2}^{-1}(u_2)$ . Figure 4.3b-d graphically depicts the previous procedure which is a step-by-step equivalent to Eq. (4.10). The simulated data ( $x_1$  and  $x_2$ ) in the actual domain attain the target correlation (i.e., 0.7) as well as the specified marginal distributions. Furthermore, in order to further validate this statement, we used the maximum likelihood estimation (MLE) method to identify the distribution parameters of the simulated data. Their estimates were found equal to  $\hat{a}_1 = 0.697$ ,  $\hat{b}_1 = 9.960$  for  $x_1$  and  $\hat{a}_2 = 0.701$ ,  $\hat{b}_2 = 10.060$ , for  $x_2$ , which indicate a very close approximation of the specified parameters.



**Figure 4.3** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) identical Gamma-distributed variables with  $\mathcal{G}(0.7, 10)$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients (i.e., Eq. (4.23)). Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\hat{\rho}}_{12} = 0.7504$  and  $\hat{\rho}_{12} = 0.7073$  respectively.

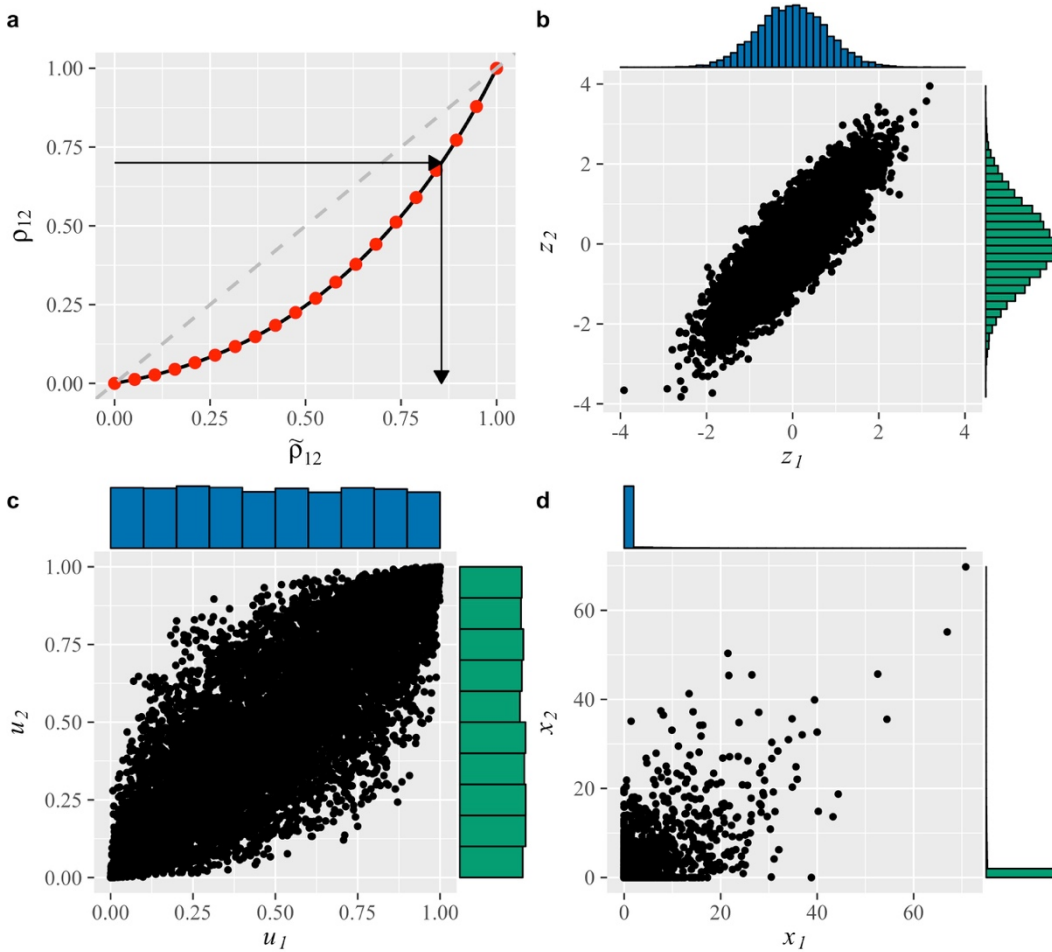
#### 4.1.4.2 Discrete-continuous-type marginal distributions

Let us further extended this example by employing a zero-inflated distribution model such as the one discussed later in section 4.4. This type of distribution is comprised by an atom of mass at zero and a continuous part for positive values. Its CDF is given by (after some slight notational modifications),

$$F_{\underline{x}}(x) = \begin{cases} p_0, & x \leq 0 \\ p_0 + (1 - p_0)G_{\underline{x}}(x), & x > 0 \end{cases} \quad (4.24)$$

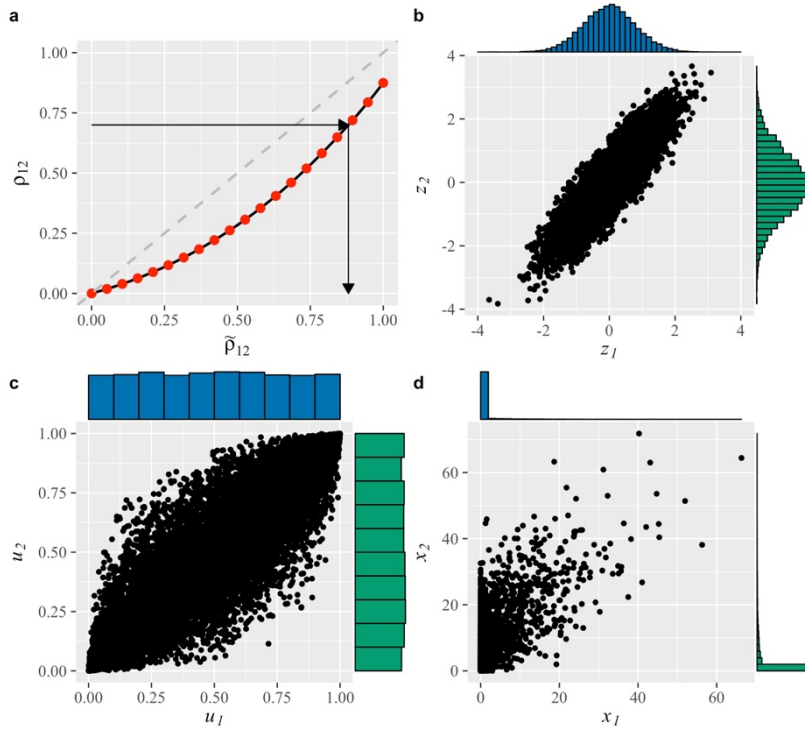
where,  $p_0$  denotes the probability of observing a zero value, i.e.,  $p_0 := P(\underline{x} = 0)$  and  $G_{\underline{x}}(x)$  stands for the continuous distribution part, that entails values greater than zero, i.e.,  $G_{\underline{x}}(x) := F_{\underline{x}|\underline{x}>0}(x) = P(\underline{x} \leq x | \underline{x} > 0)$ . Using this zero-inflated CDF model, let us consider three scenarios with different values of  $p_0$ , where in all three we assume a target correlation  $\rho_{12} = 0.7$  and that the continuous parts of  $\underline{x}_1$  and  $\underline{x}_2$  are Gamma-distributed with  $\mathcal{G}(0.7, 10)$ . i.e.,  $G_{\underline{x}_1} \equiv G_{\underline{x}_2} \equiv \mathcal{G}(0.7, 10)$ . The scenarios differentiate on the specified values of  $p_0$ . Specifically, in the first scenario it is assumed that they have identical probability zero  $p_0 = p_{0;\underline{x}_1} = p_{0;\underline{x}_2} =$

0.9. In the second that  $p_{0;x_1} = 0.9$  and  $p_{0;x_2} = 0.6$ , while in the third that  $p_{0;x_1} = 0.9$  and  $p_{0;x_2} = 0$ . For each scenario 10 000 data were simulated. The results from the three scenarios are visually summarized in **Figure 4.4**, **Figure 4.5** and **Figure 4.6** respectively. It is interesting to observe how the required equivalent correlation coefficient increases compared to the previous example, from  $\tilde{\rho}_{12} \cong 0.75$  (which had the same continuous marginal; thus can be viewed as the limiting case of  $p_0 = 0$ ) to  $\tilde{\rho}_{12} \cong 0.85$  (1<sup>st</sup> case),  $\tilde{\rho}_{12} \cong 0.88$  (2<sup>nd</sup> case) and  $\tilde{\rho}_{12} \cong 0.94$  (3<sup>rd</sup> case). It also remarkable to notice how the non-linearity of  $\mathcal{F}(\cdot)$  evolves as the marginal distribution depart from Gaussianity. Finally, and in line with the continuous-type case, it was validated, using the MLE method, that the continuous part of the zero-inflated, mixed distribution was correctly and accurately simulated. In addition, it was empirically confirmed (simply by estimating the portion of zero-valued data) that the realized data had the desired atom at zero (i.e.,  $p_0$ ).

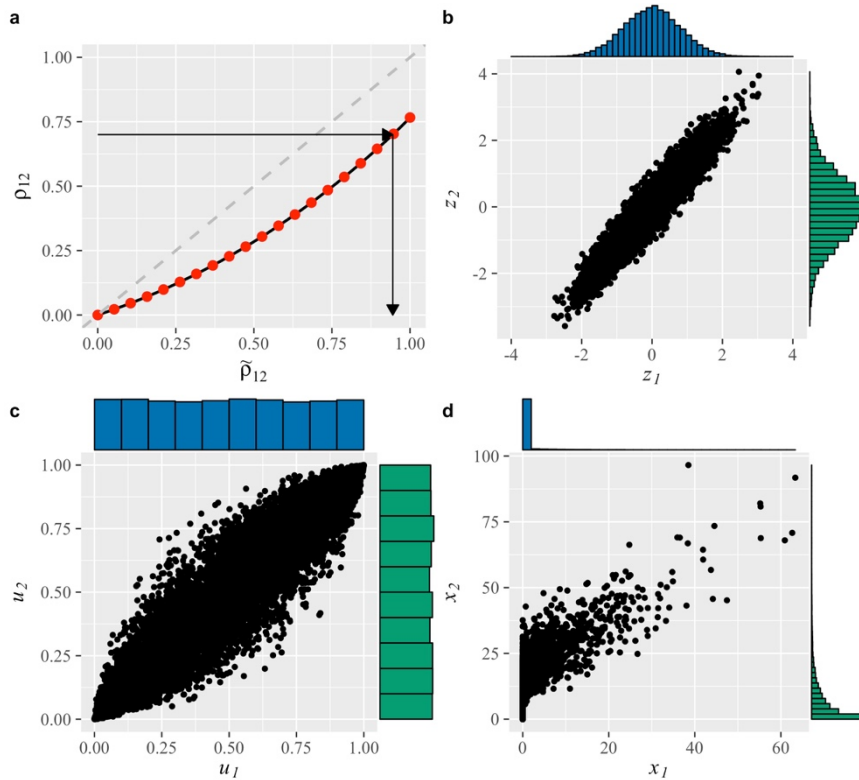


**Figure 4.4** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) zero-inflated Gamma-distributed variables with identical continuous part  $\mathcal{G}(0.7, 10)$  and  $p_{0;x_1} = p_{0;x_2} = 0.9$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\hat{\tilde{\rho}}_{12} = 0.8431$  and  $\hat{\rho}_{12} = 0.6989$  respectively.





**Figure 4.5 |** Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) zero-inflated Gamma-distributed variables with identical continuous part  $\mathcal{G}(0.7, 10)$ ,  $p_{0;x_1} = 0.9$  and  $p_{0;x_2} = 0.6$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\rho}_{12} = 0.8795$  and  $\hat{\rho}_{12} = 0.7071$  respectively.



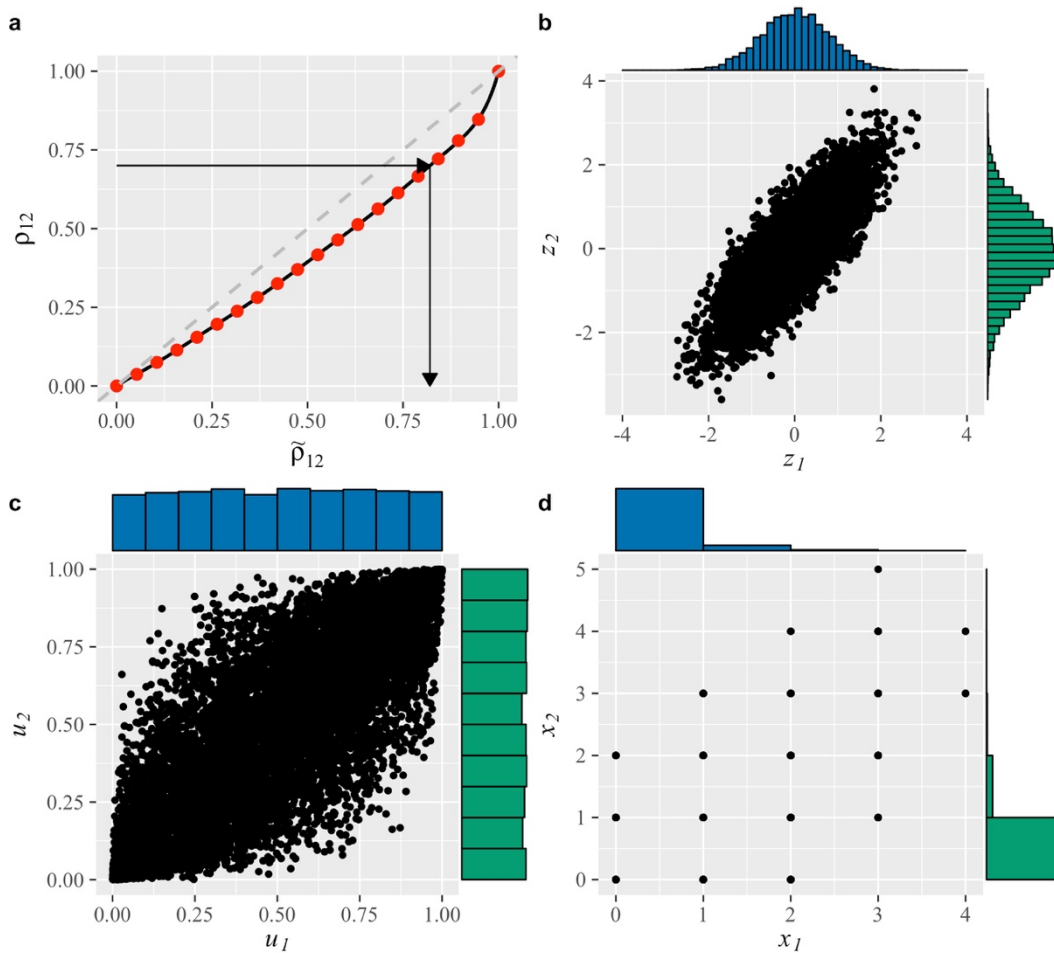
**Figure 4.6 |** Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) zero-inflated Gamma-distributed variables with identical continuous part  $\mathcal{G}(0.7, 10)$ ,  $p_{0;x_1} = 0.9$  and  $p_{0;x_2} = 0$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\rho}_{12} = 0.9422$  and  $\hat{\rho}_{12} = 0.7096$  respectively.

4.1.4.3 Discrete-type marginal distributions

As a final example, let us consider the bivariate case of correlated RVs from Poisson distribution. The probability mass function (PMF) of the Poisson distribution is given by,

$$P_{\mathcal{P}oi}(x; \lambda) = \frac{\exp(-\lambda) \lambda^x}{x!}, \quad x = 0, 1, 2, \dots \tag{4.25}$$

where  $\lambda > 0$  is the distribution parameter; and has the meaning of average number of occurrences within a time interval. Specifically, it was imposed a target correlation value equal to 0.7 and was it assumed that the distribution of  $\underline{x}_1$  and  $\underline{x}_2$  is identical, with  $\lambda_1 = \lambda_2 = 0.5$ . In a similar vein with the previous examples, **Figure 4.7** summarized the outcomes of this demonstration and validates the ability of the scheme to generate correlated discrete-type RVs. Also, in this case, the MLE of the simulated data closely resembled the theoretical parameters. Specifically, it was found that its estimates for  $\underline{x}_1$  and  $\underline{x}_2$  were,  $\hat{\lambda}_1 = 0.498$  and  $\hat{\lambda}_2 = 0.503$  respectively.



**Figure 4.7** | Hypothetical example of generating two correlated ( $\rho_{1,2} = 0.7$ ) identical Poisson-distributed variables with  $\mathcal{P}oi(0.5)$ . a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. Scatter plots with histograms in the b) Gaussian c) uniform and d) actual domain. The estimate of correlation coefficient of the simulate variables in the Gaussian and the actual domain is  $\tilde{\rho}_{12} = 0.8193$  and  $\hat{\rho}_{12} = 0.7054$  respectively.

## 4.2 A NATAF-BASED CONDITIONAL DISTRIBUTION MODEL

### 4.2.1 Theoretical background

This section extends the rationale of NDM for the derivation of conditional distributions, as well as conditional simulation of RVs (and processes) with pre-specified distributions and correlation matrix.

Similarly to the previous sections, let  $\underline{\mathbf{x}} = [\underline{x}_1, \dots, \underline{x}_m]^T$  be a  $m$ -dimensional vector of RVs, with known distributions  $F_{\underline{x}_1}, \dots, F_{\underline{x}_m}$  and correlation matrix  $\mathbf{R}$ , partitioned in a  $n$ -dimensional column-vector  $\underline{\mathbf{x}}_1^* = [\underline{x}_1, \dots, \underline{x}_n]^T$  and in a  $(m - n) \times 1$  column-vector  $\underline{\mathbf{x}}_2^* = [\underline{x}_{n+1}, \dots, \underline{x}_m]^T$ . Let also  $\mathbf{h} = [x_{n+1}, \dots, x_m]^T$  denote a vector of realizations of  $\underline{\mathbf{x}}_2^*$  on which we wish to condition the derivation of the distribution of  $\underline{\mathbf{x}}_1^* | \underline{\mathbf{x}}_2^* = \mathbf{h}$ .

As will be shown, in order to derive the conditional distribution it suffice to derive the one of the auxiliary RVs  $\underline{\mathbf{z}}$ . This can be done by using well-known properties of the auxiliary multivariate standard normal distribution [e.g., [Eaton, 1983](#)]. Particularly, let the auxiliary  $m$ -dimensional vector  $\underline{\mathbf{z}} = [z_1, \dots, z_m]^T$  with  $\underline{\mathbf{z}} \sim \mathcal{N}_m(\mathbf{0}, \tilde{\mathbf{R}})$  be similarly partitioned in  $\underline{\mathbf{z}}_1^* = [z_1, \dots, z_n]^T$  and  $\underline{\mathbf{z}}_2^* = [z_{n+1}, \dots, z_m]^T$  with sizes  $n \times 1$  and  $(m - n) \times 1$  respectively. This allow us to partition the equivalent correlation matrix  $\tilde{\mathbf{R}}$  as follows (it is also noted that,  $\tilde{\mathbf{R}}_{12} = \tilde{\mathbf{R}}_{21}^T$ ),

$$\tilde{\mathbf{R}} = \begin{bmatrix} \tilde{\mathbf{R}}_{11} & \tilde{\mathbf{R}}_{12} \\ \tilde{\mathbf{R}}_{21} & \tilde{\mathbf{R}}_{22} \end{bmatrix} \text{with sizes } \begin{bmatrix} n \times n & n \times (m - n) \\ (m - n) \times n & (m - n) \times (m - n) \end{bmatrix} \quad (4.26)$$

Furthermore, if  $\underline{\mathbf{z}}_2^* = \tilde{\mathbf{h}} = [\Phi^{-1}(F_{\underline{x}_{n+1}}(h_{n+1})), \dots, \Phi^{-1}(F_{\underline{x}_m}(h_m))]^T$  then the conditional distribution of  $\underline{\mathbf{z}}_1^* | \underline{\mathbf{z}}_2^* = \tilde{\mathbf{h}}$  is also multivariate normal, i.e.,  $P(\underline{\mathbf{z}}_1^* \leq \mathbf{z}_1^* | \underline{\mathbf{z}}_2^* = \tilde{\mathbf{h}}) \sim \mathcal{N}_n(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ , where  $\tilde{\boldsymbol{\mu}} = \tilde{\mathbf{R}}_{12} \tilde{\mathbf{R}}_{22}^{-1} \tilde{\mathbf{h}}$  and  $\tilde{\boldsymbol{\Sigma}} = \tilde{\mathbf{R}}_{11} - \tilde{\mathbf{R}}_{12} \tilde{\mathbf{R}}_{22}^{-1} \tilde{\mathbf{R}}_{21}$  denote the conditional mean vector and covariance matrix. The matrix  $\tilde{\boldsymbol{\Sigma}}$  can be easily calculated by exploiting the fact that it is Schur complement of  $\tilde{\mathbf{R}}_{22}$  in  $\tilde{\mathbf{R}}$ . This allows the calculation of  $\tilde{\boldsymbol{\Sigma}}$  via the inversion of the matrix  $\tilde{\mathbf{R}}$ , the subsequent removal of columns and vectors that correspond to the variables conditioned upon (i.e.,  $\underline{\mathbf{z}}_2^*$ ), and finally  $\tilde{\boldsymbol{\Sigma}}$  is obtained by the inversion of the remaining matrix.

Nevertheless, since Eq. (4.1) holds true, and similar to Eq. (4.2), the conditional CDF of  $\underline{\mathbf{x}}_1^* | \underline{\mathbf{x}}_2^* = \mathbf{h}$  can be written as,

$$F_{\underline{\mathbf{x}}_1^* | \underline{\mathbf{x}}_2^* = \mathbf{h}}(\mathbf{x}_1^*) = P(\underline{\mathbf{x}}_1^* \leq \mathbf{x}_1^* | \underline{\mathbf{x}}_2^* = \mathbf{h}) = \Phi_{n; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}} \left( \Phi^{-1} \left( F_{\underline{x}_1}(x_1) \right), \dots, \Phi^{-1} \left( F_{\underline{x}_n}(x_n) \right); \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}} \right) \quad (4.27)$$

where  $\Phi_{n; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}}(\cdot)$  denotes the multivariate joint CDF of  $\mathcal{N}_n(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ .

Furthermore, in order derive the conditional joint PDF of  $\underline{\mathbf{x}}_1^* | \underline{\mathbf{x}}_2^* = \mathbf{h}$  let us recall the following general property regarding conditional distributions [[Papoulis, 1991 p. 192](#)],

$$f_{\underline{\mathbf{x}}_1^* | \underline{\mathbf{x}}_2^* = \mathbf{h}}(\mathbf{x}_1^* | \underline{\mathbf{x}}_2^* = \mathbf{h}) = \frac{f_{\underline{\mathbf{x}}_1^*, \underline{\mathbf{x}}_2^*}(\mathbf{x}_1^*, \mathbf{x}_2^*)}{f_{\underline{\mathbf{x}}_2^*}(\mathbf{x}_2^*)} = \frac{f_{\underline{\mathbf{x}}}(x_1, \dots, x_m)}{f_{\underline{\mathbf{x}}_2^*}(x_{n+1}, \dots, x_m)} \quad (4.28)$$

thus by substituting Eq. (4.9) in the previous equation, the following relationship is obtained,

$$\begin{aligned}
& f_{\underline{x}_1^* | \underline{x}_2^* = \mathbf{h}}(\mathbf{x}_1^* | \mathbf{x}_2^* = \mathbf{h}) = \\
& \frac{\varphi_m \left( \Phi^{-1} \left( F_{\underline{x}_1} (x_1) \right), \dots, \Phi^{-1} \left( F_{\underline{x}_m} (x_m) \right); \tilde{\mathbf{R}} \right)}{\prod_{i=1}^m \varphi \left( \Phi^{-1} \left( F_{\underline{x}_i} (x_i) \right) \right)} \cdot \prod_{i=1}^m f_{\underline{x}_i} (x_i) \\
& \frac{\varphi_{(m-n)} \left( \Phi^{-1} \left( F_{\underline{x}_{n+1}} (x_{n+1}) \right), \dots, \Phi^{-1} \left( F_{\underline{x}_m} (x_m) \right); \tilde{\mathbf{R}}_{22} \right)}{\prod_{i=n+1}^m \varphi \left( \Phi^{-1} \left( F_{\underline{x}_i} (x_i) \right) \right)} \cdot \prod_{i=n+1}^m f_{\underline{x}_i} (x_i)
\end{aligned} \tag{4.29}$$

which after the cancelations of several terms reduces to the following relationship,

$$\begin{aligned}
& f_{\underline{x}_1^* | \underline{x}_2^* = \mathbf{h}}(\mathbf{x}_1^* | \mathbf{x}_2^* = \mathbf{h}) = \\
& \frac{\varphi_m \left( \Phi^{-1} \left( F_{\underline{x}_1} (x_1) \right), \dots, \Phi^{-1} \left( F_{\underline{x}_m} (x_m) \right); \tilde{\mathbf{R}} \right) \cdot \prod_{i=1}^n f_{\underline{x}_i} (x_i)}{\varphi_{(m-n)} \left( \Phi^{-1} \left( F_{\underline{x}_{n+1}} (x_{n+1}) \right), \dots, \Phi^{-1} \left( F_{\underline{x}_m} (x_m) \right); \tilde{\mathbf{R}}_{22} \right) \cdot \prod_{i=1}^n \varphi \left( \Phi^{-1} \left( F_{\underline{x}_i} (x_i) \right) \right)}
\end{aligned} \tag{4.30}$$

that can be further simplified to (since the left part of the products is essentially the conditional Gaussian PDF of  $\underline{z}_1^* | \underline{z}_2^* = \tilde{\mathbf{h}}$ ),

$$\begin{aligned}
& f_{\underline{x}_1^* | \underline{x}_2^* = \mathbf{h}}(\mathbf{x}_1^* | \mathbf{x}_2^* = \mathbf{h}) = \\
& \prod_{i=1}^n f_{\underline{x}_i} (x_i) \cdot \frac{\varphi_{n; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}} \left( \Phi^{-1} \left( F_{\underline{x}_1} (x_1) \right), \dots, \Phi^{-1} \left( F_{\underline{x}_n} (x_n) \right); \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}} \right)}{\prod_{i=1}^n \varphi \left( \Phi^{-1} \left( F_{\underline{x}_i} (x_i) \right) \right)}
\end{aligned} \tag{4.31}$$

where  $\varphi_{n; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}}(\cdot)$  denotes the PDF of  $\mathcal{N}_n(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ .

Additionally, in case of  $n = 1$  ( $\underline{x}_1^*$  is a single RV) it is also possible to derive a direct expression for the conditional ICDF, which reads,

$$x_{1(u_1)}^* = F_{\underline{x}_1^* | \underline{x}_2^* = \mathbf{h}}^{-1}(u_1) = F_{\underline{x}_1^*}^{-1} \left( \Phi \left( \Phi_{1; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}}^{-1} \left( u_1; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}} \right) \right) \right) \tag{4.32}$$

where  $\Phi_{1; \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}}^{-1}(\cdot)$  denotes the ICDF of the conditional standard normal distribution, while  $u_1 \in [0,1]$  is a scalar that denotes probability. Using Eq. (4.32) it is possible to estimate the desired conditional quantiles by simply plugging the target value of  $u_1$  (e.g.,  $u_1 = 0.5$ , would return the conditional median of  $\underline{x}_1^* | \underline{x}_2^* = \mathbf{h}$ ). To the best of our knowledge, this formulation is new, and it is argued that it can be viewed as a  $m$ -dimensional generalization of the expression given in [Kelly and Krzysztofowicz \[1997\]](#), for the derivation of conditional quantiles of bivariate meta-Gaussian distributions. Although it is important to note that the above authors did not employed the concept of equivalent correlations and instead relied on the use of rank-based

dependence measures for the parameter identification of the auxiliary model, i.e., the multivariate the Gaussian distribution (see also the discussion of section 4.5.3). Finally, it is remarked that while the discussion of the previous paragraphs was mainly focused on conditional distributions of RVs, their extension for stochastic processes and time series probabilistic forecasting is straightforward throughout the concepts of stationarity and cyclostationarity. In this vein, it is also interesting to note that the literature offers techniques for the optimal infilling of partially completed correlation matrices [e.g., Papoulis, 1991; Georgescu et al., 2017], which may be particularly useful in time series forecasting problems that involve large matrices and/or exogenous variables. Nowadays, hydrological forecasting is performed using point-based (hence not probabilistic) machine learning methods, such as neural networks [e.g., Hsu et al., 1995; Lekkas et al., 2001, 2004; Jain and Kumar, 2007; Wang et al., 2009].

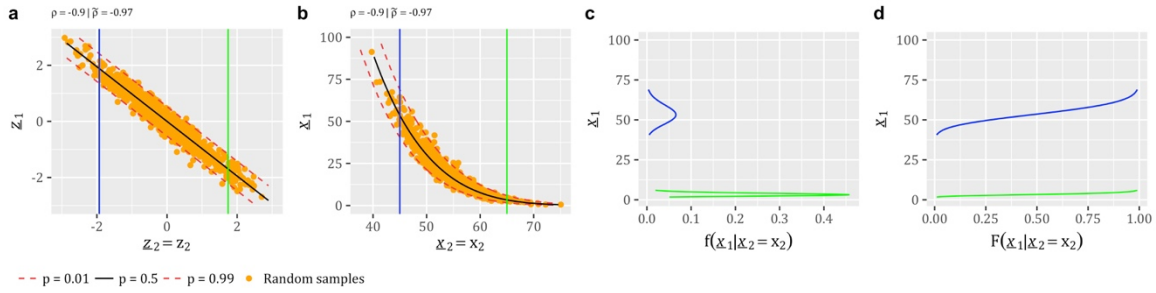
#### 4.2.2 Conditional Monte Carlo Simulation

Based on these developments and provided that the required inputs are already specified, i.e., the target marginal distributions  $F_{\underline{x}_1}, \dots, F_{\underline{x}_m}$ , the target correlation matrix  $\mathbf{R}$ , as well as the equivalent matrix  $\tilde{\mathbf{R}}$ , it is straightforward to establish a simulation algorithm for the conditional simulation of  $\underline{x}_1^* | \underline{x}_2^* = \mathbf{h}$  (using the same notation and dimensions as in the previous section). The first step consists of partitioning the matrix  $\tilde{\mathbf{R}}$  as in Eq. (4.44) and estimation of the auxiliary vector  $\tilde{\mathbf{h}} = \left[ \Phi^{-1} \left( F_{\underline{x}_{n+1}}(h_{n+1}) \right), \dots, \Phi^{-1} \left( F_{\underline{x}_m}(h_m) \right) \right]^T$ . Next, the conditional mean  $\tilde{\boldsymbol{\mu}}$  and covariance matrix  $\tilde{\boldsymbol{\Sigma}}$  of the auxiliary conditional Gaussian distribution have to be determined. Given the latter, estimate a matrix  $\tilde{\mathbf{B}}$  such that  $\tilde{\mathbf{B}} \tilde{\mathbf{B}}^T = \tilde{\boldsymbol{\Sigma}}$  (see section 4.1.3) and generate an auxiliary vector  $\underline{z}_1^*$  by  $\underline{z}_1^* = \tilde{\boldsymbol{\mu}} + \tilde{\mathbf{B}} \underline{\mathbf{w}}$ , where  $\underline{\mathbf{w}} = [w_1, \dots, w_n]^T$  is an i.i.d. vector with  $\mathcal{N}(0,1)$ . Finally, obtain the conditioned RVs  $\underline{x}_1^* | \underline{x}_2^* = \mathbf{h}$  via mapping  $\underline{z}_1^*$  to the actual domain using the corresponding ICDFs of  $F_{\underline{x}_1}, \dots, F_{\underline{x}_n}$ . This operation reads,  $\underline{x}_1^* = [\underline{x}_1^*, \dots, \underline{x}_n^*]^T = \left[ F_{\underline{x}_1}^{-1} \left( \Phi(\underline{z}_1^*) \right), \dots, F_{\underline{x}_n}^{-1} \left( \Phi(\underline{z}_n^*) \right) \right]^T$ . Apparently, when  $n = 1$ , the above equation reduces to  $\underline{x}_1^* = F_{\underline{x}_1}^{-1} \left( \Phi(\underline{z}_1^*) \right)$ , which in turn allows its expression in terms of probability  $u_1$ , i.e.,  $\underline{x}_1^* = F_{\underline{x}_1}^{-1}(u_1)$ . This formulation is identical to Eq. (4.32) and can be used to derive the conditional quantiles of interest.

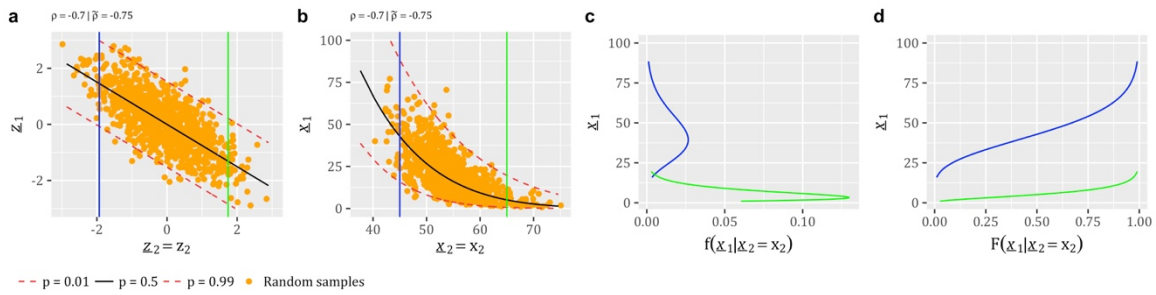
#### 4.2.3 Numerical examples

To elaborate on the previously described Nataf-based conditional sampling method, as well as on the derivation of conditional quantiles (i.e., Eq. (4.32)), let us consider a bivariate example of  $\underline{x}_1 \sim \mathcal{G}(2, 10)$  and  $\underline{x}_2 \sim \mathcal{LN}(0.10, 4)$  for various values of correlation coefficient  $\rho = \rho_{12} \in \{-0.90, -0.70, -0.50, 0.0, 0.50, 0.70, 0.90\}$ . Apparently, in the bivariate case  $\underline{x}_1^* \equiv \underline{x}_1$  and  $\underline{x}_2^* \equiv \underline{x}_2$ , hence for the sake of simplicity, the *star* notation will be hereafter omitted. Nonetheless, in order to construct a proper conditional distribution it is required to estimate the equivalent correlation coefficients, which for these cases they were found,  $\tilde{\rho} = \tilde{\rho}_{12} = \{-0.974, -0.754, -0.536, 0.0, 0.522, 0.728, 0.931\}$ . Figure 4.8 concerns the first case of  $\rho = -0.9$  and particularly panel a) depicts the 0.01, 0.5 (i.e., median) and 0.99 probability quantiles of the auxiliary conditional Gaussian distribution of  $\underline{z}_1 | \underline{z}_2$  for various values of  $\underline{z}_2 = \Phi^{-1}(F_{\underline{x}_2}(x_2))$ , which is also homoscedastic. Panel b) illustrates the same probability quantiles for the actual Nataf-based conditional distribution  $\underline{x}_1 | \underline{x}_2$  for  $\underline{x}_2 = x_2$ . Interestingly, this plot is

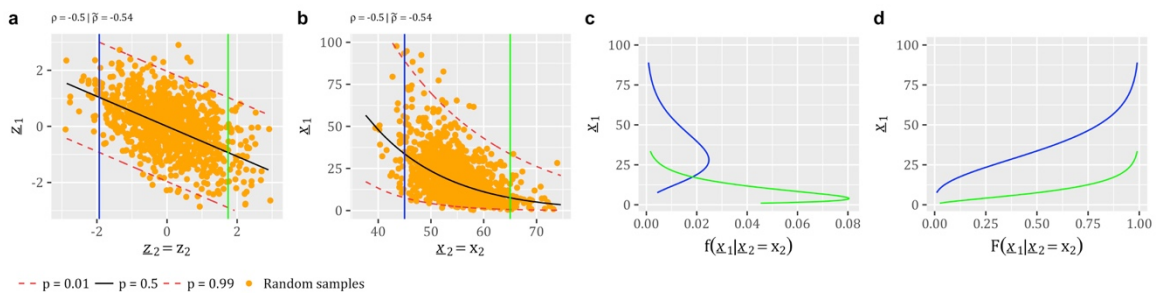
characterized by a non-linear heteroscedastic behavior, which highlights the ability of the model to capture a wider range of dependence forms. Moreover, the plots, depict some randomly generated samples from the corresponding conditional distributions. Finally, panel c) and d) visualize the complete conditional PDF and CDF of  $\underline{x}_1|\underline{x}_2$  respectively for two arbitrary selected values of  $\underline{x}_2$  (i. e.,  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ ), which emphasize on the variety of distribution shapes (e.g., bell- or J-shaped) that can arise from the Nataf-conditional distribution model. Similarly, to the previous analysis, **Figure 4.9** to **Figure 4.14** provide further examples and illustrations of the conditional model, for different values of correlation coefficient  $\rho$ .



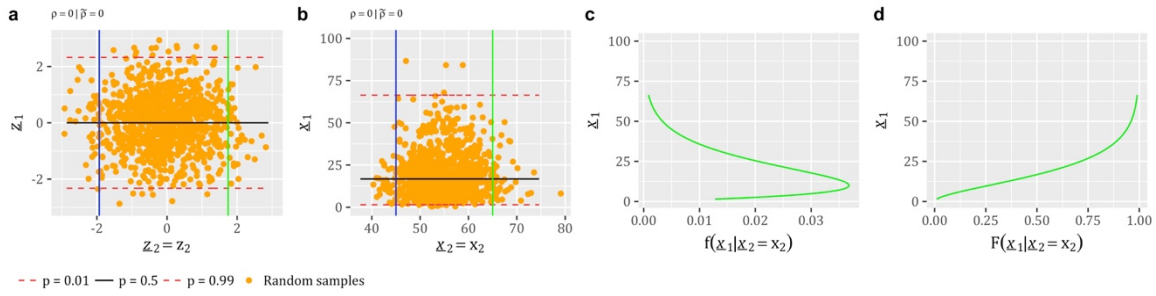
**Figure 4.8** | Bivariate example of  $\underline{x}_1 \sim \mathcal{G}(2, 10)$  and  $\underline{x}_2 \sim \mathcal{LN}(0.10, 4)$  with  $\rho = -0.9$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ .



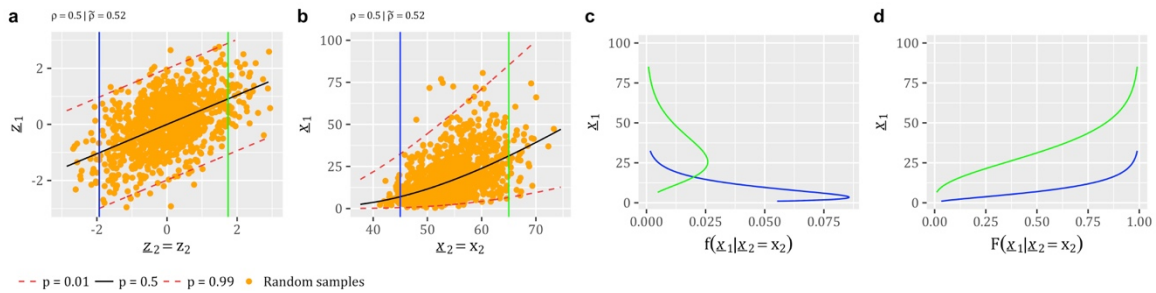
**Figure 4.9** | Bivariate example of  $\underline{x}_1 \sim \mathcal{G}(2, 10)$  and  $\underline{x}_2 \sim \mathcal{LN}(0.10, 4)$  with  $\rho = -0.7$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ .



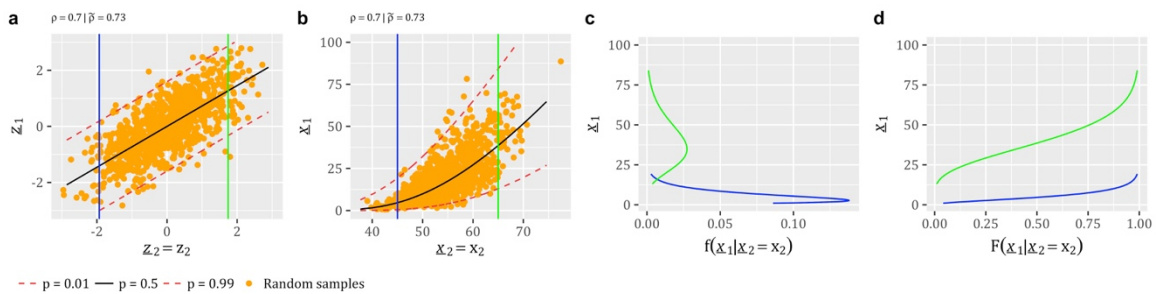
**Figure 4.10** | Bivariate example of  $\underline{x}_1 \sim \mathcal{G}(2, 10)$  and  $\underline{x}_2 \sim \mathcal{LN}(0.10, 4)$  with  $\rho = -0.5$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $\underline{x}_2 = 45$  and  $\underline{x}_2 = 65$ .



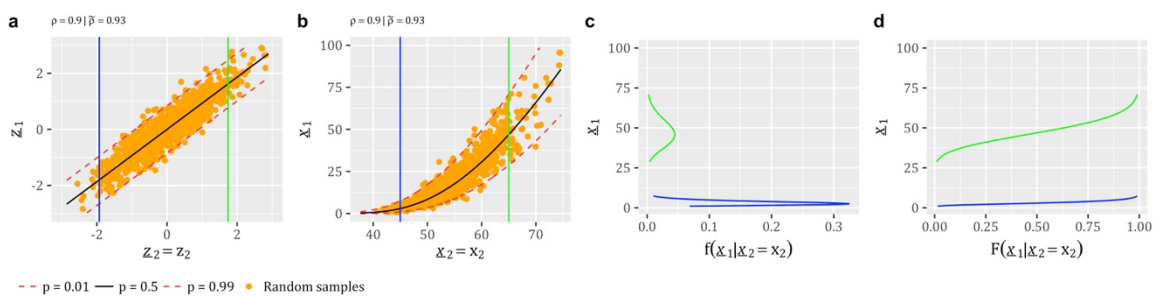
**Figure 4.11** | Bivariate example of  $x_1 \sim \mathcal{G}(2, 10)$  and  $x_2 \sim \mathcal{LN}(0.10, 4)$  with  $\rho = 0.0$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $x_2 = 45$  and  $x_2 = 65$ .



**Figure 4.12** | Bivariate example of  $x_1 \sim \mathcal{G}(2, 10)$  and  $x_2 \sim \mathcal{LN}(0.10, 4)$  with  $\rho = 0.5$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $x_2 = 45$  and  $x_2 = 65$ .

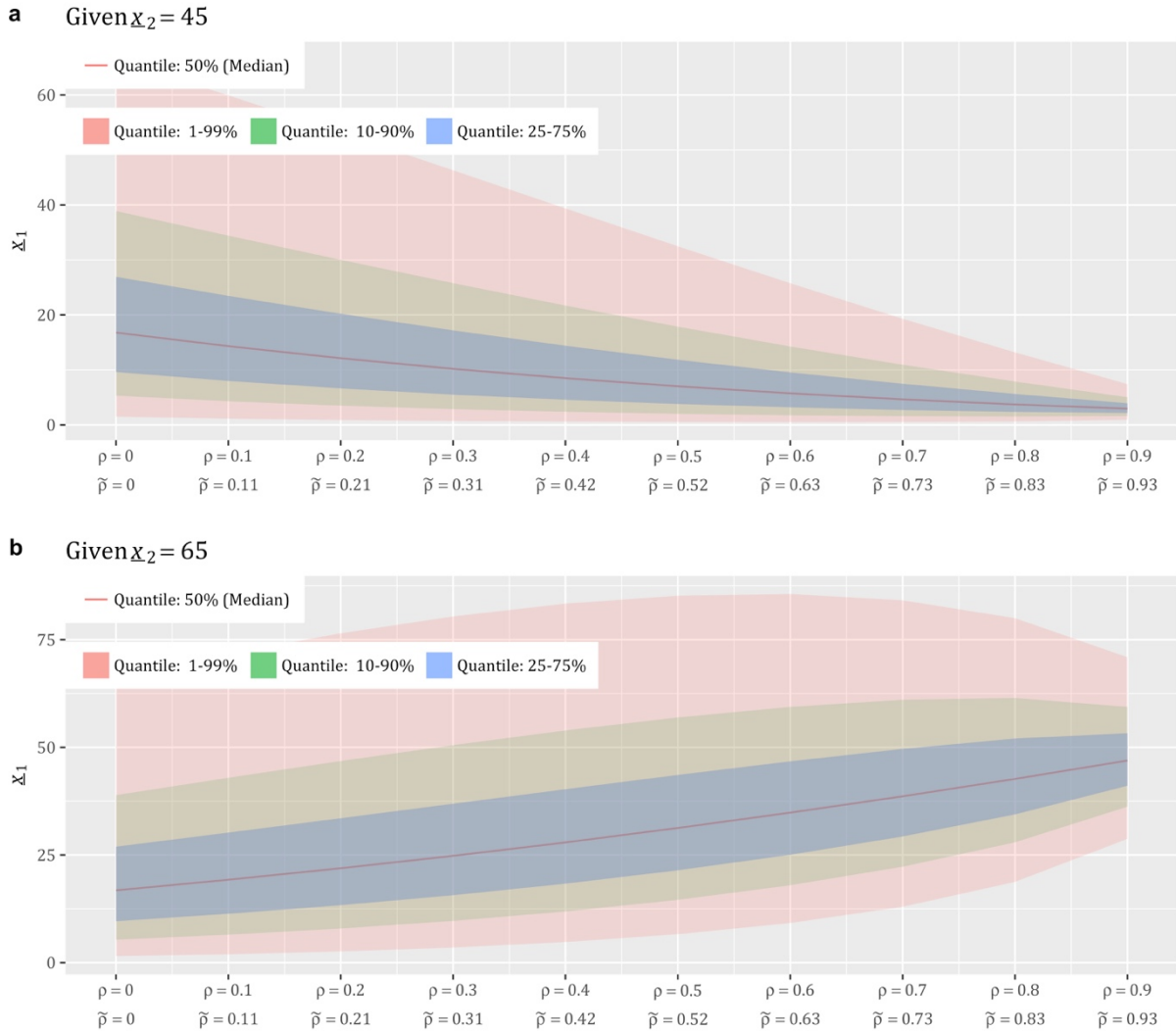


**Figure 4.13** | Bivariate example of  $x_1 \sim \mathcal{G}(2, 10)$  and  $x_2 \sim \mathcal{LN}(0.10, 4)$  with  $\rho = 0.7$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $x_2 = 45$  and  $x_2 = 65$ .



**Figure 4.14** | Bivariate example of  $x_1 \sim \mathcal{G}(2, 10)$  and  $x_2 \sim \mathcal{LN}(0.10, 4)$  with  $\rho = 0.9$ . Probability quantiles in the a) Gaussian domain b) actual domain. Conditional c) PDF and d) CDF for  $x_2 = 45$  and  $x_2 = 65$ .

A final demonstration, also related with the previous example, concerns the derivation of 0.01, 0.10, 0.25, 0.50, 0.75, 0.90 and 0.99 probability quantiles for two arbitrary selected values of  $x_2$  (i.e.,  $x_2 = 45$  and  $x_2 = 65$ ) for a sequence of values of  $\rho = \rho_{12} \in \{0.0, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90\}$ . **Figure 4.15** depicts the effect of dependence parameter (i.e., Pearson’s correlation) on the calculation of latter quantiles, through the visualization of the derived quantiles as a function of  $\rho$ .



**Figure 4.15** | Effect of dependence parameter (i.e., Pearson’s correlation coefficient) on the derived quantiles of  $x_1|x_2$ ; visualized as a function of  $\rho$  for a)  $x_2 = 45$  and b)  $x_2 = 65$ .

### 4.3 NATAF-BASED STOCHASTIC PROCESSES WITH ARBITRARY MARGINAL DISTRIBUTIONS AND CORRELATION STRUCTURE

In a recent work, *Tsoukalas et al. [2018e]* highlighted the need for generalized generation schemes, which are able to represent processes from *any* distribution and *any* correlation structure. This has been also regarded as a shift in classical stochastic modelling, emphasizing on the reproduction of a finite set of *essential* statistical characteristics [cf. *Matalas and Wallis, 1976*], estimated from the historical data.

An effective and efficient handling of this requirement is offered by the so-called Nataf-based models [*Tsoukalas et al., 2017a, 2018e, 2018b, 2018d*]. As the name suggests, these are built



upon the idea by *Nataf* [1962] and the associated concept of NDM. Using a similar rationale, it is possible to establish stochastic processes with any target marginal distribution and correlation structure (expressed in terms of Pearson's correlation coefficient) through the mapping (similar to Eq. (4.10)) of an appropriately specified auxiliary (stationary or cyclostationary) standard Gaussian process (Gp) with zero mean and unit variance, to which an *equivalent* correlation structure is assigned (see details below). The mapping operation is typically a non-linear function, often implemented through the inverse cumulative distribution function (ICDF). These approaches can be viewed as Gaussian copula-based schemes (since they rely on the mapping of a Gaussian process) or non-linear versions of the classic (i.e., Gaussian) linear stochastic schemes [Tsoukalas et al., 2018d]. Nataf-based stochastic models or approaches with common rationale, have been used within the domain of operations research [e.g., *Cario and Nelson*, 1996; *Biller and Nelson*, 2003] and probabilistic engineering mechanics [e.g., *Grigoriu*, 1998; *Deodatis and Micaletti*, 2001]. Their employment within hydrological sciences was, until recently, formally unexplored, yet, *post factum* linked with other approaches in hydrological domain (see section 4.6, as well as [Tsoukalas et al., 2018e, 2018d]). More specifically, *Cario and Nelson* [1996] and *Biller and Nelson* [2003] employed this notion for the simulation of stationary non-Gaussian univariate and multivariate autoregressive (AR) processes respectively. In this vein, herein, by building-upon the aforementioned two paradigms, the concept of NDM is being aligned for the description of stationary (not necessarily AR) univariate and multivariate processes, as well as further extended for the particularly interesting cases, from a hydrological point of view, of univariate and multivariate cyclostationary processes [Tsoukalas et al., 2017a, 2018e].

In this section we briefly discuss the theoretical background and key implementation steps of Nataf-based, simulation schemes, also providing guidelines for its *optimal* use. For convenience, we first present the most involved modelling case of multivariate cyclostationary processes, and next deal with the simpler case of stationary processes.

#### 4.3.1 Multivariate and univariate cyclostationary processes

In general, cyclostationarity is regarded as a special type of non-stationarity that implies a cyclic switching on the marginal and joint characteristics of the process over a period (e.g., year). To elaborate, let  $\{\underline{x}_{s,n}\}$  be a  $m$ -dimensional multivariate cyclostationary process. Each individual process  $\{\underline{x}_{s,n}^i\}$  is consisted of  $s = 1, \dots, S$  sub-periods (e.g., months), while  $n \in \mathbb{Z}^>$ , denotes the time index. The sub-period (i.e., season) that corresponds to a time step  $n$  may be recovered by  $s = n \bmod(S)$ , while when  $n \bmod(S) = 0$ ,  $s = S$ . Furthermore, due to cyclostationarity, each one of them is characterized by seasonally varying (herein referred to as target) marginal distributions  $F_{\underline{x}_s^i} = P(\underline{x}_s^i \leq x)$ , while their correlation structure is expressed through the Pearson's correlation coefficient  $\rho_{s,s-\tau}^{i,j} := \text{Corr}[\underline{x}_s^i, \underline{x}_{s-\tau}^j]$ , where  $\tau$  denotes the time lag (the index  $n$  is omitted for simplicity). Also let  $\{\underline{z}_{s,n}\}$  denote an auxiliary  $m$ -dimensional cyclostationary standard Gp with  $\underline{z}_s^i \sim \mathcal{N}(0,1)$ . Due to cyclostationarity, the Gp is completely defined by its correlation structure, which is expressed through the so-called equivalent correlation coefficients  $\tilde{\rho}_{s,s-\tau}^{i,j} := \text{Corr}[\underline{z}_s^i, \underline{z}_{s-\tau}^j]$ . We note that both the target and the auxiliary Gaussian process can be expressed in sup-period/period notation, i.e.,  $\{\underline{x}_{s,t}\}$  and  $\{\underline{z}_{s,t}\}$ , where  $s = 1, \dots, S, 1, \dots, S, \dots$  denotes the season (e.g., month) and  $t = 1 + (n - s)/S$ , denotes the period (e.g., year). Herein, for convenience, we will employ this notation. Nonetheless, by employing the concept of NDM and by analogy to RVs case (section 4.1.2), the target process  $\{\underline{x}_{s,t}\}$  can be established through the auxiliary process  $\{\underline{z}_{s,t}\}$  via the mapping function,

$$\underline{x}_{s,t}^i = F_{\underline{x}_s^i}^{-1} \left( \Phi(\underline{z}_{s,t}^i) \right) \quad (4.33)$$

where  $F_{\underline{x}_s^i}^{-1}$  denotes the ICDF of  $F_{\underline{x}_s^i}$  and  $\Phi(\cdot)$  denotes the cumulative density function (CDF) of the standard Gaussian distribution. This mapping, eventually relates the target correlation coefficients  $\rho_{s,s-\tau}^{i,j}$  with the equivalent correlation coefficients  $\tilde{\rho}_{s,s-\tau}^{i,j}$  of the auxiliary process [Tsoukalas et al., 2017a, 2018e]. Specifically, since Eq. (4.33) holds true, we can write,

$$\rho_{s,s-\tau}^{i,j} = \text{Corr}[\underline{x}_s^i, \underline{x}_{s-\tau}^j] = \text{Corr} \left[ F_{\underline{x}_s^i}^{-1} \left( \Phi(\underline{z}_s^i) \right), F_{\underline{x}_{s-\tau}^j}^{-1} \left( \Phi(\underline{z}_{s-\tau}^j) \right) \right] \quad (4.34)$$

Using the definition of Pearson's correlation coefficient, we can also write,

$$\rho_{s,s-\tau}^{i,j} = \text{Corr}[\underline{x}_s^i, \underline{x}_{s-\tau}^j] = \frac{\mathbb{E}[\underline{x}_s^i \underline{x}_{s-\tau}^j] - \mathbb{E}[\underline{x}_s^i] \mathbb{E}[\underline{x}_{s-\tau}^j]}{\sqrt{\text{Var}[\underline{x}_s^i] \text{Var}[\underline{x}_{s-\tau}^j]}} \quad (4.35)$$

where  $\mathbb{E}[\underline{x}_s^i]$ ,  $\mathbb{E}[\underline{x}_{s-\tau}^j]$  and  $\text{Var}[\underline{x}_s^i]$ ,  $\text{Var}[\underline{x}_{s-\tau}^j]$  denote the mean and variance of  $\underline{x}_s^i$  and  $\underline{x}_{s-\tau}^j$  respectively which are known from the corresponding distributions  $F_{\underline{x}_s^i}$  and  $F_{\underline{x}_{s-\tau}^j}$  and have to be finite. Subsequently, the first cross-product moment of  $\underline{x}_s^i$  and  $\underline{x}_{s-\tau}^j$  can be expressed as,

$$\begin{aligned} \mathbb{E}[\underline{x}_s^i \underline{x}_{s-\tau}^j] &= \mathbb{E} \left[ F_{\underline{x}_s^i}^{-1} \left( \Phi(\underline{z}_s^i) \right) F_{\underline{x}_{s-\tau}^j}^{-1} \left( \Phi(\underline{z}_{s-\tau}^j) \right) \right] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}_s^i}^{-1} \left( \Phi(\underline{z}_s^i) \right) F_{\underline{x}_{s-\tau}^j}^{-1} \left( \Phi(\underline{z}_{s-\tau}^j) \right) \varphi_2(\underline{z}_s^i, \underline{z}_{s-\tau}^j, \tilde{\rho}_{s,s-\tau}^{i,j}) d\underline{z}_s^i d\underline{z}_{s-\tau}^j \end{aligned} \quad (4.36)$$

where  $\varphi_2(\underline{z}_s^i, \underline{z}_{s-\tau}^j, \tilde{\rho}_{s,s-\tau}^{i,j})$  is the bivariate standard normal probability density function. Thus, finally we obtain,

$$\rho_{s,s-\tau}^{i,j} = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}_s^i}^{-1} \left( \Phi(\underline{z}_s^i) \right) F_{\underline{x}_{s-\tau}^j}^{-1} \left( \Phi(\underline{z}_{s-\tau}^j) \right) \varphi_2(\underline{z}_s^i, \underline{z}_{s-\tau}^j, \tilde{\rho}_{s,s-\tau}^{i,j}) d\underline{z}_s^i d\underline{z}_{s-\tau}^j - \mathbb{E}[\underline{x}_s^i] \mathbb{E}[\underline{x}_{s-\tau}^j]}{\sqrt{\text{Var}[\underline{x}_s^i] \text{Var}[\underline{x}_{s-\tau}^j]}} \quad (4.37)$$

Eq. (4.37) shows that  $\rho_{s,s-\tau}^{i,j}$  is a function of the equivalent correlation coefficient  $\tilde{\rho}_{s,s-\tau}^{i,j}$ , and the target (i.e., given) distributions  $F_{\underline{x}_s^i}$  and  $F_{\underline{x}_{s-\tau}^j}$ . i.e.,

$$\rho_{s,s-\tau}^{i,j} = \mathcal{F} \left( \tilde{\rho}_{s,s-\tau}^{i,j} \mid F_{\underline{x}_s^i}, F_{\underline{x}_{s-\tau}^j} \right) \quad (4.38)$$

where  $\mathcal{F}(\cdot)$  denotes an abbreviation of the function defined by Eq. (4.37). Seemingly, for a univariate cyclostationary processes Eq. (4.37) simplifies to,

$$= \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}_s}^{-1}(\Phi(z_s)) F_{\underline{x}_{s-\tau}}^{-1}(\Phi(z_{s-\tau})) \varphi_2(z_s, z_{s-\tau}, \tilde{\rho}_{s,s-\tau}) dz_s dz_{s-\tau} - E[\underline{x}_s] E[\underline{x}_{s-\tau}]}{\sqrt{\text{Var}[\underline{x}_s] \text{Var}[\underline{x}_{s-\tau}]}} \quad (4.39)$$

and consequently Eq. (4.38) reads,

$$\rho_{s,s-\tau} = \mathcal{F}(\tilde{\rho}_{s,s-\tau} | F_{\underline{x}_s}, F_{\underline{x}_{s-\tau}}) \quad (4.40)$$

### 4.3.2 Multivariate and univariate stationary processes

A similar relationship can be established between a target multivariate stationary process  $\{\underline{x}_t\}$  and an auxiliary multivariate stationary standard Gp  $\{\underline{z}_t\}$ . Particularly, let  $\{\underline{x}_t\}$  be comprised by  $m$  univariate stationary processes  $\{x_t^i\}$ , indexed using  $t \in \mathbb{Z}^>$ . Furthermore, let each one described by a target CDF,  $F_{\underline{x}^i} = P(x^i \leq x)$  and let their correlation structure be expressed by  $\rho_{t,t+\tau}^{i,j} := \text{Corr}[\underline{x}_t^i, \underline{x}_{t+\tau}^j]$ . Similarly, the process  $\{\underline{z}_t\}$  is a  $m$ -dimensional stationary standard Gp, with equivalent correlation structure,  $\tilde{\rho}_{t,t+\tau}^{i,j} := \text{Corr}[\underline{z}_t^i, \underline{z}_{t+\tau}^j]$ .

Using a similar rationale to the cyclostationary case, each target process  $\{x_t^i\}$  is established via  $\{\underline{z}_t\}$  by,  $x_t^i = F_{\underline{x}^i}^{-1}(\Phi(z_t^i))$ . Following the same reasoning with the previous section, the relationship between the target and equivalent correlation coefficients reads [e.g., *Biller and Nelson, 2003; Tsoukalas et al., 2018d*],

$$= \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}^i}^{-1}(\Phi(z_t^i)) F_{\underline{x}^j}^{-1}(\Phi(z_{t+\tau}^j)) \varphi_2(z_t^i, z_{t+\tau}^j, \tilde{\rho}_{t,t+\tau}^{i,j}) dz_t^i dz_{t+\tau}^j - E[x^i] E[x^j]}{\sqrt{\text{Var}[x^i] \text{Var}[x^j]}} \quad (4.41)$$

Which for simplicity is abbreviated as,

$$\rho_{t,t+\tau}^{i,j} = \mathcal{F}(\tilde{\rho}_{t,t+\tau}^{i,j} | F_{\underline{x}^i}, F_{\underline{x}^j}) \quad (4.42)$$

In the univariate case the previous equations simplify to,

$$\rho_\tau = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}}^{-1}(\Phi(z_t)) F_{\underline{x}}^{-1}(\Phi(z_{t+\tau})) \varphi_2(z_t, z_{t+\tau}, \tilde{\rho}_\tau) dz_t dz_{t+\tau} - (E[\underline{x}])^2}{\text{Var}[\underline{x}]} \quad (4.43)$$

and abbreviated as,

$$\rho_\tau = \mathcal{F}(\tilde{\rho}_\tau | F_{\underline{x}}) \quad (4.44)$$

The relationship  $\mathcal{F}(\cdot)$  imply that the correlation structure of the target process depends on the target distributions and the equivalent correlation structure of the auxiliary Gp. We underline that the term *equivalent* is used to highlight the fact that the correlation coefficients of the target

process and those of the auxiliary  $G_p$ , rarely coincide (due to the non-linear mapping operation), since the lemmas of section 4.1.2 also hold for the case of processes.

### 4.3.3 Selection of the target marginal distributions and correlation structures

As already explained, Nataf-based models can be used for the simulation of processes with arbitrary (continuous, discrete or mixed-type) marginal distributions and valid correlation structures, provided that their combination is feasible (i.e., leads to a positive definite correlation structure) and the variance of the distributions is finite (which is the typical assumption when modelling hydrometeorological processes; see section 4.3.9).

Regarding the marginal distributions, and in contrast to the classical working paradigm of stochastic hydrology, it is stressed that by design, Nataf-based models do not aim at resembling the process's moments; in fact, they aim to simulate processes with target, *a priori* specified, distributions, in order to fully describe its marginal properties (cf. discussion by *Tsoukalas et al. [2018e]*). In this respect, questions about skewness handling or *how many moments should be reproduced for approximating the distribution of a specific process?* become out of interest.

For instance, within Nataf-based schemes, simulating a process following Gamma or Log-Normal distribution requires the identification of just two parameters (shape and scale), which can be easily determined by straightforward methods. Even the classical method of product moments, would ensure reliable estimations, since in these specific cases it only requires computations up to second order (a safe upper bound as argued by *Lombardo et al. [2014]*).

In a more general context, the assignment of a specific distribution model for each modelled process is not a straightforward task, since the true distribution will always be unknown. For a given data sample one can fit a plethora of distributions, combined with different parameter estimation procedures, and use typical statistical tests to assess the *optimal* scheme.

The Nataf-based approach offers the flexibility to employ robust fitting methods for parameter estimation, that rely on alternative notions, such as, probability weighted moments [*Greenwood et al., 1979*], L-moments [*Hosking, 1990*] or maximum likelihood. In our view, this is a major advantage, since it can avoid the data-driven estimation of high-order product moments (e.g., kurtosis or higher), which it is well known that are prone to sample uncertainties and bias [*Matalas, 1967 p. 945; Lombardo et al., 2014*].

In any case, particularly when the historical samples are short or not so much reliable, the selection of the most suitable distribution may be supported by hydrological evidence. For instance, one may take advantage of the statistical behavior of the underlying processes in the broader area, as validated by large-scale regional studies [e.g., *Blum et al., 2017*].

The specification of the above inputs is not a straightforward decision neither it is advised to be made automatically, especially when considering the flexibility of Nataf-based methods regarding the selection of the distribution function and hence the fitting method. Overall, in operation context, the modeler could (and should) account for multilateral information, based both on historical data and expert judgment, in order to establish a realistic formulation of the stochastic simulation model.

Moving to the correlation structures, classical stochastic modelling strategies are designed to reproduce a limited number of low-order dependence metrics in space and time, typically expressed in terms of Pearson's correlation coefficients. Actually, most of them still follow the specifications posed by *Matalas and Wallis [1976]*, thus aiming to reproduce just two dependencies, i.e. lag-1 autocorrelations and lag-0 cross-correlations. It is remarked that

herein, the term spatial correlation will denote any dependence between different processes, either referring to different geographical locations or not.

More modern approaches suggest the use of theoretical models for the mathematical description of the auto- and cross-dependence structures that span over any lag [e.g., [Gneiting, 2000](#); [Koutsoyiannis, 2000, 2016](#); [Gneiting and Schlather, 2004](#)]. These typically concern stationary processes, and are based on the notions of correlation, spectrum or variance over aggregated time scales, which are all interrelated [see, [Beran, 1994](#); [Koutsoyiannis, 2016](#)]. The use of theoretical dependence models instead of sample statistics is mostly implied from the significant uncertainties and biases of data-driven estimates.

Arguably, the most popular type of theoretical dependence models are correlation-based ones. These can be further classified to full spatio-temporal models [[Chilès and Delfiner, 1999](#); [Gneiting et al., 2010](#); [Genton and Kleiber, 2015](#)], which simultaneously model the auto- and cross-correlation structure of the process, and separable [[Rodríguez-Iturbe and Mejía, 1974](#); [Mardia and Goodall, 1993](#); [Genton, 2007](#)], which describe the two correlation structures independently, as product of two functions (i.e., one for the spatial and one for the temporal component).

Throughout this Thesis, and without loss of generality (since alternative models can be used), we will employ the separable approach for stationary process. Specifically, we model directly the lag-0 contemporaneous cross-correlations of the processes, while the auto-dependence structure of each individual stationary process is modelled using the two-parameter Cauchy-type autocorrelation structure (CAS), introduced by [Koutsoyiannis \[2000\]](#), i.e.,  $\rho_{\tau}^{\text{CAS}}(\kappa, \beta) = (1 + \kappa\beta\tau)^{-1/\beta}$ ,  $\tau \geq 0$  where  $\beta \geq 0$  and  $\kappa > 0$  are model parameters. By construction, CAS can resemble a wide spectrum of processes, characterized by both short- and long-range dependence, i.e., SRD and LRD (for more details see also, section 5.2). SRD refers to a stochastic process with a weak autocorrelation structure (e.g., exponential) that decays rapidly, while LRD implies the exact opposite (see section 2.2). These properties and its parsimonious character (as the model has only two parameters), make CAS a good candidate model for modelling hydrometeorological processes. Regarding parameter identification, the most straightforward option is to fit CAS to the empirical estimates of autocorrelation coefficients. However, this simple approach neglects the estimator's biases [e.g., [Marriott and Pope, 1954](#); [Beran, 1994](#); [Koutsoyiannis, 2016](#)], which are considered to be significant in the presence of LRD and for large time lags (due to small sample sizes). In such cases, it may be advantageous to explicitly account for bias by using alternative robust parameter identification procedures, such as the climacogram [e.g., [Dimitriadis and Koutsoyiannis, 2015](#); [Koutsoyiannis, 2016](#)], or even through empirical approaches, accounting for regional information and user expertise [[Efstratiadis et al., 2014a](#)].

In summary, the combined use of Nataf-based models along with theoretical distribution functions and theoretical correlation structures (e.g., CAS), offers several advantages, such as, the easy alternative scenario exploration (by perturbing the models parameters), regional transferability (through spatial interpolation), improved model stability (since a valid correlation structure owes to be positive definite; a fact guaranteed by a proper theoretical model), and the decoupling of parameter identification (involving the parameters of the distribution model and the theoretical correlation structure) with the generation mechanism.

#### 4.3.4 The auxiliary Gaussian processes

In order to deploy a Nataf-based stochastic simulation scheme, it is required to employ and simulate realizations from an auxiliary Gp. Regardless this choice, it is important to estimate its parameters using the equivalent correlation coefficients. This way, the realizations of the auxiliary Gp will preserve the equivalent correlation coefficients, which in turn, after the mapping procedure, reproduces the target stochastic structure. The Gp could be modeled using simple mechanisms, especially in the case of univariate stationary processes, such as the well-known decomposition-based simulation scheme discussed in section 4.1.3 or by utilizing more advanced schemes.

An apparent option, extensively discussed in the following Chapters, is the use of Gaussian linear stochastic models (also called time series models). Characteristic examples, adapted from operations research, are the works of *Cario and Nelson [1996]* and *Biller and Nelson [2003]*, who used as an auxiliary Gp, univariate and multivariate stationary AutoRegressive (AR) processes, respectively. The resulting Nataf-based models are termed AutoRegressive To anything (ARTA) and Vector AutoRegressive To Anything (VARTA). A notable difference of these works compared to the approach described herein, lies in the fact that the previous works did not employ the notion of theoretical correlation structures. This implies that the order  $p$  of the associated AR model dictates the correlation structure of the process to simulate. This may be also the reason for the typical use of low order models. On the other hand, if the auto-correlation structure has been *a priori* specified (e.g., using CAS), it is possible to employ high-order models (e.g.,  $AR(p)$ ), even multivariate, without sacrificing parsimony. In this case, the order of the Gp model solely controls the degree of resemblance of the correlation structure up to the desired lag  $\tau$  (since a higher order model provides more flexibility), while the associated model's parameters can be viewed as internal coefficients (for bivariate examples using high-order AR models, see section 5.6.2). In the water resources domain, a comprehensive treatment of multivariate and univariate Nataf-based schemes, based on stationary and cyclostationary Gaussian linear stochastic models, is presented in Chapters 5 and 6 respectively. A simple simulation example using the classical decomposition scheme is given in section 4.3.7.

#### 4.3.5 Estimation of the equivalent correlation coefficients

An important step of any Nataf-based simulation scheme, is the identification of equivalent correlation coefficients, which in turn allows the reproduction of the target correlation structure. It is reminded that the equivalent correlations (i.e., in the Gaussian domain) typically differ from the target ones (in the real domain), and they are estimated on the basis of the NDM approach. Nevertheless, regardless the case, multivariate or univariate, stationary or cyclostationary, the identification of equivalent correlation coefficients (i.e., by inverting the corresponding  $\mathcal{F}(\cdot)$  relationship – see sections 4.3.1 and 4.3.2) can be accomplished in a pairwise basis using the methods of section 4.5. Particularly, the proposed algorithm of section 4.5.1 can significantly simplify and thus accelerate the identification procedure since it establishes a functional relationship between the target and equivalent correlation coefficients that can be used multiple times (it also avoids the use of integration methods). For example, in the simple case of a univariate stationary process, this procedure has to be employed only once, to establish the relationship  $\rho_\tau = \mathcal{F}(\tilde{\rho}_\tau | F_{\underline{x}})$  and then reused multiple times for different target values of  $\rho_\tau$ .

#### 4.3.6 Mapping auxiliary processes to the actual domain

After simulating a realization of the auxiliary processes (i.e.,  $\underline{z}_t, \underline{z}_t, \underline{z}_{s,t}$  or  $\underline{z}_{s,t}$ ), the last step is its mapping to the actual domain (i.e.,  $\underline{x}_t, \underline{x}_t, \underline{x}_{s,t}$  or  $\underline{x}_{s,t}$ ), through the ICDF(s). It is noted that this procedure is implemented for each individual process and season (in the case of multivariate processes and cyclostationarity). Due to the use of the ICDF, as well as the use of equivalent coefficients of correlation within the auxiliary Gaussian model, final realization will preserve both the target marginal distributions (for all seasons and variables) as well as the target correlation structure.

#### 4.3.7 Brief overview via a step-by-step procedure

For a given stochastic process (univariate case), or a set of processes (multivariate case) to simulate, the required methodological steps of any Nataf-based model are:

**Step 1.** Identify the type (i.e., stationary or cyclostationary) of the processes, accounting for process properties and the time scale of simulation.

**Step 2.** Based on the available information (e.g., historical data), as well as the user expertise, assign appropriate target marginal distributions to all processes and identify the target correlation structure, in time and (case of multivariate simulation) space.

**Step 3.** Select a suitable linear stochastic model to simulate the auxiliary Gp.

**Step 4.** Estimate the equivalent correlation coefficients for all pairs of variables that are required by the parameter estimation procedure of the auxiliary model, i.e. Gp.

**Step 5.** Estimate the parameters of the Gp model through the equivalent correlation coefficients.

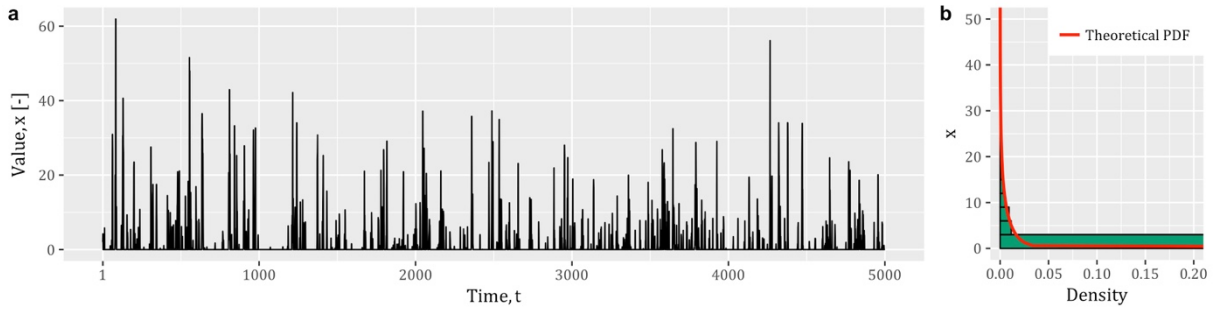
**Step 6.** Generate a synthetic time series by employing the Gp (i.e.,  $\underline{z}_t, \underline{z}_t, \underline{z}_{s,t}$  or  $\underline{z}_{s,t}$ ).

**Step 7.** Map the auxiliary (i.e. Gaussian) time series to the actual domain in order to attain a realization of the target process (i.e.,  $\underline{x}_t, \underline{x}_t, \underline{x}_{s,t}$  or  $\underline{x}_{s,t}$ ).

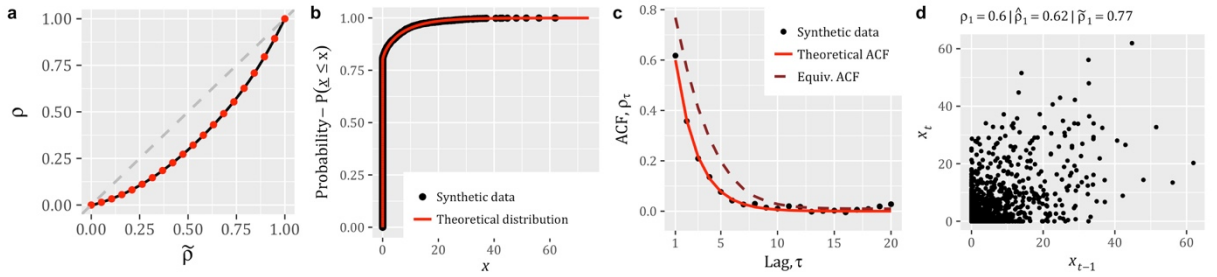
#### 4.3.8 Numerical examples

As mentioned in the previous section, the decomposition-based simulation scheme of section 4.1.3, which is typically employed for the simulation of correlated RVs, can be easily used for the simulation of stationary and non-stationary, non-Gaussian processes. The simulation procedure is exactly the same, with the only difference that given an autocorrelation function  $\rho_\tau, \tau \geq 0$ , and assuming that we wish to generate a process with  $T$  time steps, the elements of the now  $T \times T$  matrix  $\mathbf{R}$  are being determined by  $\mathbf{R}_{[i,j]} = \rho_{|i-j|}$ . Particularly, in the case of stationarity, the marginal distribution of the process  $\underline{x}_t$  is constant for any  $t \in \mathbb{Z}$ , i.e.,  $F_{\underline{x}_t} \equiv F_{\underline{x}}$ . As first example, we employ a simple theoretical autocorrelation structure, that of the first order autoregressive model (i.e., AR(1)), which is given by,  $\rho_\tau = \alpha_1^{|\tau|}$ , where  $\alpha_1$  is a model parameter in  $[-1,1]$ . Assuming  $\alpha_1 = 0.6$  and a zero-inflated discrete-continuous marginal distribution (as in the previous example of section 4.1.4.2) with  $p_0 = 0.8$  and  $G_{\underline{x}}(x) = F_G(x; 0.7,10)$  for the continuous distribution part (i.e., a Gamma distribution) we simulate one realization of 5 000 time steps. Given this information, the first step is to identify the relationship,  $\rho_\tau = \mathcal{F}(\tilde{\rho}_\tau | F_{\underline{x}})$  (see Figure 4.17a), and subsequently invert it, in order to obtain the equivalent correlation coefficients; hence the structure of the auxiliary Gaussian process (brown dashed line Figure 4.17c). Next, a realization of the auxiliary process is generated, and

then transformed using a mapping function similar to Eq. (4.33). Figure 4.16a depicts the simulated sequence, as well as Figure 4.16b a comparison between the theoretical PDF and the empirical histogram. Furthermore, Figure 4.17b-d synopsise some of the key marginal and stochastic properties of the simulated time series, which arguably closely resemble the theoretical ones. However, it is noted that this scheme was mainly employed for demonstration purposes due to the fact that often suffers from numerical difficulties (as a result of large matrix operations), thus it is often preferable to resort to alternative modelling approaches, such as those of Chapter 4, 6 and 7.



**Figure 4.16** | Hypothetical example of a zero-inflated stationary process with  $p_0 = 0.8$  and continuous Gamma-distributed part  $\mathcal{G}(0.7, 10)$ . a) Simulated time series of 5 000 time steps. b) Comparison between empirical histogram and theoretical PDF.



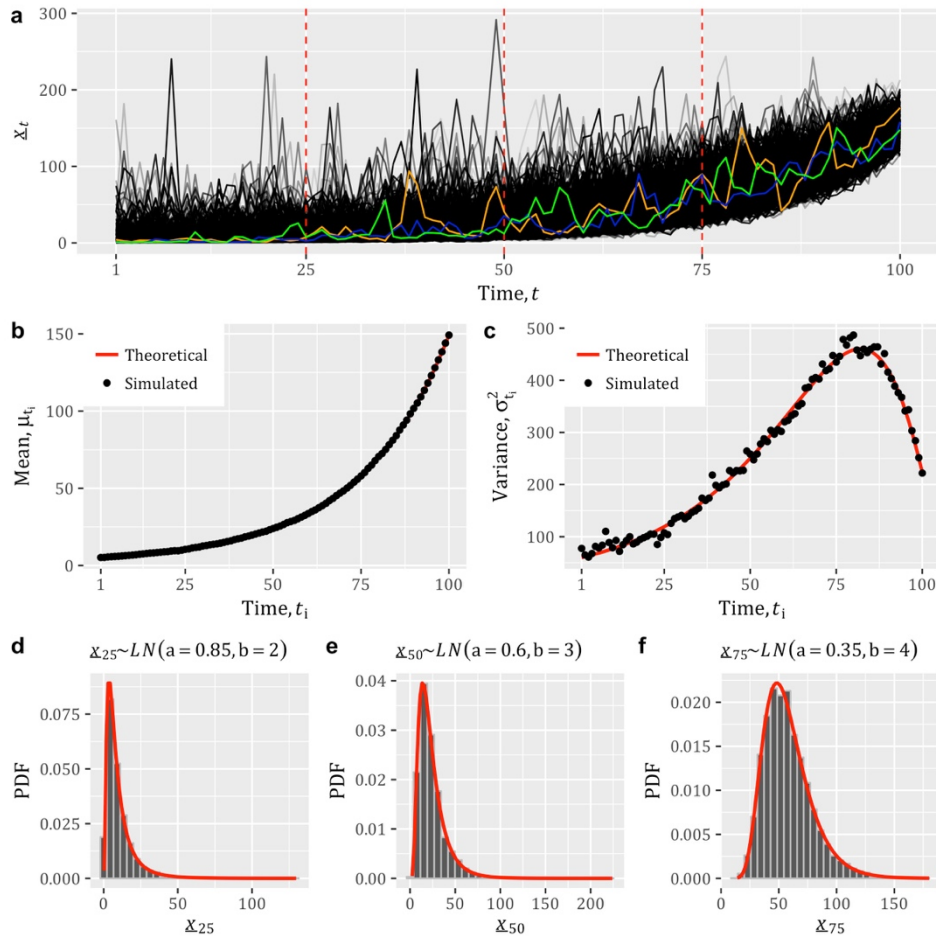
**Figure 4.17** | a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. b) Theoretical and simulated CDFs (using Weibull's plotting position). c) Theoretical, equivalent and simulated autocorrelation functions (ACF). d) Scatter plot depicting the established lag-1 dependence pattern among consecutive time steps.

To further explore the applicability of the method, let us consider the simulation of a non-Gaussian, non-stationary process  $\underline{x}_t$ . In this case the marginal distribution of the process is modeled using the two-parameter Log-Normal (i.e.,  $\mathcal{LN}(a, b)$ ) (see section 4.5.2) with its parameters depending on absolute time  $t$ , i.e.,  $\underline{x}_t \sim \mathcal{LN}(a(t), b(t))$ . Specifically, the shape parameter  $a$  is related with time  $t$  by,  $a(t) = -0.01t + 1.1$ , while the scale  $b$  by,  $b(t) = 0.04t + 1$ . These parameters are also related with the mean and variance (that also depend on absolute time) of the process by,  $\mu_t := E[\underline{x}_t] = \exp(b(t) + 0.5(a(t))^2)$  and  $\sigma_t^2 := \text{Var}[\underline{x}_t] = \exp(2b(t) + (a(t))^2) (\exp((a(t))^2) - 1)$  respectively. The  $i^{\text{th}}$  and  $j^{\text{th}}$  elements of the correlation matrix  $\mathbf{R}$  were set as  $\mathbf{R}_{[i,j]} = 0.6^{|i-j|}$ . Of course, this implies that the  $i^{\text{th}}$  and  $j^{\text{th}}$  elements of the equivalent matrix  $\tilde{\mathbf{R}}$  are being determined using,  $\tilde{\mathbf{R}}_{[i,j]} = \mathcal{F}^{-1}(\mathbf{R}_{[i,j]} | F_{\underline{x}^i}, F_{\underline{x}^j})$ , where  $F_{\underline{x}^i}$  and  $F_{\underline{x}^j}$  denote the distribution of the process at time  $t_i$  and  $t_j$  respectively. Using this setup, 5 000 realizations, each of 100 time steps, are synthesized (hereafter called ensemble) and the results from this example are summarized in Figure 4.18 and Figure 4.19, which illustrate that the model is capable to fulfill its promises and reproduce the time-varying, 1)

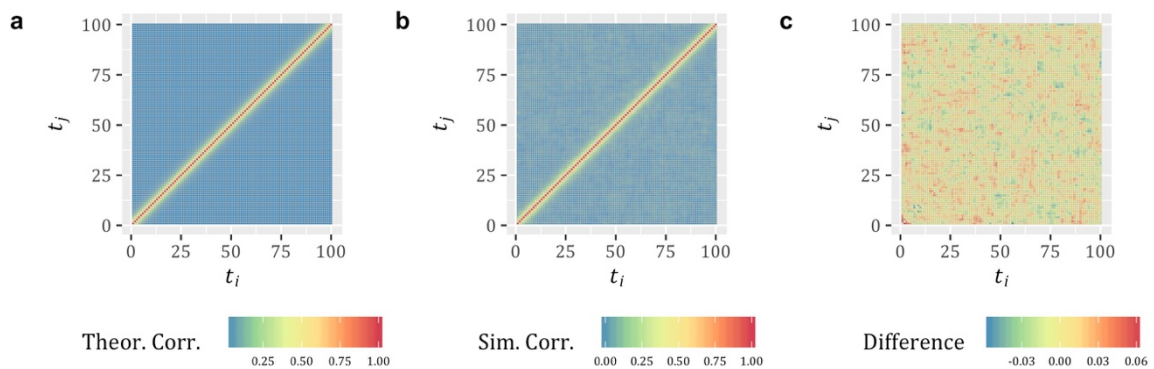


### 4.3 NATAF-BASED STOCHASTIC PROCESSES WITH ARBITRARY MARGINAL DISTRIBUTIONS AND CORRELATION STRUCTURE

distribution function, 2) mean, and 3) variance of the process, as well the target correlation structure. It is remarked, that in practice, deterministic relationships, such as those employed in this example are always unknown, hence the concept of non-stationarity should always be used with extreme caution.



**Figure 4.18** | Hypothetical example of a non-stationary process with Log-Normal marginal distribution. a) All 5 000 realizations (three of which are depicted with distinct colors), each consisted of 100 time steps. Comparison of theoretical and ensemble b) mean and c) variance as a function of time. Comparison between empirical histogram and theoretical PDF for time d)  $t = 25$ , e)  $t = 50$  and f)  $t = 75$ .



**Figure 4.19** | a) Theoretical and b) simulated correlation as a function of absolute time. c) Absolute difference between theoretical and simulated correlation coefficients.

#### 4.3.9 A brief note on Nataf-based stochastic models

By now, it should be clear that the theoretical developments presented herein allow the construction of explicit, in terms of reproducing the distribution function, stochastic simulation methods (relieved from the limitations and constraints of such schemes; see section 2.3) that fundamentally differ from the other two typical schemes (implicit and transformation-based; see section 2.3.1) used in hydrology, which also employ linear stochastic models.

Compared to the implicit approaches that employ non-Gaussian white noise, Nataf-based schemes alleviate several notable limitations. Among them, the approximation of the distribution function, the generation of negative values, the bounded dependence patterns and the (often) narrow type of possible correlation structures, which is attributed to the limited number of schemes for which analytical equations can be derived to link the moments of the process with those of the white noise.

Additionally, in contrast to transformation-based approaches, that aim to *normalize* the data, Nataf-based schemes explicitly model them using target marginal distributions. Though, it has to be noted, that in principle, the rationale of transformation-based approaches can be easily aligned with the theoretical background of Nataf's distribution model by using the concept of equivalent (i.e., adjusted) correlation coefficients. This modification would mitigate their major weakness (i.e., the introduction of bias) but still will not be equivalent with the reproduction of a certain, pre-specified, distribution functions. On top of this, since the ICDF is employed, a unique advantage of Nataf-based approaches, over the aforementioned schemes, is that it can be used for the simulation of both univariate and multivariate stationary processes with continuous, discrete, or mixed-type distributions, which implies that explicitly avoid the generation of unwanted negative values (see Chapter 5 and 6).

Regarding parameterization, the Nataf-based approaches exhibit a parsimonious character, as it is evident by the small number of required parameters, which are equal or lower than those required by the aforementioned schemes (for a comparison see section 5.6.1.1, 6.5.1 and 6.5.3).

As discussed earlier, NDM and the associated methods can be used for the simulation of correlated RVs with continuous, discrete or mixed-type marginal distributions, provided that their combination is feasible and the variance of the distributions is finite. The latter requirement stems for the definition of Pearson's correlation coefficient, which implies that that both the mean and variance of the involved distributions are finite. This assumption also holds for Nataf-based stochastic models, and of course, for any other model that relies on Pearson's correlation coefficient. In such situations the use of alternative simulation methods is required [e.g., Samoradnitsky, 2017].

Random variables with infinite moments typically arise when heavy-tailed distribution functions with power-type tails are employed. For instance, a Pareto type-I distribution with CDF,  $F(x) = 1 - (x/b)^{-a}$ , where  $b > 0$  (scale),  $a > 0$  (shape) and  $x \geq b$ , has finite variance only for  $a > 2$ . The literature offers a plethora of studies indicating the suitability of heavy-tailed distributions for both precipitation [e.g., Papalexiou and Koutsoyiannis, 2013, 2016; Papalexiou et al., 2013; Cavanaugh et al., 2015; Koutsoyiannis and Papalexiou, 2016] and streamflow [e.g., Anderson and Meerschaert, 1998; Bowers et al., 2012; Basso et al., 2015; Blum et al., 2017] processes, especially regarding the description of their extreme behavior. By reviewing the outcomes of these studies, which involve the analysis of numerous worldwide

historical records, it was found that the majority of them agree that the such variables are characterized by distribution functions (with either exponential or power-type tails) with finite variance [Koutsoyiannis and Papalexiou, 2016]. On top of the empirical evidence provided by the aforementioned works, theoretical reasoning (related with entropy and energy production) further supports the finite variance hypothesis for natural processes [Koutsoyiannis, 2016, 2017]. In this vein, it is regarded that this *infinite variance* assumption poses a practical barrier of limited impact, if any, on the application of abovementioned methods for the simulation of hydrometeorological processes.

Moreover, due to changing environmental and hydroclimatic conditions, the statistical information contained in historical data may not be fully representative of the *projected* future conditions. In this context, aiming to explore the effects of *change*, several researchers suggest perturbing the values of the statistical characteristics to be reproduced within synthetic data [e.g., Baltas, 2007; Baltas and Karaliolidou, 2008; Nazemi et al., 2013; Borgomeo et al., 2015], which implies employing parameters different than the data-driven ones. Nevertheless, wherever it is necessary to manually assign target input values, these have to be checked against both physical consistency and hydrological evidence. In this vein, it is remarked that Nataf-based models are able to synthesize data from any distribution hence allowing their straightforward use in such studies. This can be easily accomplished by changing the parameters of the distribution functions (even the distribution functions themselves) or the correlation structure of the process and subsequently investigate the effects of such changes to the system under study.

#### 4.4 THE CASE OF MIXED MARGINAL DISTRIBUTIONS

Herein we highlight the case of mixed-distributions (which can be used within Nataf-based methods), often advocated within hydrological applications, either to better represent the tails of the understudy hydrometeorological variable [e.g., Foufoula-Georgiou and Lettenmaier, 1987; Wilks, 1998; Furrer and Katz, 2008; Li et al., 2012, 2013; Evin et al., 2018], or to simultaneously represent the dual character of intermittent processes [e.g., Williams, 1998; Cannon, 2008; Serinaldi, 2009a; Serinaldi and Kilsby, 2014; Bárdossy and Pegram, 2016; Papalexiou, 2018; Tsoukalas et al., 2018d, 2018b]. Herein we briefly describe the second case, which can be accomplished using a *zero-inflated* (also referred to as *zero-augmented* or *discrete-continuous*) distribution model. This model is composed by both a discrete and a continuous part, and its CDF is given by,

$$F_{\underline{x}}(x) = \begin{cases} p_D, & x \leq 0 \\ p_D + (1 - p_D)G_{\underline{x}}(x), & x > 0 \end{cases} \quad (4.45)$$

The discrete part is represented by  $p_D := P(\underline{x} = 0)$ , and denotes the probability of a zero value. The continuous part is given by  $G_{\underline{x}} := F_{\underline{x}|\underline{x}>0} = P(\underline{x} \leq x | \underline{x} > 0)$ , which denotes a continuous distribution function for the non-zero data. For instance, within the context of intermittent hydrometeorological processes (e.g., rainfall),  $p_D$  stands for the probability of a dry interval (i.e., probability dry), and  $G_{\underline{x}}$  represents the distribution of positive amounts. In real-world situations, the most straightforward way to specify  $p_D$  and  $G_{\underline{x}}$  is through the available data. Specifically,  $p_D$  is estimated as the ratio of dry occurrences to the total number of observations, while  $G_{\underline{x}}$  can be identified by fitting a continuous distribution function to the positive amounts. For completeness, the ICDF of the zero-inflated model, which can be used for RVs generation, is given by,

$$F_{\underline{x}}^{-1}(u) = \begin{cases} 0, & 0 \leq u \leq p_D \\ G_{\underline{x}}^{-1}\left(\frac{(u - p_D)}{(1 - p_D)}\right), & p_D < u \leq 1 \end{cases} \quad (4.46)$$

where  $u \in [0, 1]$  denotes probability. In this formulation values less or equal to  $x_D$  (that arise with probability  $p_D$ ) are assumed equal to zero. For real-world applications of this distribution model within the context of hydrometeorological processes simulation see Chapter 5 and 7.

#### 4.5 IDENTIFICATION OF EQUIVALENT CORRELATION COEFFICIENTS

An important part of every Nataf-based method (i.e., unconditional and conditional simulation of RVs, as well as for stochastic processes simulation) is the identification of equivalent correlations. As already mentioned, for the RVs case, in order to preserve the target correlations  $\rho_{i,j}$  in the actual domain, after mapping the Gaussian variables with their prescribed distributions  $F_{\underline{x}_i}$  and  $F_{\underline{x}_j}$ , it is essential to establish a suitable relationship between  $\tilde{\rho}_{i,j}$  and  $\rho_{i,j}$ , i.e.,  $\mathcal{F}(\cdot)$ . The following discussion is centered towards the generic case of RVs, while it is easily adopted for stochastic processes by simply changing the associated RVs. For instance, in the cyclostationary case, we just have to set  $\underline{x}_i := \underline{x}_s^i$  and  $\underline{x}_j := \underline{x}_{s-\tau}^j$ , and approximate the required (by the auxiliary model) equivalent correlation coefficients  $\tilde{\rho}_{s,s-\tau}^{i,j}$  of the target correlations  $\rho_{s,s-\tau}^{i,j}$ .

The literature offers a variety of approaches to establish  $\mathcal{F}(\cdot)$ , including empirically derived relationships [Der Kiureghian and Liu, 1986; Liu and Der Kiureghian, 1986; Ditlevsen and Madsen, 2007], crude search procedures [Cario and Nelson, 1996, 1997], methods based on the Gauss-Kronrod quadrature rule [Cario, 1996], root finding methods [Li and Hammond, 1975; Chen, 2001; Macke et al., 2009] as well as Gauss-Hermite quadrature and Monte Carlo methods [Zhou and Nowak, 1988; Li et al., 2008; Xiao, 2014]. Herein, in contrast to most of the aforementioned procedures, which are suitable only for continuous marginal distributions, we present a recently developed, simple and easy to implement method, which is applicable for any-type marginal distributions  $F_{\underline{x}_i}$  and  $F_{\underline{x}_j}$ , regardless if they are continuous, discrete, or mixed-type (see section 4.4), since its only requirement is the target ICDFs. In a nutshell, the proposed method is based on the evaluation of few pairs of  $\rho_{i,j}$  and  $\tilde{\rho}_{i,j}$  using Monte Carlo simulation and subsequently, the establishment of the relationship of Eq. (4.15) through polynomial interpolation [Tsoukalas et al., 2017a, 2018e, 2018d]. Throughout this Thesis, unless stated otherwise, the relationship of Eq. (4.15) is established using the above method. The proposed method is particularly useful when the concept of NDM is employed for the simulation of stochastic processes, since it significantly reduces the required computational load. See Chapter 5 and 6 for a more thorough discussion on the subject.

##### 4.5.1 A hybrid Monte Carlo approach

In this context, the following generic procedure has been developed. Let  $\underline{x}_i$  and  $\underline{x}_j$  be two random variables while  $\tilde{\rho}_{i,j}$  and  $\rho_{i,j}$  stand for the equivalent (in Gaussian domain) and the target correlation coefficients respectively. Furthermore, let  $F_{\underline{x}_i}$  and  $F_{\underline{x}_j}$ , denote the corresponding target distributions, whose variance is assumed finite. The developed procedure is comprised by the following steps (the indices  $i$  and  $j$  are omitted for simplicity):

**Step 1:** Create a  $\Omega$ -dimensional vector  $\tilde{\boldsymbol{\rho}} = [\tilde{\rho}^1, \dots, \tilde{\rho}^k, \dots, \tilde{\rho}^\Omega]$  of equally spaced values in the interval  $[\tilde{\rho}_{\min}, \tilde{\rho}_{\max}]$ . Here, lemma 2 can be accounted for in order to determine the boundaries  $\tilde{\rho}_{\min}$  and  $\tilde{\rho}_{\max}$ , since it provides insights regarding the sign of  $\tilde{\rho}_{i,j}$ . For example, if the target correlation  $\rho_{i,j}$  is positive, then we set  $\tilde{\rho}_{\min} = 0$  and  $\tilde{\rho}_{\max} = 1$ .

**Step 2:** For each element of  $\tilde{\boldsymbol{\rho}}$ , generate  $N$  samples from the bivariate standard normal distribution, with correlation  $\tilde{\rho}^k$ .

**Step 3:** Map the generated data to the actual domain through  $\underline{x}_i = F_{\underline{x}_i}^{-1}(\Phi(Z_i))$ , using the associated target marginal distributions  $F_{\underline{x}_i}$  and  $F_{\underline{x}_j}$ .

**Step 4:** Calculate the empirical correlations  $\hat{\rho}^k$  and store them in the vector  $\hat{\boldsymbol{\rho}} = [\hat{\rho}^1, \dots, \hat{\rho}^k, \dots, \hat{\rho}^\Omega]$ .

**Step 5:** Approximate the relationship between target ( $\rho_{i,j}$ ) and equivalent ( $\tilde{\rho}_{i,j}$ ) correlation by establishing a polynomial function of order  $p$ , among the values of  $\tilde{\boldsymbol{\rho}}$  and  $\hat{\boldsymbol{\rho}}$  i.e.,

$$\rho = \mathcal{F}\left(\tilde{\rho} \middle| F_{\underline{x}_i}, F_{\underline{x}_j}\right) \cong \hat{\rho} = \theta_p \tilde{\rho}^p + \theta_{p-1} \tilde{\rho}^{p-1} + \dots + \theta_1 \tilde{\rho}^1 + \theta_0 \quad (4.47)$$

**Step 6:** Evaluate the equivalent correlation  $\tilde{\rho}_{i,j}$  by inverting the relationship between the fitted polynomial and the target correlation  $\rho_{i,j}$ .

We highlight that, according to Weierstrass approximation theorem, the formulation of the polynomial expression of Eq. (4.47) is theoretically feasible, since  $\mathcal{F}(\cdot)$  is continuous and  $\tilde{\rho}$  is by definition bounded on the interval  $[-1, 1]$ . Moreover, we remark that the constant term  $\theta_0$  could be omitted, as indicated by Lemma 2.

The above procedure, which is hybrid combination of Monte Carlo simulation and numerical interpolation through polynomial regression, uses three input arguments, i.e., the vector dimension  $\Omega$ , the sample size  $N$ , and the polynomial order  $p$ . The first two influence the accuracy and computational effort of the Monte Carlo procedure, while the third influences the accuracy of the interpolation approach. Preliminary analysis detected that a good balance between accuracy and computational efficiency is ensured for  $\Omega$  around 10 - 20, and  $N$  around 50 000 - 100 000 trials. Regarding the polynomial order, [Xiao \[2014\]](#) conducted an extensive analysis, with distributions exhibiting a wide range of skewness and kurtosis coefficients, and concluded that  $\mathcal{F}(\cdot)$  can be accurately approximated by a polynomial of less than ninth degree ( $p \leq 9$ ). Apparently, for  $p = \Omega - 1$ , the polynomial passes exactly through all simulated points, yet, in order to ensure parsimony, it may be preferable employing a less complicated expression. In this vein, in order to avoid over-fitting, we propose adjusting the order of the polynomial with the use of cross-validation techniques or the Akaike information criterion [[Akaike, 1974](#)]. We note that on the basis of systematic investigations, instead of polynomials, alternative functions could be employed [e.g., [Serinaldi and Lombardo, 2017](#); [Papalexiou, 2018](#)].

The key advantages of the proposed methodology, are its generality (it can be used for continuous, discrete or mixed-type distributions) and simplicity, as well as the fact that it doesn't depend on specialized algorithms to solve the double integral embedded in Eq. (4.15), in order to obtain a valid expression  $\mathcal{F}(\cdot)$ . Despite the iterative nature of the algorithm, its implementation in high-level programming languages, such as R or MATLAB, requires less

than 1/2 second (assuming  $N = 150\,000$ ,  $\Omega = 9$  and  $p = 8$ ) on a modest 3.0 GHz Intel Dual-Core i5 processor with 4 GB RAM.

Note that, the proper and accurate identification of the relationship  $\mathcal{F}(\cdot)$  has a crucial role in NDM-based schemes, since its misspecification may lead to simulation errors. Hence, to assess the suitability of the developed algorithm, which is extensively used in this work, we employed it and recreated the cases depicted in **Figure 4.2**; which concerned the identification of equivalent correlation coefficients of two Gamma-distributed variables  $\underline{x}_1$  and  $\underline{x}_2$ , for various values of shape parameters. The parameters of the algorithm were set as follows,  $N = 150\,000$ ,  $p = 8$  and  $\Omega = 9$ . After the specification of the relationship  $\mathcal{F}(\cdot)$  by the latter algorithm, the target correlations were evaluated for values of  $\tilde{\rho}_{1,2} \in [-1,1]$  sampled by 0.01. To provide a quantitative comparison, we estimated the MSE and maximum square error (Max(SE)) between the estimates of the numerical integration method (i.e., **Figure 4.2**) and those of the aforementioned algorithm. A synopsis of the results is given on **Table 4-1**, where the panels (a) and (b) corresponds to those of **Figure 4.2**. This analysis illustrates the potential of the employed method to resemble the asymmetric and non-linear nature of  $\mathcal{F}(\cdot)$  with high accuracy.

**Table 4-1** | Comparison between numerical integration and the algorithm of section 4.5.1 for the numerical example illustrated in **Figure 4.2**. Panels a) and b) correspond to those of **Figure 4.2**.

| a) | $a := a_1 = a_2 \mid b := b_\xi = b_2 = 1$ |                       |                       | b) | $a_1 = 5 \mid b := b_1 = b_2 = 1$ |                       |                       |
|----|--|-----------------------|-----------------------|----|-----------------------------------|-----------------------|-----------------------|
|    | Shape ( $a$ )                              | MSE                   | Max(SE)               |    | Shape ( $a_1$ )                   | MSE                   | Max(SE)               |
|    | 0.01                                       | $8.03 \times 10^{-5}$ | $7.75 \times 10^{-4}$ |    | 0.01                              | $2.12 \times 10^{-5}$ | $3.79 \times 10^{-4}$ |
|    | 0.05                                       | $5.81 \times 10^{-5}$ | $3.08 \times 10^{-4}$ |    | 0.05                              | $6.46 \times 10^{-6}$ | $2.70 \times 10^{-5}$ |
|    | 0.1  | $2.44 \times 10^{-6}$ | $9.89 \times 10^{-6}$ |    | 0.1                               | $6.26 \times 10^{-6}$ | $4.15 \times 10^{-5}$ |
|    | 0.5  | $4.33 \times 10^{-6}$ | $1.59 \times 10^{-5}$ |    | 0.5                               | $1.51 \times 10^{-5}$ | $9.37 \times 10^{-5}$ |
|    | 1  | $3.31 \times 10^{-6}$ | $1.88 \times 10^{-5}$ |    | 1                                 | $2.54 \times 10^{-6}$ | $1.13 \times 10^{-5}$ |
|    | 2  | $1.22 \times 10^{-6}$ | $8.47 \times 10^{-6}$ |    | 2                                 | $7.19 \times 10^{-7}$ | $3.20 \times 10^{-6}$ |
|    | 5  | $3.70 \times 10^{-6}$ | $1.80 \times 10^{-5}$ |    | 5                                 | $5.24 \times 10^{-7}$ | $1.77 \times 10^{-6}$ |

#### 4.5.2 The Log-Normal case

As mentioned earlier, Eq. (4.15) has a closed-form solution for the Log-Normal case, which is of particular interest from a hydrological perspective. The PDF of the 3-parameter Log-Normal distribution ( $\mathcal{LN}$ ) is given by,

$$f_{\mathcal{LN}}(x; a, b, c) = \frac{1}{(x - c)a\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(x - c) - b}{a}\right)^2\right), \quad x > c \quad (4.48)$$

where  $a > 0$ ,  $b \in \mathbb{R}$  and  $c \in \mathbb{R}$  denote the shape (i.e., log standard deviation), scale (i.e., log mean) and location parameters respectively; while when  $c = 0$ , the distribution reduces to the 2-parameter Log-Normal distribution. As shown in *Mostafa and Mahmoud [1964]*, yet without direct reference to NDM, for two random variables  $\underline{x}_i$  and  $\underline{x}_j$  that are Log-Normally distributed, Eq. (4.13), hence Eq. (4.15) simplifies to,

$$\rho_{i,j} = \frac{\exp(\tilde{\rho}_{i,j} a_i a_j) - 1}{\sqrt{(\exp(a_i^2) - 1)(\exp(a_j^2) - 1)}} \quad (4.49)$$

Which can be easily inverted in order to directly provide the equivalent correlation coefficient  $\tilde{\rho}_{i,j}$ , given the target value of  $\rho_{i,j}$ . i.e.,

$$\tilde{\rho}_{i,j} = \frac{\text{Ln} \left( 1 + \rho_{i,j} \sqrt{(\exp(a_i^2) - 1)(\exp(a_j^2) - 1)} \right)}{a_i a_j} \quad (4.50)$$

It is remarked that Eq. (4.50) is identical with the one employed in the celebrated multivariate lag-1 Log-Normal model of *Matalas [1967]*, to adjust the correlation coefficients, which interestingly can be identified as a Nataf-based approach [cf. *Tsoukalas et al., 2018d*].

#### 4.5.3 A cautionary note

A delicate point worth stating concerns the use of alternative, rank-based dependence measures such as, Spearman's  $r_s$  and Kendall's  $t$  within NDM (or Gaussian copula). Under the assumption that both marginal distributions and copula are Gaussian (or more generally elliptical distributions) there is a one-to-one relationship between these dependence measures and Pearson's correlation, which can be expressed as [e.g., *Esscher, 1924; Kruskal, 1958; Embrechts et al., 1999; Lebrun and Dutfoy, 2009*] (notice that the indices have been omitted for the sake of simplicity),

$$\rho = 2 \sin \left( \frac{\pi r_s}{6} \right) \leftrightarrow r_s = \left( \frac{6}{\pi} \right) \arcsin \left( \frac{\rho}{2} \right) \quad (4.51)$$

$$\rho = \sin \left( \frac{\pi t}{2} \right) \leftrightarrow t = \left( \frac{2}{\pi} \right) \arcsin(\rho) \quad (4.52)$$

Both  $r_s$  and  $t$  are measures of concordance and are invariant to non-linear monotonic transformations, such as those imposed by  $\underline{x}_i = F_{x_i}^{-1} \left( \Phi(\underline{z}_i) \right)$ . Specifying NDM or Gaussian copula with estimates of  $\rho$  based on the conversion of empirical estimates of  $r_s$  or  $t$  will inevitably preserve the target values of  $r_s$  or  $t$  after the application of the mapping procedure (due to the property of invariance) but it will lead to misspecification of the underlying model (i.e., NDM or Gaussian copula) due to Eq. (4.15), and of course the target values of  $\rho$  won't be preserved. Unfortunately, the underlying assumptions of Eq. (4.51) and Eq. (4.52) are often relaxed or not fully considered in practice since these equations have been employed in several works (that involve an auxiliary Gaussian model) within the hydrological domain [e.g., *Kelly and Krzysztofowicz, 1997; Herr and Krzysztofowicz, 2005; Renard and Lang, 2007; Serinaldi, 2009a, 2009b; Srikanthan and Pegram, 2009; Mhanna and Bauwens, 2012; Serinaldi and Kilsby, 2014*], just to name a few.

## 4.6 BITS AND PIECES OF NDM IN HYDROLOGY

NDM-based approaches have been widely applied in industrial, financial and operations research applications, as indicated from the popularity of the original article by *Nataf [1962]* and the related publications [*Liu and Der Kiureghian, 1986; Cario and Nelson, 1996, 1997; Grigoriu, 1998; Deodatis and Micaletti, 2001; Biller and Nelson, 2003*].

While the hydrological community does not make direct reference to NDM, the concept of equivalent correlations (which are often neglected; see section 4.5.3) and the associated models, such as NORTA, ARTA, VARTA, etc. it actually shares the same rationale, even from the

geneses of hydrological stochastics [see, *Tsoukalas et al., 2018e*]. Loosely speaking, the core idea of NDM comprises the initiation from the Gaussian domain, with properly adjusted correlation coefficients, and then a mapping to the desirable domain; an idea that can be retrospectively associated with several well-known hydrological approaches.

In particular, *Matalas [1967]* has studied the effects of logarithmic transformations in the context of synthesizing log-normally distributed processes, concluding that the so far prevailing transformation approach failed to resemble the historical statistics. To reestablish consistency, he developed a framework based on the generation of normal processes, and provided a set of theoretical equations to estimate the statistical parameters (including adjusted correlation coefficients) in the Log-Normal domain. *Moran [1969]* developed a bivariate Gamma distribution using as main building block an auxiliary bivariate Gaussian distribution. Later, *Klemeš and Borůvka [1974]* developed a generation scheme for gamma-distributed univariate first-order Markov chains, through a mapping procedure of Gaussian processes with the use of adjusted correlation coefficients. *Mejía and Rodríguez-Iturbe [1974]* discuss the link between Gaussian and log-Normal processes, while they also comment that as formulated, the log-Normal model of *Matalas [1967]*, is able to resemble only the lag-1 autocorrelation coefficient and approximate the Markovian autocorrelation structure. A fact attributed to the use of adjusted correlations only for the lag-1 correlations. Yet, it is noted that this deviation from the Markovian structure is typically minimal. More recently, *Kelly and Krzysztofowicz [1997]* proposed and illustrated through several hydrology-related applications, a flexible bivariate distribution model, termed meta-Gaussian, which builds upon the bivariate standard normal distribution and the normal quantile transformation. *Wilks [1998]*, in the context of his widely known rainfall generation model (which is also associated with many weather generator schemes; see section 2.3.3), and in an effort to simulate cross-correlated random variates, representing either the precipitation occurrence or amount process (neglecting temporal dependence), proposed the simulation of cross-correlated Gaussian variables and their subsequent mapping via their ICDF. Wilks empirically observed that a monotonic relationship exists which links the correlation coefficients of the Gaussian and *real* domain. Hence, the use of inflated correlation coefficients was proposed within the multivariate Gaussian distribution, in order to attain random variates with the required cross-correlation and distribution. This seminal work has triggered the development of improved schemes, supporting more distributions and correlation structures (see also section 2.3.3).

Additionally, advances in stochastic hydrology are also in alignment with NDM, since it seems that presently, Nataf-based approaches are gaining momentum. In particular, in a similar vein, *Serinaldi and Lombardo [2017]* proposed a fast procedure for autocorrelated univariate binary processes. *Lee [2017]* introduced a Gaussian copula, simulation-based method for cross-correlated, yet serially independent, Gamma-distributed precipitation. *Papalexiou [2018]* provided a framework for synthetic data generation using autoregressive models, also accounting for intermittency using mixed distributions. *Tsoukalas et al. [2017a, 2018e]*, employed the notion of NDM and provided a cyclostationary generalization of the models ARTA and VARTA, termed SPARTA (Stochastic Periodic AutoRegressive To Anything), for the simulation of univariate and multivariate periodic processes with arbitrary marginal distributions (see Chapter 6). Furthermore, *Tsoukalas et al. [2018d]* proposed a model termed Symmetric Moving Average (nearLy) To Anything (SMARTA), which is capable of simulating univariate and multivariate stationary stochastic processes with any distribution and correlation structure (see Chapter 5, which presents also an additional model). It is noted that the designation *nearly* in the model is included to emphasize that the target marginal distributions



ought to have finite variance. Finally, *Tsoukalas et al. [2018b]*, presented a multivariate multi-level disaggregation-based approach, designed for the pairwise coupling of Nataf-based stochastic models that operate independently of each other at certain key time scales. The coupling approach, as well as the solid theoretical basis of these models, enable the development of modular stochastic simulation schemes, that can synthesize multivariate time series with any distribution and correlation structure that are also statistically consistent across multiple temporal scales (see Chapter 7).

#### 4.7 SUMMARY

This Chapter provided a comprehensive treatment on the Nataf's joint distribution (NDM) model, starting from its theoretical basis and the establishment of the multivariate joint distribution of RVs. The analysis also highlighted its relationship with the Gaussian copula, which in turn allow us to extend NDM for the derivation of the multivariate conditional distribution. The applicability of these concepts has been demonstrated through several examples, including continuous, discrete and mixed-type distributions.

Subsequently, the concept of NDM, has been adapted for the simulation of non-Gaussian stochastic processes, and general guidelines that can be used for the development of Nataf-based stochastic simulation models have been provided (see below).

An additional contribution of this Chapter is the development of a simple and versatile Monte Carlo procedure for the identification of equivalent correlation coefficients, which have an essential, yet often neglected, role in the establishment of NDM-based constructs.

In the following Chapters, 5 and 6, the focus is given on the development of non-Gaussian Nataf-based models for stationary and cyclostationary processes respectively, while Chapter 7 provides a fusion of these new developments, into integrated stochastic simulation schemes, capable of simultaneously accounting for the peculiar characteristics of hydrometeorological processes at multiple time scales.

## SIMULATION OF STATIONARY STOCHASTIC PROCESSES EXHIBITING ANY-RANGE DEPENDENCE AND ARBITRARY MARGINAL DISTRIBUTIONS

♣

### PREAMBLE

This Chapter presents a novel approach for synthetic time series generation. In particular it presents two models, termed Symmetric Moving Average (nearLy) To Anything (SMARTA) and Contemporaneous Multivariate Autoregressive (nearLy) to Anything (CMARTA), able to simulate stationary univariate and multivariate contemporaneously cross-correlated processes with any-range dependence and arbitrary marginal distributions; provided that the former is feasible and the latter have finite variance. This is accomplished by utilizing a mapping procedure in combination with the relationship that exists between the correlation coefficients of an auxiliary Gaussian process and a non-Gaussian one, formalized through the Nataf's joint distribution model. The generality of the two models is validated through several hypothetical simulation studies (univariate and multivariate), characterized by different dependencies and distributions. We demonstrate the practical aspects of the proposed approach through two real-world cases, one that concerns the generation of annual non-Gaussian streamflow time series at four stations and another that involves the synthesis of intermittent, non-Gaussian, daily rainfall series at a single location.

The structure of the Chapter is as follows: Section 5.1 introduces the problem. Section 5.2 presents some key concepts regarding modelling of auto-dependence structure in general. Section 5.3 provides the theoretical background of the proposed models; next, section 5.4 describes the auxiliary SMA and CMAR models and, section 5.5 summarizes the overall approach and provides the generation mechanism of the two models in an algorithmic step-by-step manner. The generality of SMARTA and CMARTA models is illustrated through a series of numerical examples, hypothetical (section 5.6) and real-world (section 5.7), including the simulation of both univariate and multivariate time series. Finally, in section 5.8 we summarize and discuss the proposed modelling approach.

---

\* Based on:

Tsoukalas, I., C. Makropoulos, and D. Koutsoyiannis (2018d), Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions, *Water Resour. Res.*, doi:10.1029/2017WR022462.

## 5.1 INTRODUCTION

A typical characteristic encountered in hydrometeorological processes is auto-dependence (persistence), either short or long-range. The former, short-range dependence (SRD), has been extensively discussed in literature [e.g., *Box and Jenkins, 1970*] and implies an exponential autocorrelation structure that diminishes after few time lags. On the contrary, the second, long-range dependence (LRD), also known as long-term persistence (sometimes referred to as long-memory), implies an auto-dependence structure that strongly extends for large lags. This behavior is also related to the so-called Hurst phenomenon, also known as Joseph effect, fractional Gaussian noise (fGn), scaling in time or Hurst-Kolmogorov dynamics [HK; *Koutsoyiannis and Montanari, 2007; Koutsoyiannis, 2011a*]; see also the review work of *Molz et al. [1997]*. Its discovery is usually credited to *Hurst [1951]* who while studying long records of streamflow and other data noticed that extreme events tend to exhibit a clustering behavior in terms of temporal occurrence. However, as pointed out by *Koutsoyiannis [2011a]*, it was *Kolmogorov [1940]* who first proposed its mathematical description. Eventually, after the seminal work of Hurst and the extensive documentation of *Mandelbrot and Wallis [1969a, 1969b, 1969c]* it is now acknowledged that LRD (and HK) processes are omnipresent in geophysics, hydrology, climate and other scientific disciplines [*Beran, 1994; Koutsoyiannis, 2002; O'Connell et al., 2016*]. These publications provide further examples and details regarding the interpretation and identification of such behavior.

As far as it concerns modelling and application of SRD or LRD in hydrological studies, the former type (SRD) has been systematically studied and employed in numerous cases for the simulation of a variety of hydrometeorological processes [*Matalas, 1967; Srikanthan and McMahon, 2001; Brissette et al., 2007; Thompson et al., 2007; Khalili et al., 2009; Srikanthan and Pegram, 2009; Mhanna and Bauwens, 2012; Breinl et al., 2013; Mehrotra et al., 2015*]. On the other hand, it is well recognized that proper representation of LRD is of high importance, especially in reservoir-related studies, since their operation and regulation is performed in over-annual scale, where LRD is mostly encountered [*Bras and Rodríguez-Iturbe, 1985; Koutsoyiannis, 2002; Iliopoulou et al., 2016*]. Other notable hydrology-related applications of LRD include the stochastic simulation of precipitation or streamflow at finer time-scales, from monthly and daily [*Montanari et al., 1997, 2000; Maftai et al., 2016; e.g., Detzel and Mine, 2017*] to 10-second interval [*Papalexiou et al., 2011; e.g., Lombardo et al., 2012*], as well as the generation of synthetic storm hyetographs [e.g., *Koutsoyiannis and Foufoula-Georgiou, 1993*].

In general, SRD can be easily captured with the so-called AutoRegressive Moving Average (ARMA) family of models, while we note, that even though such models have a long history and are still popular, today the literature offers other powerful and flexible options [cf. *Koutsoyiannis, 2016*]. On the other hand, LRD, hence HK behavior, requires the use of alternative generation schemes [see, *Bras and Rodríguez-Iturbe, 1985; O'Connell et al., 2016*], such as, fractional Gaussian noise models [*Mandelbrot and Wallis, 1969a, 1969b, 1969c*], fast fractional Gaussian noise (ffGn) models [*Mandelbrot, 1971*], broken line models [*Ditlevsen, 1971; Mejia et al., 1972*] and Fractional AutoRegressive Integrated Moving-Average (FARIMA) models [*Granger and Joyeux, 1980; Hosking, 1984*]. In contrast to the abovementioned specialized simulation schemes, a notable exception, that can simulate any type of autocorrelation function of a process, is the symmetric moving average (SMA) model of *Koutsoyiannis [2000, 2002, 2016]*, coupled with theoretical autocorrelation (or autocovariance) structures. This flexibility is achieved by decoupling the parameter identification of the autocorrelation structure and the generation mechanism (i.e., the model).

In addition to temporal dependence, hydrometeorological processes are often characterized by non-Gaussian and skewed distribution functions (see the discussion in section 2.2), especially in fine time scales (e.g., daily or finer), where intermittency is omnipresent. Regarding stochastic hydrology and simulation through linear stochastic models, many efforts have been made towards that direction (i.e., simulating non-Gaussian processes) which can be broadly classified in three main categories [Tsoukalas et al., 2018e]: a) Explicit methods that are able to generate data from specific marginal distributions [e.g., Matalas, 1967; Klemeš and Borůvka, 1974; Lawrance and Lewis, 1981a; Lombardo et al., 2012, 2017] b) Implicit approaches, pioneered by Thomas and Fiering [1963], that treat skewness via employing non-Gaussian white noise (typically from Pearson type-III distribution) for the innovation term [e.g., Matalas, 1967; Matalas and Wallis, 1971, 1976; Lettenmaier and Burges, 1977; Todini, 1980; Koutsoyiannis, 1999, 2000; Efstratiadis et al., 2014a; Detzel and Mine, 2017]. c) Transformation-based approaches that employ appropriate functions (e.g., Box-Cox) in order to *normalize* the observed data; next simulate realizations using typical Gaussian stochastic models and finally *de-normalize* the generated data in order to attain the process of interest [e.g., Salas et al., 1980].

However, as discussed in Tsoukalas et al. [2018e], as well as in section 2.2, most of these schemes exhibit a number of limitations that still remain unresolved. Particularly, approaches of category (a) are limited to a narrow type of autocorrelation functions and non-Gaussian distributions (e.g., Gamma or Log-Normal), while they are typically able to simulate only univariate processes. On the other hand, approaches of category (b) are prone to the generation of negative values, provide an approximation of the marginal distributions, while encounter difficulties when modelling highly skewed (univariate or multivariate) processes [Todini, 1980; Koutsoyiannis, 1999]. It is noted thought, that some recent schemes are able to capture moments higher than skewness (e.g., kurtosis), by the inclusion of additional model parameters [Koutsoyiannis et al., 2018 and references therein]. On top of these issues, only few schemes (e.g., SMA) are able to simultaneously model a variety of temporal correlation structures, while it is also possible to establish bounded dependence patterns which are far from natural ones (see Tsoukalas et al. [2018e, 2018a], Chapter 3 and section 6.5). Finally, regarding the schemes of category (c), they require the specification of a non-trivial normalization function (due to the inadequacy of simple transformations; such as, Box-Cox) that often entail several parameters (usually determined through optimization techniques). Further to this, even if the transformation function is properly identified, it is acknowledged that it introduces bias in the simulated marginal and joint characteristics [Salas et al., 1980 p. 73; Bras and Rodríguez-Iturbe, 1985].

In this Chapter, in an effort to simultaneously address these challenges and provide flexible tools for the generation of hydrometeorological synthetic time series, we build upon the concept of Nataf-based processes (see section 4.3) and develop two particularly flexible models. The models follow rationale employed within the scientific field of operations research and particularly by Cario and Nelson [1996], as well as, Biller and Nelson [2003] who proposed the AutoRegressive To Anything (ARTA) and the Vector AutoRegressive To Anything (VARTA) methods respectively for the explicit simulation of stationary autoregressive (AR) processes with arbitrary marginal distributions. It is remarked that (to the extent of our knowledge) despite their wide acceptance, the aforementioned approaches (and their variants) have been unknown to the hydrological community and have never been used for the simulation of hydrometeorological processes until very recently (see section 4.6).

Herein we move beyond the simulation low-order AR autocorrelation structures, and introduce two generic, yet simple and theoretically justified, models for the simulation of univariate and multivariate contemporaneously cross-correlated stationary processes exhibiting any-range dependence and arbitrary marginal distributions (continuous, discrete or mixed-type). More specifically, the first model uses as an auxiliary model the SMA scheme of *Koutsoyiannis [2000]*, hence termed Symmetric Moving Average (nearLy) To Anything (SMARTA). The second employs a Contemporaneous Multivariate AutoRegressive model (CMAR), hence termed Contemporaneous Multivariate AutoRegressive (nearLy) To Anything (CMARTA). Both SMARTA and CMARTA can explicitly model the autocorrelation structure and distribution of each individual process, provided that the former is feasible and the latter have finite variance, while simultaneously they can preserve the lag-0 cross-correlation structure. This assumption, which significantly simplifies the parameter estimation procedure, is often regarded adequate within hydrological domain, and can be found in several other (stationary and cyclostationary; typically Gaussian) stochastic simulation schemes [e.g., *Pegram and James, 1972*; e.g., *Camacho et al., 1985, 1987*; *Koutsoyiannis and Manetas, 1996*; *Rasmussen et al., 1996*; *Efstratiadis et al., 2014a*; *Tsoukalas et al., 2018e*].

The main components of the models are, 1) a theoretical autocorrelation structure, 2) an auxiliary model for simulating Gaussian processes, and 3) the pivotal concept of Nataf's joint distribution model [*NDM, Nataf, 1962*]. The key idea of our approach lies in mathematically describing the (target) autocorrelation structure of the process to simulate using a theoretical model and subsequently, employing an auxiliary Gaussian stochastic process, with such parameters that reproduce the target auto- (i.e., temporal; SRD or LRD) and lag-0 cross-correlation (i.e., spatial) coefficients of the target process after its subsequent mapping to the actual domain via the target inverse cumulative density functions (ICDFs).

## 5.2 MODELLING THE AUTO-DEPENDENCE STRUCTURE OF STATIONARY PROCESSES

Prior to describing the proposed models it is considered useful to provide a brief introduction to the tools that allow the mathematical description of the auto-dependence structure of a stochastic process. To elaborate, let  $\underline{x}_t, t \in \mathbb{Z}$  be a discrete-time stationary process, indexed using  $t$ , with finite variance  $\sigma^2 := \text{Var}[\underline{x}_t]$  and autocorrelation function  $\rho_\tau := \text{Corr}[\underline{x}_t, \underline{x}_{t+\tau}]$ , where  $\tau$  denotes the time lag. The autocovariance function (ACVF) of the process can be obtained by,  $c_\tau := \text{Cov}[\underline{x}_t, \underline{x}_{t+\tau}] = \sigma^2 \rho_\tau$ . Note that a valid autocorrelation structure has to be positive definite [e.g., *Lindgren, 2013*], which can be readily checked by formulating, and testing for positive definiteness, the so-called  $(n \times n)$  autocorrelation matrix  $\mathbf{R}$ , whose  $i^{\text{th}}, j^{\text{th}}$  elements are being determined by,  $\mathbf{R}_{[i,j]} = \rho_{|i-j|}$ .

Besides the ACF and ACVF, particularly useful stochastic tool, is the climacogram [*CG, Koutsoyiannis, 2010, 2016*], which is typically depicted using a log-log plot, and expresses the variance of the aggregated  $(\underline{X}_l^{(k)})$  or time averaged  $(\underline{x}_l^{(k)})$  process at scale  $k \in \mathbb{Z}^+$ . We remark that the notation employed herein slightly differs from the typical one, since we restrict our attention to discrete-time processes. Assuming that  $\underline{x}_t$  denotes a discrete-time stationary process at the basic time scale  $k = 1$ , the discrete-time aggregated process at scale  $k > 1$  can be obtained by,

$$\underline{X}_l^{(k)} := \sum_{t=(l-1)k+1}^{kl} \underline{x}_t \quad (5.1)$$

while the averaged discrete-time process is obtained by,  $\underline{x}_l^{(k)} = \underline{X}_l^{(k)}/k$ . Hence the corresponding climacograms of the discrete-time aggregated and averaged process can be defined as  $\Gamma^{(k)} := \text{Var}[\underline{X}_l^{(k)}]$  and  $\gamma^{(k)} := \text{Var}[\underline{x}_l^{(k)}]$  respectively. Moreover, as shown by [Beran \[1994 p. 3\]](#), as well as by [Koutsoyiannis \[2010, 2016\]](#), the variance over scales (i.e., the CG) and the ACVF (and therefore ACF) are interrelated. Specifically, if the theoretical ACVF (or ACF),  $c_\tau$  at the basic time scale ( $k = 1$ ) is known, the corresponding theoretical discrete-time climacogram of the aggregated process can be calculated through the following equation,

$$\Gamma^{(k)} = c_0 k + 2 \sum_{\tau=1}^{k-1} (k - \tau) c_\tau \quad (5.2)$$

while the averaged one can be obtained by,  $\gamma^{(k)} = \Gamma^{(k)}/k^2$ . The recursive application of the following equation facilitates the calculation of the climacogram  $\Gamma^{(k)}$ ,

$$\Gamma^{(k)} = 2\Gamma^{(k-1)} - \Gamma^{(k-2)} + 2c_{k-1} \quad (5.3)$$

It is noted that,  $\Gamma^{(1)} = \gamma^{(1)} = c_0 = \sigma^2$ , while  $\Gamma^{(0)} = 0$ . The inverse relationship that calculates the ACVF of the aggregated discrete-time process ( $\underline{X}_l^{(k)}$ ), denoted  $C_\tau^{(k)} := \text{Cov}[\underline{X}_l^{(k)}, \underline{X}_{l+\tau}^{(k)}]$ , at time scale  $k$  given the theoretical climacogram is given by [\[Koutsoyiannis, 2017\]](#),

$$C_\tau^{(k)} = \frac{\Gamma^{(|\tau+1|k)} + \Gamma^{(|\tau-1|k)}}{2} - \Gamma^{(|\tau|k)}, \quad \forall \tau \in \mathbb{Z} \quad (5.4)$$

Furthermore, the ACVF,  $C_\tau^{(k)}$  at scale  $k$  is linked with the ACVF,  $c_\tau$ , of the basic time scale  $k = 1$ , through the following relationship,

$$C_\tau^{(k)} = \sum_{t=1}^k \sum_{r=\tau k+1}^{(1+\tau)k} \text{Cov}[\underline{x}_t, \underline{x}_r] = \sum_{t=1}^k \sum_{r=\tau k+1}^{(1+\tau)k} c_{|t-r|}, \quad \tau \geq 0 \quad (5.5)$$

Analogously, the ACVF of the time averaged discrete-time process ( $\underline{x}_l^{(k)}$ ) at scale  $k$ , denoted  $c_\tau^{(k)} := \text{Cov}[\underline{x}_l^{(k)}, \underline{x}_{l+\tau}^{(k)}]$ , is obtained by  $c_\tau^{(k)} = C_\tau^{(k)}/k^2$ . Hence, the ACF of the aggregated discrete-time process at time scale  $k$  can be obtained by,  $\rho_\tau^{(k)} = C_\tau^{(k)}/\Gamma^{(k)}$  while the ACF of the time averaged discrete-time process by,  $\rho_\tau^{(k)} = c_\tau^{(k)}/\gamma^{(k)}$ . Note that the ACF of the aggregated and time averaged process are identical, due to standardization of the corresponding ACVF with the variance. It is also noted that  $C_0^{(k)} = \Gamma^{(k)}$  and  $C_\tau^{(1)} = c_\tau$ , while similarly,  $c_0^{(k)} = \gamma^{(k)}$  and  $c_\tau^{(1)} = c_\tau$ . It is also noted that a similar relationship to Eq. (5.5), can be derived for the case of two processes. Particularly let  $\underline{x}_t^i$  and  $\underline{x}_t^j$  be two discrete time processes, at the basic time scale  $k = 1$ , and  $c_\tau^{i,j} = \text{Cov}[\underline{x}_t^i, \underline{x}_{t+\tau}^j]$  denote the lag- $\tau$  cross-covariance function at scale  $k = 1$ . The lag- $\tau$  (for  $\tau = 0, 1, 2, \dots$ ) cross-covariance  $C_\tau^{i,j} = \text{Cov}[\underline{X}_l^{i(k)}, \underline{X}_{l+\tau}^{j(k)}]$  of the aggregated processes  $\underline{X}_l^{i(k)} := \sum_{t=(l-1)k+1}^{lk} \underline{x}_t^i$  and  $\underline{X}_l^{j(k)} := \sum_{t=(l-1)k+1}^{lk} \underline{x}_t^j$  at scale  $k$  is given by,

$$C_{\tau}^{i,j(k)} = \sum_{t=1}^k \sum_{r=\tau k+1}^{(1+\tau)k} \text{Cov}[x_t^i x_r^j] = \sum_{t=1}^k \sum_{r=1}^{(1+\tau)k} c_{t-r}^{i,j} \quad (5.6)$$

For instance, for  $\tau = 0$ , the equation simplifies to,

$$C_0^{i,j(k)} = \text{Cov}[X_l^{i(k)}, X_l^{j(k)}] = \sum_{t=1}^k \sum_{r=1}^k \text{Cov}[x_t^i x_r^j] = \sum_{t=1}^k \sum_{r=1}^k c_{t-r}^{i,j} \quad (5.7)$$

Undoubtedly, the most commonly-employed tool to characterize the auto-dependence structure is the autocorrelation function (ACF). The literature offers a plethora of theoretical models in both continuous and discrete time [Gneiting, 2000; Koutsoyiannis, 2000, 2016; Gneiting and Schlather, 2004; Papalexiou et al., 2011; Dimitriadis and Koutsoyiannis, 2015; Papalexiou, 2018], that can be easily combined with the proposed approach (see next section). In this Chapter we focus our attention to the discrete-time Cauchy-type autocorrelation structure (CAS) of Koutsoyiannis [2000] due to its simple and parsimonious form (a desired property of stochastic modelling), which however does not hinder its ability to model a wide range of short (ARMA-type) and long-range dependence structures (including HK behavior). CAS is a two-parameter power-type autocorrelation structure which in its simplest form, if the ACF has constant and positive sign, (as in the case of geophysical and hydrometeorological processes) is given by,

$$\rho_{\tau}^{\text{CAS}} = (1 + \kappa\beta\tau)^{-1/\beta}, \quad \tau \geq 0 \quad (5.8)$$

where  $\beta \geq 0$  and  $\kappa > 0$  are parameters that control the degree of dependence (or persistence) of the process. It is remarked that the autocorrelation function of an HK (i.e., fGn) process consists a special case (or a very good approximation) of the CAS model (i.e., Eq. (5.8)). The theoretical ACF of an HK process is given by,

$$\rho_{\tau}^{\text{HK}} = \frac{1}{2} (|\tau - 1|^{2H} - 2|\tau|^{2H} + |\tau + 1|^{2H}) \quad (5.9)$$

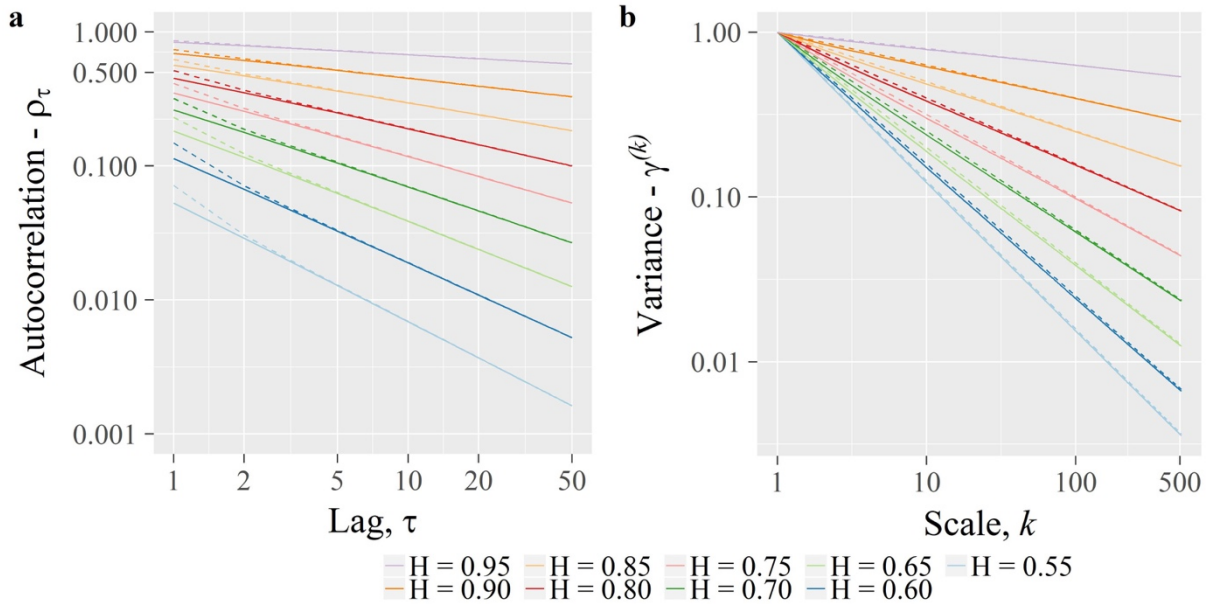
where  $H$  is the Hurst coefficient ( $0 \leq H \leq 1$ ), which loosely speaking, controls the degree of long-term dependence (or persistency) of the process. It has been shown that for large time lags and  $H > 0.5$ , the parameter  $\beta$  of CAS is related to the  $H$  coefficient of an HK ACF through the relationship  $\beta = 1/(2 - 2H) > 1$ , thus asymptotically resembling the right tail of the HK theoretical model. More specifically, for  $\beta > 1$  and when  $\kappa$  is set equal to  $\kappa_0$ , see Eq. (5.10), CAS closely approximates the theoretical ACF of an HK process, even from small time lags.

$$\kappa = \kappa_0 := \frac{1}{\beta \left[ \left(1 - \frac{1}{\beta}\right) \left(1 - \frac{1}{2\beta}\right) \right]^{\beta}} \quad (5.10)$$

In addition, the ACF of an SRD process (ARMA-type) can be obtained through CAS, by setting  $\beta = 0$  and applying the L' Hôpital's rule. The resulting SRD ACF is given by,

$$\rho_{\tau}^{\text{SRD}} = \exp(-\kappa\tau) \quad (5.11)$$

Furthermore, when  $\kappa = -\ln(\rho_1)$ , and  $0 \leq \rho_1 \leq 1$ , Eq. (5.10) reduces to the classic Markovian ACF of an AR(1) process, given by,  $\rho_\tau^{\text{AR}(1)} = \rho_1^{|\tau|}$ . For other parameter values, CAS resembles a plethora of alternative autocorrelation structures, that differ from the aforementioned classic models [see, [Koutsoyiannis, 2000](#)]. The flexibility of CAS is illustrated in [Figure 5.1a](#) where we depict (in a log-log scale) the theoretical ACF of various HK processes, characterized by different values of Hurst coefficient,  $H$ , as well as, their approximation with CAS. The close agreement of the two theoretical models is further validated in [Figure 5.1b](#) where we plot (also in log-log scale) their climacograms (assuming  $\sigma^2 = c_0 = 1$ ), which are practically indistinguishable. It is noted that for an HK process, which exhibits simple and constant scaling laws, the slope,  $s$ , of the climacogram  $\gamma^{(k)}$ , i.e., the log-log derivative  $s := d(\ln(\gamma^{(k)}))/d(\ln(k))$ , is related with  $H$  parameter by  $s = 2H - 2$ . The resemblance of the HK and CAS is confirmed by estimating the average mean square error (MSE) of the depicted processes by means of both ACF and climacogram. In terms of ACF, the average MSE value is 0.01 and the corresponding value in terms of climacogram is 0.66.



**Figure 5.1** | a) Autocorrelation functions and b) climacograms of HK processes exhibiting different Hurst coefficients (dashed lines) and their approximation with the CAS (continuous line).

Considering the practical aspects of the identification procedure of the auto-dependence structure (e.g., estimation of the parameters of CAS or any other theoretical structure, given a sample time series), it is remarked that it is a challenging task due the fact that the sample estimates of variance and autocorrelation coefficients (i.e., empirical variance and ACF - calculated from the historical time series) are negatively biased [e.g., [Marriott and Pope, 1954](#); [Beran, 1994](#); [Koutsoyiannis, 2003, 2016, 2017](#)], especially in the presence of LRD (e.g., HK behavior). A thorough treatment on the subject lies beyond the scope of this study, as it has been extensively documented by the aforementioned authors, as well as by [Dimitriadis and Koutsoyiannis \[2015\]](#) who highlighted the advantages of using the climacogram, in comparison with the ACF and power spectrum, for the identification of the auto-dependence structure. The authors via an extended analysis of a wide range of SRD and LRD processes showed that the climacogram exhibits less uncertainty and bias in its estimation, while bias can be easily estimated *a priori*, thus providing an attractive alternative to the latter classic approaches. Further to this, the climacogram can be used as a basis for LRD identification algorithms [e.g.,



[Tyrallis and Koutsoyiannis, 2011], as well as for the development additional tools (e.g., the climacospectrum) that provide further insights regarding the asymptotic behavior of the process [Koutsoyiannis, 2016, 2017]. It is noted that in this work, the above mentioned stochastic tools (i.e., ACF and CG) are mainly employed for *diagnostic*, and not for identification purposes, i.e., to verify that the simulated processes exhibit the desired dependence properties.

### 5.3 THEORETICAL BACKGROUND OF THE MODELS

The central idea of the proposed approach is based on the Nataf's joint distribution model [NDM, Nataf, 1962] which has been originally implemented for the generation of cross-correlated, yet serially independent, random vectors with arbitrary distributions. One of its key assumptions, which consequently holds for SMARTA and CMARTA or any other Nataf-based method, is that the employed distributions owe to have finite variance. This assumption is implied throughout this work.

In this work, we employ the concept of NDM, but in a different context, i.e., for the simulation of stationary any-range-dependent stochastic processes. Particularly, the rationale of NDM is combined with an auxiliary Gaussian process in order to capture the stochastic structure (in terms of autocorrelation and cross-correlation coefficients) of the target process and simultaneously preserve the desired marginal distributions after the use of the ICDF.

Suppose that the goal is to generate a  $m$ -dimensional discrete-time stationary process  $\underline{x}_t = [x_t^1, \dots, x_t^i, \dots, x_t^m]^T$ , where  $t$  is the time index and the indices  $i, j = 1, \dots, m$  are used to refer to individual process  $x_t^i$  and  $x_t^j$  respectively. Also let,  $\mathbf{x}_t = [x_t^1, \dots, x_t^i, \dots, x_t^m]^T$  denote a realization of it. Furthermore, let us assign a target cumulative distribution function (CDF), denoted by,  $F_{x^i} := P(x^i \leq x)$  to each individual process  $x_t^i$ , and let  $\rho_{t,t+\tau}^{i,j} := \text{Corr}[x_t^i, x_{t+\tau}^j]$  denote the target Pearson's correlation coefficient between  $x_t^i$  and  $x_{t+\tau}^j$  for time lag  $\tau$ .

Likewise, and using the same notation as above, let  $\underline{z}_t = [z_t^1, \dots, z_t^i, \dots, z_t^m]^T$  be an auxiliary  $m$ -dimensional stationary standard Gaussian process with zero mean and unit variance. Also, let  $\tilde{\rho}_{t,t+\tau}^{i,j} := \text{Corr}[z_t^i, z_{t+\tau}^j]$  denote the Pearson's correlation coefficient of the auxiliary process between  $z_t^i$  and  $z_{t+\tau}^j$  for time lag  $\tau$ , hereafter, referred to as equivalent correlation coefficient. It is noted that throughout the Chapter the superscripts or subscripts of  $\rho_{t,t+\tau}^{i,j}$  or  $\tilde{\rho}_{t,t+\tau}^{i,j}$  may be omitted when possible. For example, the target autocorrelation of the process  $x_t^i$  will be denoted  $\rho_\tau^i$  and its lag- $\tau$  cross-correlation with  $x_t^j$  as  $\rho_\tau^{i,j}$ .

As mentioned earlier, the idea behind SMARTA and CMARTA models (and any other Nataf-based stochastic model; see section 4.3) lies in simulating an auxiliary standard Gaussian process  $\underline{z}_t$  using the auxiliary model (i.e., SMA or CMAR) with such parameters that after applying the inverse of their distribution function, results in a process  $\underline{x}_t$  with the desired correlation structure and marginal distributions. This mapping operation can be written as follows,

$$\underline{x}_t^i = F_{x^i}^{-1} \left( \Phi(\underline{z}_t^i) \right) \quad (5.12)$$

where  $\Phi(\cdot)$  denotes the standard normal CDF and  $F_{x^i}^{-1}(\cdot)$  stands for the ICDF of process  $x_t^i$ . An advantage of the above scheme is that since the ICDFs of the target distributions are

employed (given that they can be analytically or numerically evaluated), the process  $\underline{x}_t^i$  will inevitably have the desired marginal properties. Although, the Pearson's correlation coefficient is not invariant under such non-linear monotonic transformations [Embrechts et al., 1999], hence  $\rho_{t,t+\tau}^{i,j}$  will differ from  $\tilde{\rho}_{t,t+\tau}^{i,j}$ .

Therefore, the main challenge of such approaches, lies in identifying the *equivalent* correlations coefficients that should be used within the generation procedure (Gaussian domain) in order to attain the target correlation structure in the actual (i.e., real) domain. The relationship between equivalent and target correlations can be expressed theoretically through the following double infinite integral (see [Nataf, 1962; Liu and Der Kiureghian, 1986; Biller and Nelson, 2003], as well as section 4.3.2),

$$\rho_{t,t+\tau}^{i,j} = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}^i}^{-1}(\Phi(z_t^i)), F_{\underline{x}^j}^{-1}(\Phi(z_{t+\tau}^j)) \varphi_2(z_t^i, z_{t+\tau}^j, \tilde{\rho}_{t,t+\tau}^{i,j}) dz_t^i dz_{t+\tau}^j - E[\underline{x}^i] E[\underline{x}^j]}{\sqrt{\text{Var}[\underline{x}^i] \text{Var}[\underline{x}^j]}} \quad (5.13)$$

where  $E[\underline{x}^i], E[\underline{x}^j]$  and  $\text{Var}[\underline{x}^i], \text{Var}[\underline{x}^j]$  denote the mean and variance of  $\underline{x}^i$  and  $\underline{x}^j$  respectively; which are known from the corresponding distributions  $F_{\underline{x}^i}$  and  $F_{\underline{x}^j}$  and have to be finite. Furthermore,  $\varphi_2(z_t^i, z_{t+\tau}^j, \tilde{\rho}_{t,t+\tau}^{i,j})$  denotes the bivariate standard normal probability density function.

Inspection of Eq. (5.13) indicates that  $\rho_{t,t+\tau}^{i,j}$  is a function of  $\tilde{\rho}_{t,t+\tau}^{i,j}$ , since all other quantities are already known from the target (i.e., given) distributions  $F_{\underline{x}^i}$  and  $F_{\underline{x}^j}$ . Therefore, it can be compactly written as,

$$\rho_{t,t+\tau}^{i,j} = \mathcal{F}(\tilde{\rho}_{t,t+\tau}^{i,j} | F_{\underline{x}^i}, F_{\underline{x}^j}) \quad (5.14)$$

where  $\mathcal{F}(\cdot)$  is an abbreviation of the function defined by Eq. (5.13), which can be approximated with the use of numerical techniques such as the one employed herein (see section 4.5.1).

This relationship implies that prior to the estimation of the auxiliary Gaussian model's parameters it is essential to identify, and next use within parameter estimation, the equivalent correlations,  $\tilde{\rho}_{t,t+\tau}^{i,j}$ , that result to the target correlations,  $\rho_{t,t+\tau}^{i,j}$ , after the subsequent mapping of the auxiliary process to the actual domain. This can be achieved through inversion of Eq. (5.14), i.e.,  $\tilde{\rho}_{t,t+\tau}^{i,j} = \mathcal{F}^{-1}(\rho_{t,t+\tau}^{i,j} | F_{\underline{x}^i}, F_{\underline{x}^j})$ .

## 5.4 THE AUXILIARY GAUSSIAN MODELS

Having described the theoretical background of the proposed Nataf-based models, this section provides a brief introduction a) to the univariate and multivariate symmetric moving average (SMA) model of Koutsoyiannis [2000], which is used within SMARTA as an auxiliary standard Gaussian process model, as well as b) to the univariate and contemporaneous multivariate autoregressive (CMAR) model, which is employed within CMARTA model.

It is also remarked that instead of SMA or CMAR, any other linear stochastic model (e.g., ARMA-type) could be employed in order to mathematically describe the auxiliary Gaussian

process, yet, it is anticipated that the resulting simulation scheme will inherit its properties regarding the simulation of univariate and multivariate processes. For instance, since cross-correlation coefficients other than lag-0 are not directly modeled by these models, in the case of hydrometeorological processes characterized by strongly lagged cross-correlations (e.g., rainfall-runoff at fine time scales), it may be advantageous to employ the full MAR model (preferably, for parsimony and stability, in combination with suitable theoretical auto- and cross-correlation structures; e.g., similar to CAS), which apart from the lag-0 cross-correlations, is capable to directly model the lagged cross-correlation coefficients.

Note that the notation slightly differs from the typical one (we use the tilde representation) in order to highlight the fact the models are employed in the Gaussian domain using the equivalent correlation coefficients  $\tilde{\rho}$ , instead of the target correlation coefficients,  $\rho$ .

#### 5.4.1 The univariate SMA model

SMA model is a special case of the Backward-Forward Moving Average (BFMA) model, whose key idea is that a stochastic process  $\underline{z}_t$  can be described as a weighted sum of infinite backward and forward random variables. In practice, the model slightly relaxes the assumptions of BFMA model and assumes that a stochastic process  $\underline{z}_t$  can be described as a weighted sum of a finite number of backward and forward random variables. Particularly, the generating mechanism of the SMA model is given by the following equation,

$$\underline{z}_t = \sum_{\zeta=-q}^q \tilde{a}_{|\zeta|} \underline{v}_{t+\zeta} = \tilde{a}_q \underline{v}_{t-q} + \dots + \tilde{a}_1 \underline{v}_{t-1} + \tilde{a}_0 \underline{v}_t + \tilde{a}_1 \underline{v}_{t+1} + \dots + \tilde{a}_q \underline{v}_{t+q} \quad (5.15)$$

where  $\underline{v}_t$  are standard normal i.i.d. variables and  $\tilde{a}_\zeta$  are internal model parameters (i.e., weight coefficients) that are assumed to be symmetric, i.e.,  $\tilde{a}_\zeta = \tilde{a}_{-\zeta}$  (for  $\zeta = 1, 2, \dots$ ) and approach zero after some value  $|\zeta| > q$ , where  $q$  denotes a large positive integer value. The selection of  $q$  depends on the degree of auto-dependence imposed by the target process (e.g., through CAS; Eq. (5.8)) and the desired level of accuracy. Furthermore,  $q$  cannot be greater than the length of the time series to simulate. Particularly, the parameters  $\tilde{a}_\zeta$  are related to the autocorrelation coefficients  $\tilde{\rho}_\tau$  via a  $2q + 1$  equation system of the following form,

$$\tilde{\rho}_\tau = \sum_{\zeta=-q}^{q-\tau} \tilde{a}_{|\zeta|} \tilde{a}_{|\tau+\zeta|}, \quad \tau = 0, 1, 2, \dots, q \quad (5.16)$$

$$\tilde{\rho}_\tau = \sum_{\zeta=\tau-q}^q \tilde{a}_\zeta \tilde{a}_{\tau-\zeta}, \quad \tau = q + 1, \dots, 2q \quad (5.17)$$

If Eq. (5.16) is honored, the model resembles the theoretical ACF up to  $\tilde{\rho}_q$ , while it decays to zero after  $2q$  (see Eq. (5.17)). In order to estimate the parameters  $\tilde{a}_\zeta$ , [Koutsoyiannis \[2000\]](#) proposed two solutions, one closed-form and one based on a formulation of an optimization problem. The interested reader is referred to the above publication for a thorough and in-depth description of the two methods. In this work we restrict our attention in briefly describing only the first one since it is a fast and direct method. The aforementioned author showed that the discrete Fourier transformation (DFT) of  $\tilde{a}_\zeta$ , i.e.,  $S_{\tilde{a}}(\omega)$ , is related to the power spectrum of the autocorrelation function, i.e.,  $S_{\tilde{\rho}}(\omega)$ , by,  $S_{\tilde{a}}(\omega) = \sqrt{2S_{\tilde{\rho}}(\omega)}$ .

If the autocorrelation structure  $\tilde{\rho}_\tau$  is known (or specified), its power spectrum can be calculated using the DFT, hence estimate  $S_{\tilde{\alpha}}(\omega)$ . Then, by applying the inverse Fourier transformation one can obtain the parameters  $\tilde{\alpha}_\zeta$ . It is remarked that algorithms that facilitate these calculations are nowadays built-in in most high-level programming languages (e.g., R or MATLAB) which in turn allow the straightforward implementation of SMA and SMARTA models in most computational environments. At this point we note that [Koutsoyiannis \[2002, 2016\]](#) proposed an even simpler and straightforward procedure for the estimation of  $\tilde{\alpha}_\zeta$  coefficients, which however is applicable only for HK (i.e., fGn) type autocorrelation structures.

#### 5.4.2 The multivariate SMA model

Furthermore, the SMA model can be extended for the multivariate simulation of contemporaneously cross-correlated processes via the explicit preservation of the lag-0 cross-correlation coefficients. Particularly, let  $\mathbf{z}_t = [z_t^1, \dots, z_t^i, \dots, z_t^m]^T$  be a  $m$ -dimensional vector of  $m$  processes and  $\tilde{\rho}_\tau^{i,j} := \text{Corr}[z_t^i, z_{t+\tau}^j]$  denote the equivalent lag- $\tau$  cross-correlation between process  $z_t^i$  and  $z_t^j$  for time lag  $\tau$ . Similar to the univariate case, each process  $z_t^i$  is represented by a weighted sum of random variables  $v_t^i$ , i.e.,

$$z_t^i = \sum_{\zeta=-q}^q \tilde{\alpha}_{|\zeta|}^i v_{t+\zeta}^i \quad (5.18)$$

In this case, the random variables  $v_t^i$  are considered serially independent but contemporaneously cross-correlated. Therefore, the problem lies in generating such variables in a way that they reproduce the equivalent lag-0 cross-correlation coefficients ( $\tilde{\rho}_0^{i,j}$ ).

It has been shown that it suffices to generate random variables  $v_t^i$  with correlation  $\tilde{g}^{i,j} := \text{Corr}[v_t^i, v_t^j]$  equal to,

$$\tilde{g}^{i,j} = \frac{\tilde{\rho}_0^{i,j}}{\sum_{\zeta=-q}^q \tilde{\alpha}_{|\zeta|}^i \tilde{\alpha}_{|\zeta|}^j} \quad (5.19)$$

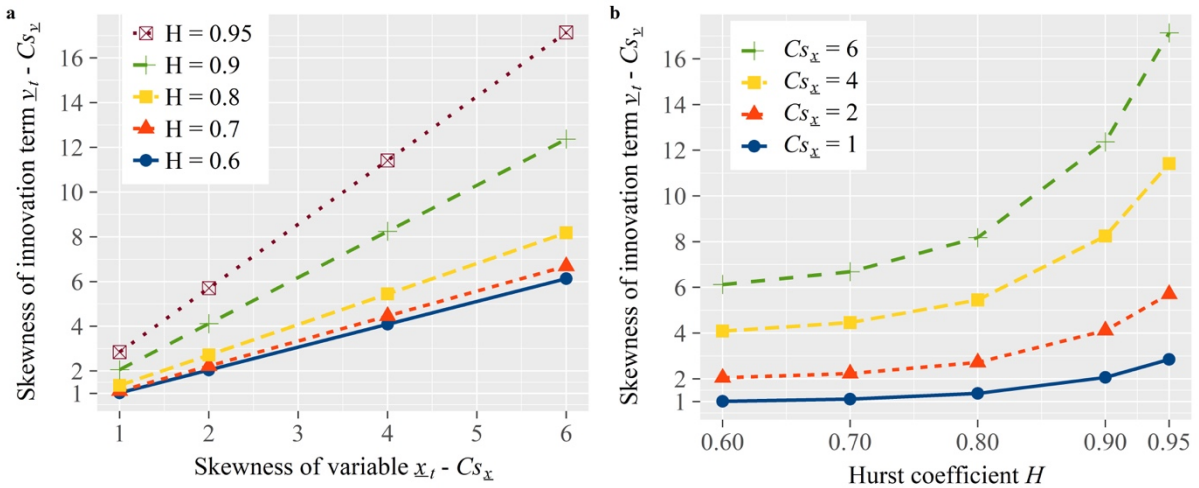
Hence, the  $(m \times m)$  correlation matrix  $\tilde{\mathbf{G}}$  is formulated, with ones in the diagonal and its  $i^{\text{th}} \neq j^{\text{th}}$  elements determined by,  $\tilde{\mathbf{G}}_{[i,j]} = \tilde{g}^{i,j}$ . Furthermore, the theoretical lag- $\tau$  cross-correlation structure of the model is given by,

$$\tilde{\rho}_\tau^{i,j} = \tilde{\rho}_0^{i,j} \frac{\sum_{\zeta=-q}^{q-\tau} \tilde{\alpha}_{|\tau+\zeta|}^i \tilde{\alpha}_{|\zeta|}^j}{\sum_{\zeta=-q}^q \tilde{\alpha}_{|\zeta|}^i \tilde{\alpha}_{|\zeta|}^j} = \tilde{g}^{i,j} \sum_{\zeta=-q}^{q-\tau} \tilde{\alpha}_{|\tau+\zeta|}^i \tilde{\alpha}_{|\zeta|}^j \quad (5.20)$$

Regarding simulation, a vector of correlated random variables  $\mathbf{v}_t = [v_t^1, \dots, v_t^i, \dots, v_t^m]^T$  can be generated by,  $\mathbf{v}_t = \tilde{\mathbf{B}}\mathbf{w}_t$ , where  $\mathbf{w}_t = [w_t^1, \dots, w_t^i, \dots, w_t^m]^T$  is a vector of standard normal i.i.d. variables, and  $\tilde{\mathbf{B}}$  is a  $m \times m$  matrix obtained by finding the so-called square root of matrix  $\tilde{\mathbf{G}}$ , i.e., Eq. (5.21). For its computation see section 5.4.5.

$$\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T = \tilde{\mathbf{G}} \quad (5.21)$$

At this point it is noted that an incidental contribution of SMARTA is the alleviation of a burden related to preservation of the skewness coefficient. As mentioned in the introduction, a broad class of linear stochastic models in an attempt to preserve the coefficients of skewness of the target process,  $\underline{x}_t$ , employ non-Gaussian white noise for the innovation term,  $\underline{v}_t$ , typically from Pearson type-III distribution. However, this practice may lead to very high coefficients of skewness for the innovation term which are hardly attainable [Todini, 1980; Koutsoyiannis, 1999]. This practice was also adopted by Koutsoyiannis [2000] in the original SMA scheme where the abovementioned distribution has been employed for the generation of skewed white noise. More specifically, as far as it concerns the univariate formulation of the SMA model (assuming  $q = 2^{10}$ ), in Figure 5.2a-b we depict (from two distinct points of view) the relationship between the skewness coefficient ( $C_{s_v}$ ) of innovation term,  $\underline{v}_t$  that is required in order to attain the target coefficient of skewness ( $C_{s_x}$ ) of the variable,  $\underline{x}_t$ , for several hypothetical HK process with characterized by different values of  $H$  coefficient. See also Eq. (29) in Koutsoyiannis [2000]. It is apparent from in Figure 5.2a-b that the higher the value of  $H$ , the higher the required skewness of the innovation term,  $\underline{v}_t$ . For example, in an HK process with  $H = 0.8$ , the skewness coefficient of innovation term  $\underline{v}_t$  has to be set twice as high as than the one of  $\underline{x}_t$ . We remark that this issue is further amplified (not shown herein) when the underlying model is used in multivariate mode [Koutsoyiannis, 1999]. On the other hand, SMARTA completely alleviates these difficulties since the SMA scheme is used as an auxiliary model in the standard Normal (i.e., Gaussian) domain and the generated data are subsequently mapped to the actual domain using the target ICDFs. Therefore, the target marginal statistics are attained without making no attempts to generate skewed innovation terms, neither in univariate nor multivariate mode. Moreover, an additional contribution of SMARTA regards the optimization problem that arises when the matrix  $\tilde{\mathbf{G}}$  is non-positive. Particularly, the problem is simplified in a nearest correlation matrix problem, since the 3<sup>rd</sup> term of Eq. (28) in Koutsoyiannis [1999], that accounts for skewness, is no longer needed.



**Figure 5.2** | Graphical illustration of the relationship between the required skewness coefficient ( $C_{s_v}$ ) of innovation term  $\underline{v}_t$  and a) the skewness ( $C_{s_x}$ ) of an fGn process  $\underline{x}_t$  for various values of  $H$  and b) the value of  $H$  of an fGn process  $\underline{x}_t$  for various values of skewness of  $C_{s_x}$  (using the SMA model with  $q = 2^{10}$ ).

### 5.4.3 The univariate AR model

An alternative, and particularly popular model for stationary processes, is the autoregressive model of order  $p$  (i.e.,  $AR(p)$ ), which has been the basis of the AutoRegressive To Anything model (ARTA; [Cario and Nelson, 1996]). A standard Gaussian  $AR(p)$  process with zero mean and unit variance can be simulated by,

$$\underline{z}_t = \sum_{l=1}^p \tilde{\alpha}_l \underline{z}_{t-l} + \underline{\varepsilon}_t \quad (5.22)$$

where  $p$  denotes the order of the model, and  $\tilde{\alpha}_l$  are the model's parameters, while  $\underline{\varepsilon}_t \sim \mathcal{N}(0, \sigma_{\underline{\varepsilon}}^2)$ . The parameters  $\tilde{\alpha}_l$  can be obtained by solving the Yule-Walker system. Specifically, given a  $p$ -dimensional vector of correlation coefficients,  $\tilde{\rho}_p = [\tilde{\rho}_1, \dots, \tilde{\rho}_p]^T$  the parameter vector  $\tilde{\alpha}_p = [\tilde{\alpha}_1, \dots, \tilde{\alpha}_p]^T$ , can be obtained by,

$$\tilde{\alpha}_p = \tilde{\mathbf{P}}_p^{-1} \tilde{\rho}_p \quad (5.23)$$

where,  $\tilde{\mathbf{P}}_p^{-1}$  denotes the inverse of  $(p \times p)$  matrix  $\tilde{\mathbf{P}}_p$  whose  $i^{\text{th}}$  and  $j^{\text{th}}$  element are  $[\tilde{\mathbf{P}}_p]_{i,j} = \tilde{\rho}_{|i-j|}$ . After the specification of  $\tilde{\alpha}_l$ ,  $\sigma_{\underline{\varepsilon}}^2$  is obtained by,  $\sigma_{\underline{\varepsilon}}^2 = 1 - \sum_{l=1}^p \tilde{\alpha}_l \tilde{\rho}_l$ . A stationary  $AR(p)$  process reproduces the autocorrelation structure of the process up to lag  $p$ , while for  $\tau \geq p + 1$  its correlation structure is given by,  $\tilde{\rho}_\tau = \tilde{\alpha}_1 \tilde{\rho}_{\tau-1} + \tilde{\alpha}_2 \tilde{\rho}_{\tau-2} + \dots + \tilde{\alpha}_p \tilde{\rho}_{\tau-p}$ , or more compactly, by,  $\tilde{\rho}_\tau = \sum_{l=1}^p \tilde{\alpha}_l \tilde{\rho}_{\tau-l}$ .

As a side note let us provide an additional relationship that will be subsequently used within the parameter estimation procedure of the auxiliary Gaussian CMAR model. According to Wold's representation theorem any covariance stationary causal process can be written as a general linear process, i.e., as a weighted linear combination of past and present i.i.d. random variables  $\underline{w}_t$ .

$$\underline{z}_t = \psi_0 \underline{w}_t + \psi_1 \underline{w}_{t-1} + \psi_2 \underline{w}_{t-2} + \dots = \sum_{\zeta=0}^{\infty} \psi_\zeta \underline{w}_{t-\zeta} \quad (5.24)$$

where  $\psi_\zeta$  are weight coefficients. This representation is also known as infinite moving average representation, i.e.,  $MA(\infty)$ . It can be shown that  $\psi_\zeta$  are related with the coefficients  $\tilde{\alpha}_\zeta$  of  $AR(p)$  model by [e.g., Cryer and Chan, 2008; Shumway and Stoffer, 2017],

$$\begin{aligned} \psi_0 &= 1 \\ \psi_1 &= \tilde{\alpha}_1 \\ \psi_2 &= \tilde{\alpha}_2 + \tilde{\alpha}_1 \psi_1 \\ &\vdots \\ \psi_\zeta &= \tilde{\alpha}_p \psi_{\zeta-p} + \tilde{\alpha}_{p-1} \psi_{\zeta-p+1} + \dots + \tilde{\alpha}_1 \psi_{\zeta-1} \end{aligned} \quad (5.25)$$

or more compactly,

$$\psi_\zeta = \sum_{l=1}^{\zeta} \tilde{\alpha}_l \psi_{\zeta-l}, \quad \text{for } \zeta = 1, 2, \dots \quad (5.26)$$

where  $\psi_0 = 1$  and  $\tilde{a}_\zeta = 0$  for  $\zeta > p$ . It is also noted that a similar relationship exists for ARMA-type models. Nevertheless, since  $\psi_j$  decay with increasing  $\zeta$  and approach zero after some large value of  $\zeta$  it is possible to truncate Eq. (5.24) at some large value  $q$  to read,

$$\underline{z}_t = \sum_{\zeta=0}^q \psi_\zeta \underline{w}_{t-\zeta} \quad (5.27)$$

#### 5.4.4 The multivariate AR model

The univariate AR( $p$ ) model can be extended for multivariate processes [e.g., Pegram and James, 1972; Kottogoda, 1980; Bras and Rodríguez-Iturbe, 1985; Cryer and Chan, 2008; Shumway and Stoffer, 2017], and it is often referred to as multivariate or vector autoregressive (MAR( $p$ ) or VAR( $p$ )) model. Assuming that we wish to model a  $m$ -dimension vector of Gaussian processes  $\underline{z}_t = [z_t^1, \dots, z_t^m]^T$  with zero and unit variance, its generating equation is given by,

$$\underline{z}_t = \tilde{A}_1 \underline{z}_{t-1} + \tilde{A}_2 \underline{z}_{t-2} + \dots + \tilde{A}_p \underline{z}_{t-p} + \underline{\varepsilon}_t = \sum_{l=1}^p \tilde{A}_l \underline{z}_{t-l} + \underline{\varepsilon}_t \quad (5.28)$$

where  $p$  denotes the order of the model,  $\tilde{A}_l$  are  $(m \times m)$  parameter matrices and  $\underline{\varepsilon}_t = [\varepsilon_t^1, \dots, \varepsilon_t^m]^T$  is a vector of  $m$  Gaussian variates with zero mean and covariance matrix  $\tilde{G} := \text{Cov}[\underline{\varepsilon}_t, \underline{\varepsilon}_t]$  (whose  $i^{\text{th}}$   $j^{\text{th}}$  element is denoted by  $\tilde{g}^{i,j}$ ). The correlation (since we assume a standard Gaussian model) matrix of time lag  $\tau$ , is denoted by  $\tilde{R}_\tau := \text{Corr}[\underline{z}_t, \underline{z}_{t-\tau}]$ , and is related with the parameter matrices  $\tilde{A}_l$  by,

$$\tilde{R}_\tau - \tilde{A}_1 \tilde{R}_{\tau-1} - \dots - \tilde{A}_p \tilde{R}_{\tau-p} = \begin{cases} \tilde{G} & \text{if } \tau = 0 \\ \mathbf{0} & \text{if } \tau > 0 \end{cases} \quad (5.29)$$

Specifically, for  $\tau = 0$ , the system reads,

$$\tilde{G} = \tilde{R}_0 - \tilde{A}_1 \tilde{R}_1^T - \dots - \tilde{A}_p \tilde{R}_p^T = \tilde{R}_0 - \sum_{l=1}^p \tilde{A}_l \tilde{R}_l^T \quad (5.30)$$

Furthermore, Eq. (5.29) can be written in a matrix notation as follows (for  $\tau = 1, \dots, p$ ),

$$[\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_p] = [\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_p] \begin{bmatrix} \tilde{R}_0 & \tilde{R}_1 & \dots & \tilde{R}_{p-1} \\ \tilde{R}_1^T & \tilde{R}_0 & \dots & \tilde{R}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{R}_{p-1}^T & \tilde{R}_{p-2}^T & \dots & \tilde{R}_0 \end{bmatrix} \quad (5.31)$$

where  $\tilde{R}_{-\tau} = \tilde{R}_\tau^T$ . Eq. (5.31) is also known as the multivariate Yule-Walker system of MAR( $p$ ) model. Provided that the matrices  $\tilde{R}_1, \tilde{R}_2, \dots, \tilde{R}_p$  are known, the Yule-Walker system of MAR( $p$ ) can be solved for  $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_p$ , i.e.,

$$[\tilde{\mathbf{A}}_1, \tilde{\mathbf{A}}_2, \dots, \tilde{\mathbf{A}}_p] = [\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_p] \begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_{p-1} \\ \tilde{\mathbf{R}}_1^T & \tilde{\mathbf{R}}_0 & \dots & \tilde{\mathbf{R}}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{R}}_{p-1}^T & \tilde{\mathbf{R}}_{p-2}^T & \dots & \tilde{\mathbf{R}}_0 \end{bmatrix}^{-1} \quad (5.32)$$

Arguably, this is a complex system of equations that requires the specification of  $p$  matrices  $\tilde{\mathbf{R}}_p$ . The overall parameter estimation procedure can be significantly simplified if we assume that the parameter matrices  $\tilde{\mathbf{A}}_1, \tilde{\mathbf{A}}_2, \dots, \tilde{\mathbf{A}}_p$  are diagonal, i.e.,

$$\tilde{\mathbf{A}}_l = \begin{bmatrix} \tilde{\alpha}_{l[1,1]} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \tilde{\alpha}_{l[m,m]} \end{bmatrix} = [\tilde{\mathbf{A}}_l]_{i,j} \quad (5.33)$$

Thereby formulating the so-called contemporaneous multivariate autoregressive model of order  $p$  (i.e., CMAR( $p$ ); [see, *Pegram and James, 1972*]). This simplification apart from the reproduction of the autocorrelation structure of the process up to time lag  $p$  (as in the case of full matrices  $\tilde{\mathbf{A}}_l$ ), implies the direct reproduction of the lag-0 cross-correlation structure, i.e., correlation matrix  $\tilde{\mathbf{R}}_0$ . Using the contemporaneous formulation, the model can be decomposed into  $m$  univariate AR( $p$ ) models, which are contemporaneously cross-correlated at lag 0, i.e.,

$$\begin{aligned} \underline{z}_t^1 &= \tilde{\alpha}_{1[1,1]} \underline{z}_{t-1}^1 + \tilde{\alpha}_{2[1,1]} \underline{z}_{t-2}^1 + \dots + \tilde{\alpha}_{l[1,1]} \underline{z}_{t-l}^1 + \dots + \tilde{\alpha}_{p[1,1]} \underline{z}_{t-p}^1 + \underline{\varepsilon}_t^1 \\ &\quad \vdots \\ \underline{z}_t^m &= \tilde{\alpha}_{1[m,m]} \underline{z}_{t-1}^m + \tilde{\alpha}_{2[m,m]} \underline{z}_{t-2}^m + \dots + \tilde{\alpha}_{l[m,m]} \underline{z}_{t-l}^m + \dots + \tilde{\alpha}_{p[m,m]} \underline{z}_{t-p}^m + \underline{\varepsilon}_t^m \end{aligned} \quad (5.34)$$

Alternatively, and assuming that  $\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T = \tilde{\mathbf{G}}$ , where  $\tilde{\mathbf{B}}$  is a  $m \times m$  matrix that denotes the square root matrix of  $\tilde{\mathbf{G}}$  (for its computation see section 5.4.5.), Eq. (5.34) can be rewritten as,

$$\underline{z}_t^i = \sum_{l=1}^p \tilde{\alpha}_{l[i,i]} \underline{z}_{t-l}^i + \sum_{j=1}^m \tilde{b}_{[i,j]} \underline{w}_t^j \quad (5.35)$$

where  $\underline{w}_t^j$  are i.i.d. standard Gaussian variates (i.e.,  $\underline{w}_t^j \sim \mathcal{N}(0,1)$ ).

In this form, and assuming that the autocorrelation structure of each process is known (e.g., specified by a theoretical model such as CAS), the parameters  $\alpha_l$  ( $l = 1, \dots, p$ ), as well as the variance ( $\sigma_{\underline{\varepsilon}}^2$ ) of  $\underline{\varepsilon}_t$ , can be easily computed through the univariate Yule-Walker system. Hence it is possible to fully estimate the matrices  $\tilde{\mathbf{A}}_1, \tilde{\mathbf{A}}_2, \dots, \tilde{\mathbf{A}}_p$  as well as the diagonal elements of  $\tilde{\mathbf{G}}$ , which are,  $\tilde{g}^{i,i} = \text{Var}[\underline{\varepsilon}_t^i, \underline{\varepsilon}_t^i] = \sigma_{\underline{\varepsilon}}^2$ .

According to *Pegram and James [1972]*, in order to estimate the off-diagonal elements of  $\tilde{\mathbf{G}}$  one can resort either to iterative methods or solve a complicated system of equations. Both solutions experience difficulties, especially when implemented in a computer software. Herein, we propose an alternative technique, which to the best of our knowledge is new. It is reminded that according to Eq. (5.24) each individual process  $\underline{z}_t^i$  can be represented in terms of a MA( $\infty$ ) process, which can be truncated in some high value of  $q$ , i.e.,



$$\underline{z}_t^i = \sum_{\zeta=0}^q \psi_{\zeta}^i \underline{w}_{t-\zeta}^i \quad (5.36)$$

The elements  $\psi_{\zeta}^i$  can be easily computed for each process  $\underline{z}_t^i$  using Eq. (5.25) or (5.26). Provided that the  $\psi_{\zeta}^i$  quantities are estimated, the off-diagonal  $i^{\text{th}} j^{\text{th}}$  elements (for  $i, j = 1, \dots, m$  and  $i \neq j$ ; since the diagonal elements are known) of matrix  $\tilde{\mathbf{G}}$  are identified as follows,

$$\tilde{g}^{i,j} = \frac{[\tilde{\mathbf{R}}_0]_{i,j}}{\sum_{\zeta=0}^q \psi_{\zeta}^i \psi_{\zeta}^j} = \frac{\tilde{\rho}_0^{i,j}}{\sum_{\zeta=0}^q \psi_{\zeta}^i \psi_{\zeta}^j} \quad (5.37)$$

It is also noted that the elements  $\psi_{\zeta}^i$  can be used for the estimation of any cross-correlation value for lag  $\tau = 0, 1, 2 \dots$  through,

$$\text{Cov}[\underline{z}_t^i, \underline{z}_{t+\tau}^j] = [\tilde{\mathbf{R}}_0]_{i,j} \frac{\sum_{\zeta=0}^{q-\tau} \psi_{\zeta}^i \psi_{\zeta+\tau}^j}{\sum_{\zeta=0}^q \psi_{\zeta}^i \psi_{\zeta}^j} = \tilde{\rho}_0^{i,j} \frac{\sum_{\zeta=0}^{q-\tau} \psi_{\zeta}^i \psi_{\zeta+\tau}^j}{\sum_{\zeta=0}^q \psi_{\zeta}^i \psi_{\zeta}^j} = \tilde{g}^{i,j} \sum_{\zeta=0}^{q-\tau} \psi_{\zeta}^i \psi_{\zeta+\tau}^j \quad (5.38)$$

#### 5.4.5 A note on the computation of the square root matrix

Both of the aforementioned models, when employed in multivariate mode, involve the problem of find the square root of a matrix, i.e.,  $\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T = \tilde{\mathbf{G}}$ . This can be obtained by standard decomposition (e.g., Cholesky or singular value decomposition) or optimization-based methods [Koutsoyiannis, 1999; Higham, 2002]. Specifically, if  $\tilde{\mathbf{G}}$  is positive definite (which indicates that the multivariate process is admissible), the problem has infinite solutions hence, both decomposition and optimization-based methods can be employed. On the other hand, when  $\tilde{\mathbf{G}}$  is non-positive definite (implying that the multivariate process is inadmissible), the former methods cannot offer a solution. In this case, optimization-based techniques can provide a potential remedy, by formulating an optimization problem, where the objective is to identify a matrix  $\tilde{\mathbf{B}}^*$  which results to a feasible and near-to-optimum matrix  $\tilde{\mathbf{G}}^* := \tilde{\mathbf{B}}^* \tilde{\mathbf{B}}^{*T}$  which is as closest (typically quantified in terms of some distance measure; e.g., Euclidean norm) as possible to the original matrix  $\tilde{\mathbf{G}}$ . Of course, in such cases, the target process will not be exactly resembled, while, the difference between  $\tilde{\mathbf{G}}$  and  $\tilde{\mathbf{G}}^*$  can be regarded as a proxy for the magnitude of approximation introduced to the simulation. Bras and Rodríguez-Iturbe [1985 p. 98], as well as Koutsoyiannis [1999] discuss several situations which may lead to a non-positive definite matrix  $\tilde{\mathbf{G}}$ . Almost all of these situations are related with the estimates of correlation coefficients from the empirical data. In the case of SMARTA and CMARTA, and provided that a feasible autocorrelation structure has been identified for each individual process, a non-positive definite matrix  $\tilde{\mathbf{G}}$  may arise due to, data-based estimates of lag-0 cross-correlation coefficients, imprecise approximation of equivalent correlation coefficients or incompatible combinations of marginal distributions, autocorrelation structures and target cross-correlations (see section 4.1.2). For instance, since the proposed scheme (in multivariate mode) treats each individual process separately of the cross-correlations, the simulation of highly cross-correlated processes with particularly different distributions and autocorrelation structures (e.g., very fast-decaying and very slow-decaying) may be infeasible (see section 5.6.1.2 for a simulation example using SMARTA, involving both positively and negative cross-correlated LRD and SRD processes), even if the employed autocorrelation structures are individually valid.

## 5.5 GENERATION PROCEDURE

Having described in detail all the key components of SMARTA and CMARTA models in the previous sections, we useful to provide the complete generation procedure in a step-by-step manner. The procedure is similar for the two models (it differs only on step 4), and can be decomposed into the following six steps:

**Step 1.** Define a target distribution  $F_{\underline{x}^i}$  for each process  $\underline{x}_t^i$ ;  $i = 1, \dots, m$ . SMARTA and CMARTA, as well as all Nataf-based methods, are flexible in terms of distribution fitting method; hence one can select a fitting method of their preference.

**Step 2.** Define a target auto-correlation structure  $(\rho_t^i)$  for each process  $\underline{x}_t^i$ ;  $i = 1, \dots, m$  using a theoretical ACF model. For instance, for each process  $\underline{x}_t^i$  identify the parameters of CAS that better fit the observed data.

**Step 3.** Identify the equivalent correlation coefficients  $(\tilde{\rho}_t^i)$  of each theoretical ACF, up to the maximum specified lag (which depends on the type of the process; LRD or SRD), for each process  $\underline{x}_t^i$ ;  $i = 1, \dots, m$ . Furthermore, in the multivariate case, the lag-0 equivalent cross-correlation coefficient  $\tilde{\rho}_0^{i,j}$  between processes,  $\underline{x}_t^i$  and  $\underline{x}_t^j$ ;  $i \neq j = 1, \dots, m$  should be also determined. Assuming that the algorithm of section 4.5.1 is employed for the identification of equivalent correlations, and given the fact that it allows the direct estimation of the equivalent ACF up to any lag, the algorithm has to be employed  $m$  times, one for each process  $\underline{x}_t^i$ ;  $i = 1, \dots, m$ . Furthermore, in order to estimate the lag-0 equivalent cross-correlation coefficient  $\tilde{\rho}_0^{i,j}$ , the same procedure should be employed  $m(m - 1)/2$  additional times. For instance, in a 4-dimensional problem ( $m = 4$ ), the algorithm of section 4.5.1 is executed in total,  $m(m + 1)/2$  times (=10).

**Step 4. (SMARTA)** Calculate the parameters of the univariate auxiliary SMA model (section 5.4.1), i.e., the weight coefficients  $(\tilde{a}_\zeta^i)$  of each auxiliary process  $\underline{z}_t^i$ ;  $i = 1, \dots, m$ . Additionally, in the multivariate case (section 5.4.2), calculate the elements of matrices  $\tilde{\mathbf{G}}$  and  $\tilde{\mathbf{B}}$  (see also, Eq. (5.20) and (5.21)).

**Step 4. (CMARTA)** Calculate the parameters of the auxiliary AR model (section 5.4.3), i.e., the autoregressive coefficients  $(\tilde{a}_\zeta^i)$  of each auxiliary process  $\underline{z}_t^i$ ;  $i = 1, \dots, m$ . Additionally, in the multivariate case (section 5.4.4), establish the MA( $\infty$ ) representation (by truncating it to some large value  $q$ ; see Eq. (5.25), Eq. (5.26) and Eq. (5.37)) of each process and obtain the quantities  $\psi_\zeta^i$ . Subsequently estimate the elements of matrices  $\tilde{\mathbf{G}}$  (see Eq. (5.37)) and  $\tilde{\mathbf{B}}$  (see also, Eq. (5.35)).

It is noted that the weight coefficients  $(\tilde{a}_\zeta^i)$  of each auxiliary SMA model and the autoregressive coefficients  $(\tilde{a}_\zeta^i)$  of each auxiliary AR model, are not essentially model parameters, since the correlation structure to simulate has already specified by a theoretical model (e.g., CAS; Step 2). Thereby, high order model specifications are parameter-parsimonious and the order of the model solely controls the degree of resemblance of the target correlation structure.

**Step 5.** Employ the auxiliary Gaussian SMA or CMAR model and generate a realization of the auxiliary univariate  $(\underline{z}_t)$  or multivariate process  $(\underline{\mathbf{z}}_t)$ .

**Step 6.** Attain the actual process  $\underline{x}_t$  (or  $\underline{x}_t$ ), by mapping the auxiliary Gaussian process  $\underline{z}_t$  (or  $\underline{z}_t$ ) to the actual domain using the ICDF,  $F_{\underline{x}^i}^{-1}$ , of each process  $\underline{x}_t^i$ ;  $i = 1, \dots, m$ , via Eq. (5.12).

## 5.6 HYPOTHETICAL SIMULATION STUDIES

Prior employing real-world datasets to demonstrate the developed approaches we decided to setup a series of hypothetical simulation studies (univariate and multivariate) of processes characterized by a variety of target distribution functions (continuous, discrete or mixed-type). The motivation regarding this choice was based on conducting experiments where all the assumptions are *a priori* known, hence allowing the comprehensive evaluation and assessment of the models without the effect of exogenous factors, such as, erroneous or short length historical data. However, it is remarked that the proposed models (SMARTA and CMARTA) are generic, and can be directly applied for the simulation of univariate and multivariate stationary processes (e.g., geophysical, hydrometeorological and beyond). In this vein, in section 5.7 we focus on SMARTA model whose applicability is demonstrated using two real-world datasets, one that concerns the simulation of annual non-Gaussian streamflow at four stations and another that involves the simulation of intermittent, non-Gaussian, daily rainfall at a single location

### 5.6.1 SMARTA model

#### 5.6.1.1 Simulation of univariate processes

The first simulation study constitutes a comparison between the original SMA and the proposed SMARTA models (with  $q = 2^{12}$  for both) for the simulation of long (i.e.,  $2^{20}$  time steps) univariate HK processes (i.e., fGn), exhibiting different Hurst coefficients, i.e.,  $H \in \{0.6, 0.7, 0.8, 0.9\}$  and Pearson type-III marginal distribution ( $\mathcal{P}III$ ). Towards this, we identified a total of 4 scenarios each one characterized by  $\mathcal{P}III$  and different  $H$  coefficients. It is reminded that the original SMA model, in order to approximate the marginal statistics employs  $\mathcal{P}III$  variates for the innovation term (hence hereafter referred to as SMA- $\mathcal{P}III$ ), while SMARTA uses the ICDF of the target distribution; in this case  $\mathcal{P}III$ . The rationale regarding the selection of this distribution intended at conducting a fair and meaningful comparison among the two models, which in this formulation have exactly the same number of parameters. i.e., three for the marginal distribution (see, Eq. (5.39)) and one (i.e.,  $H$ ) for the autocorrelation structure. It is pointed out, that the comparison does not intend to infer which model is the best, rather is used as a benchmark to highlight the merits of the proposed approach.  $\mathcal{P}III$  is essentially a Gamma distribution ( $\mathcal{G}$ ; see, Eq. (4.17)) with an additional location (else known as threshold or shift) parameter, whose PDF is given by,

$$f_{\mathcal{P}III}(x; a, b, c) = \frac{1}{|b| \Gamma(a)} \left( \frac{x-c}{b} \right)^{a-1} \exp\left(-\frac{x-c}{b}\right), \begin{cases} \text{if } b > 0 & c \leq x < \infty \\ \text{if } b < 0 & -\infty < x \leq c \end{cases} \quad (5.39)$$

where  $\Gamma(\cdot)$  denotes the gamma function, while,  $a > 0$ ,  $b \neq 0$  and  $c \in \mathbb{R}$  are shape, scale and location parameters, respectively; and they are interconnected with the mean ( $\mu_{\underline{x}}$ ), variance ( $\sigma_{\underline{x}}^2$ ), skewness ( $C_{s_{\underline{x}}}$ ) and kurtosis ( $C_{k_{\underline{x}}}$ ) coefficients of random variable  $\underline{x}$  by,

$$\mu_{\underline{x}} = c + ab, \quad \sigma_{\underline{x}}^2 = ab^2, \quad C_{s_{\underline{x}}} = \frac{2b}{|b|\sqrt{a}}, \quad C_{k_{\underline{x}}} = \frac{6}{a} + 3 \quad (5.40)$$

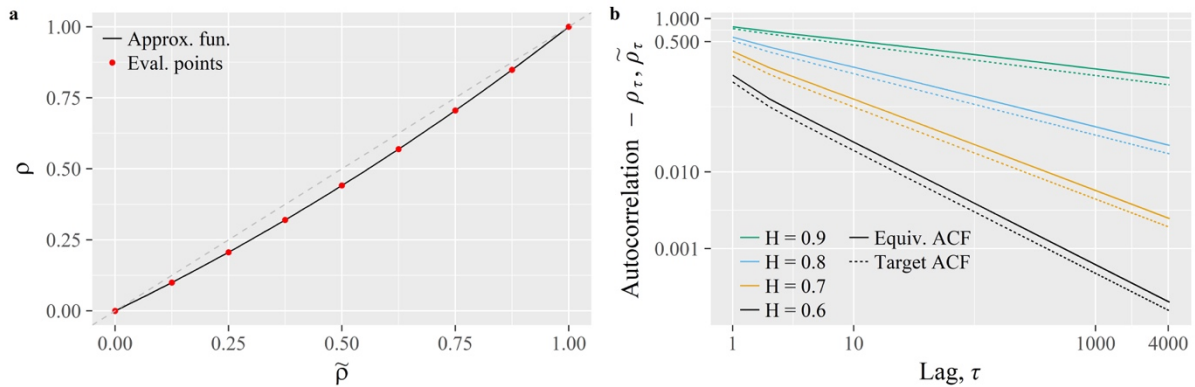
More specifically, in all scenarios, we employed a  $\mathcal{P}$ III distribution with parameters  $a = 0.75614$ ,  $b = 11.5$  and  $c = 1.30434$  whose theoretical moments are given in **Table 5-1**.

**Table 5-1** | Summary of theoretical and simulated statistics as reproduced by SMA and SMARTA models.

| Scenario                  | Theoretical            | Simulated (SMA- $\mathcal{P}$ III) |         |         |         | Simulated (SMARTA) |         |         |         |
|---------------------------|------------------------|------------------------------------|---------|---------|---------|--------------------|---------|---------|---------|
|                           | All                    | $H=0.6$                            | $H=0.7$ | $H=0.8$ | $H=0.9$ | $H=0.6$            | $H=0.7$ | $H=0.8$ | $H=0.9$ |
| Mean ( $\mu$ )            | 10                     | 9.99                               | 10.08   | 9.85    | 10.23   | 10.00              | 9.99    | 9.99    | 10.00   |
| Variance ( $\sigma^2$ )   | 100                    | 100.61                             | 100.78  | 100.04  | 99.79   | 100.03             | 99.86   | 100.07  | 101.65  |
| Skewness coeff. ( $C_s$ ) | 2.30                   | 2.35                               | 2.34    | 2.32    | 2.35    | 2.30               | 2.29    | 2.30    | 2.35    |
| Kurtosis coeff. ( $C_k$ ) | 10.93                  | 11.43                              | 11.80   | 12.62   | 15.97   | 10.94              | 10.85   | 11.00   | 11.53   |
| Hurst coeff. ( $H$ )      | 0.60, 0.70, 0.80, 0.90 | 0.61                               | 0.70    | 0.80    | 0.89    | 0.60               | 0.71    | 0.80    | 0.90    |

\*The theoretical moments correspond to  $\mathcal{P}$ III distribution ( $a = 0.75614$ ,  $b = 11.5$  and  $c = 1.30434$ ).

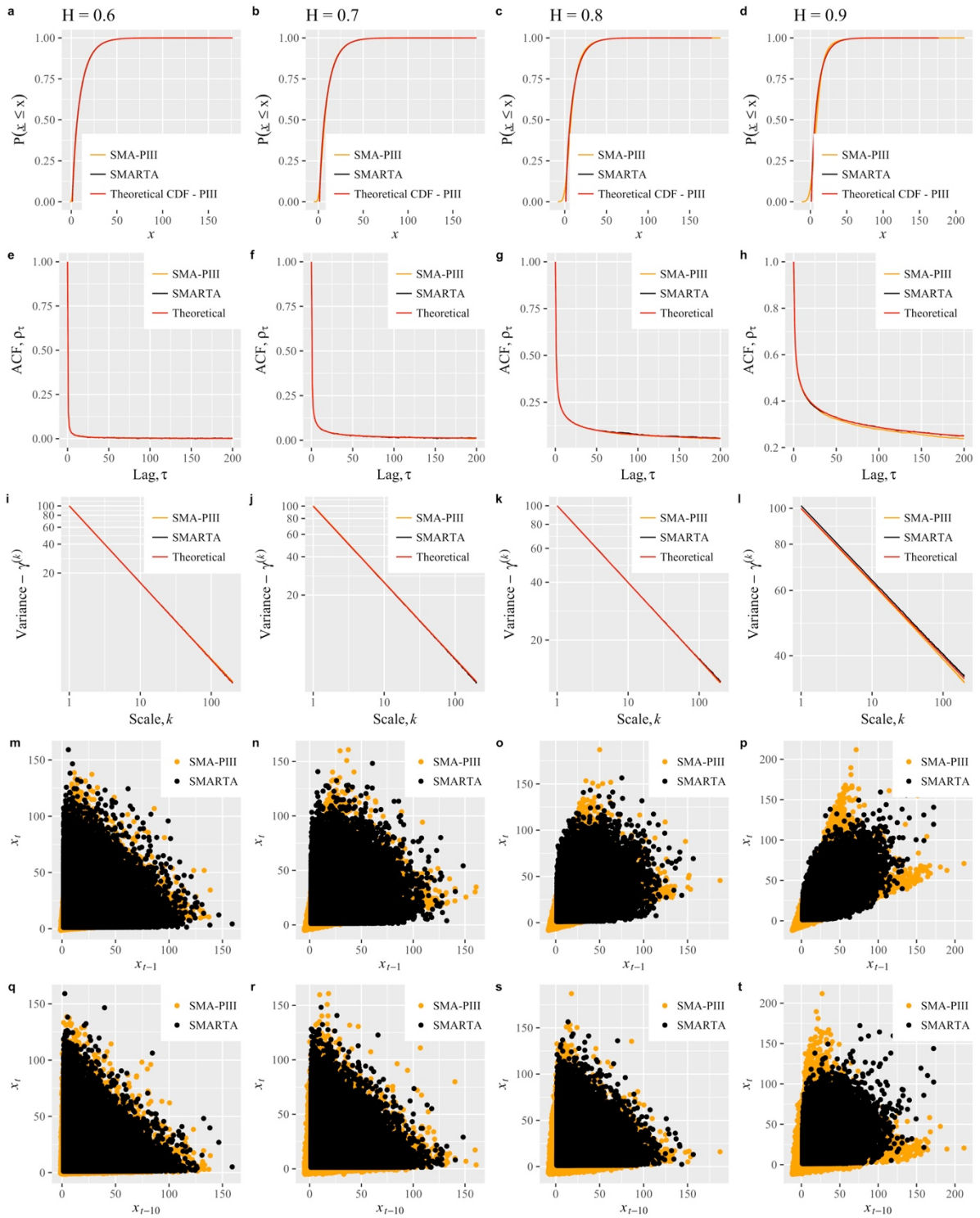
Regarding SMARTA and the given marginal distribution, **Figure 5.3a** illustrates the relationship between the equivalent correlation coefficients  $\tilde{\rho}$  and the target ones  $\rho$  (the superscripts are omitted for simplicity), while **Figure 5.3b** depicts the equivalent autocorrelation coefficients  $\tilde{\rho}_\tau$  employed by SMARTA in order to capture the target autocorrelation structure  $\rho_\tau$  of the target HK processes.



**Figure 5.3** | a) The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients. b) Comparison between the target and equivalent autocorrelation coefficients employed within the SMARTA model for HK processes with the various values of  $H$ .

**Table 5-1** presents the simulated (by the two approaches) first four moments; which are apparently well-captured by both models. It is noted that while SMA does not explicitly accounts for the kurtosis coefficient it is able to reproduce it in a satisfactory degree; especially when one considers the high uncertainty associated with its estimation [cf., *Lombardo et al., 2014*]. Nevertheless, it is reminded that the resemblance of the moments does not imply the reproduction of the marginal distribution [*Matalas and Wallis, 1976*]. This is depicted in **Figure 5.4a-d** where we compare the target theoretical cumulative distribution (CDF) with the empirically derived cumulative density functions (ECDFs) of the two models. In this case, only SMARTA was able to reproduce the target distribution, regardless of the value of  $H$  coefficient (its ECDF is almost indistinguishable with the theoretical one). On the other hand, the ECDF of SMA- $\mathcal{P}$ III departs from the theoretical one for high values of  $H$  (e.g., see **Figure 5.4d**). Furthermore, SMARTA explicitly avoids the generation of negative values; since the target distribution ( $\mathcal{P}$ III) is positively bounded at  $c = 1.30434$ . A property of high importance in hydrology due to the (often) non-negative nature of such variables (e.g., streamflow and precipitation). Regarding the resemblance of the auto-dependence structure of the processes,

it is apparent from **Figure 5.4e-h** and **Figure 5.4i-l** that both models were able to reproduce the theoretical HK ACFs as well as the corresponding climacograms, even for high values of  $H$ . These graphs also provide an empirical evidence of the theoretical consistency of both approaches. In addition, the Hurst coefficient of the synthetic realizations (see **Table 5-1**) was estimated using the climacogram-based, least squares variance (LSV) method [Tyrallis and Koutsoyiannis, 2011] and are in agreement with the theoretical values. Finally, in order to visually assess the form of the established dependencies, for both models and each HK process (i.e., scenario), we employ scatter plots of the lagged synthetic data for  $\tau = 1$  (**Figure 5.4m-p**) and  $\tau = 10$  (**Figure 5.4q-t**). It is observed that despite the fact that both models reproduced the same autocorrelation coefficient for  $\tau = 1$  and  $\tau = 10$  they establish particularly different dependence patterns. This is attributed to the underlying assumption of SMARTA regarding the joint behavior of the process which is related to the Gaussian copula (expressed through the auxiliary Gaussian model; see also [Tsoukalas et al., 2018a], as well as section 3.2.4).



**Figure 5.4** | Comparison between theoretical and simulated CDFs (using the Weibull’s plotting position) of SMA-P<sub>III</sub> and SMARTA models for HK processes with a)  $H = 0.6$ , b)  $H = 0.7$ , c)  $H = 0.8$ , d)  $H = 0.9$ . Comparison between theoretical (HK) and empirical ACF of SMA-P<sub>III</sub> and SMARTA models for HK processes with e)  $H = 0.6$ , f)  $H = 0.7$ , g)  $H = 0.8$ , h)  $H = 0.9$ . Comparison between theoretical and empirical climacograms of SMA-P<sub>III</sub> and SMARTA models for HK processes with i)  $H = 0.6$ , j)  $H = 0.7$ , k)  $H = 0.8$ , l)  $H = 0.9$ . Scatter plots of SMA-P<sub>III</sub> and SMARTA models for time lag  $\tau = 1$  for simulated HK processes with m)  $H = 0.6$ , n)  $H = 0.7$ , o)  $H = 0.8$ , p)  $H = 0.9$ . Scatter plots of SMA-P<sub>III</sub> and SMARTA models for time lag  $\tau = 10$  for simulated HK processes with q)  $H = 0.6$ , r)  $H = 0.7$ , s)  $H = 0.8$ , t)  $H = 0.9$ .

### 5.6.1.2 Simulation of multivariate processes

To further elaborate on the SMARTA approach, we setup a multivariate problem that concerns the simultaneous generation of four contemporaneously cross-correlated SRD and LRD processes. These may be seen as four (4) different processes at the same site or processes of the same variable at 4 different sites. Hereinafter, we consider the latter for convenience and refer to them as sites A-D, as well as model them in that order, i.e., as 4-dimensional stationary process  $\underline{x}_t = [\underline{x}_t^1, \underline{x}_t^2, \underline{x}_t^3, \underline{x}_t^4]^T$ , where for instance,  $i = 3$  refers to site C. In this demonstration, the target auto-dependence structure of each process is described by the two-parameter CAS (i.e., Eq. (5.8)). More specifically, sites A and B are characterized by LRD behavior (particularly HK, since we set  $\beta > 1$  and  $\kappa = \kappa_0$ ) and slowly-decaying ACF, while sites C and D by SRD (since we set  $\beta = 0$  and  $H = 0.5$ ) and fast-decaying ACF. In addition, we assigned different target distributions to the sites A-D, i.e., Burr type-XII ( $\mathcal{B}rXII$ ; [Burr, 1942; Singh and Maddala, 1976; Tadikamalla, 1980]; Eq. (5.41)), Pearson Type-III ( $\mathcal{P}III$ ; Eq. (5.39)), two-parameter Log-Normal ( $\mathcal{LN}$ ; Eq. (4.48)) and Weibull ( $\mathcal{W}EJ$ ; Eq. (5.42)). The PDF of the Burr type-XII distribution is given by,

$$f_{\mathcal{B}rXII}(x; a_1, a_2, b) = \left(\frac{a_1 a_2}{b}\right) \left(\frac{x}{b}\right)^{a_1-1} \left(1 + \left(\frac{x}{b}\right)^{a_1}\right)^{-a_2-1}, \quad x > 0 \quad (5.41)$$

where  $a_1, a_2 > 0$  are shape parameters and  $b > 0$  is a scale parameter. It is noted that  $\mathcal{B}rXII$  is a power-type distribution and its  $r^{\text{th}}$  moment exist if and only if  $a_1 a_2 > r$ . Furthermore, the PDF of the Weibull reads as follows,

$$f_{\mathcal{W}EJ}(x; a, b) = \left(\frac{a}{b}\right) \left(\frac{x}{b}\right)^{a-1} \exp\left(-\left(\frac{x}{b}\right)^a\right), \quad x \geq 0 \quad (5.42)$$

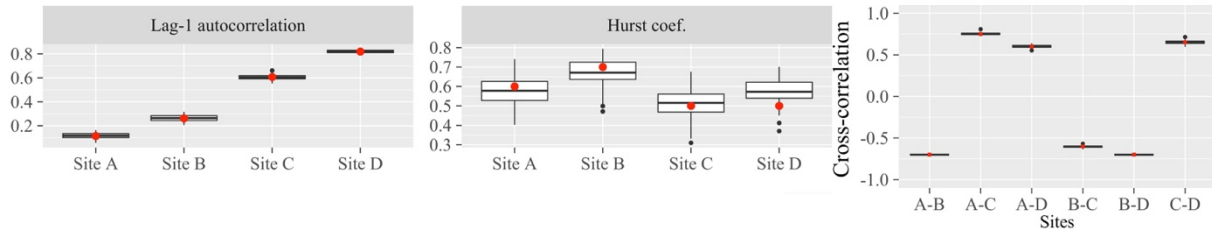
where  $a > 0$  and  $b > 0$  are shape and scale parameters respectively. **Table 5-2a** provides a synopsis of all assumptions, as well as contains the parameters of CAS and the theoretical moments of the corresponding distributions. Note that the Kurtosis coefficient of site A is infinite, since  $a_1 a_2 < 4$ . Further to this, the target and equivalent lag-0 cross-correlation coefficients (involving both positive and negative ones) are given in **Table 5-2b**. It is apparent that this is a peculiar simulation scenario, which was devised in order stress-test the SMARTA method.

In order to provide further insights regarding the consistency of the model, we generated 100 independent realizations with length  $2^{11}$  time steps and set the number of SMARTA model's internal weight coefficients equal to,  $q = 2^{10}$ . **Figure 5.5** provides a synopsis of some basic dependence statistics in terms of box-plots. SMARTA resembled with high precision the lag-1 autocorrelation and lag-0 cross-correlation coefficients (including the negative ones) despite the fact that the target processes are characterized by very different auto-dependence structures and distribution functions. Additionally, as far as the Hurst coefficient of the simulated series is concerned, it was once again estimated with the LSV method. A small discrepancy that concern site D, which is an SRD process (i.e.,  $H = 0.5$ ) is observed. This may be attributed to the associated estimation method and the high lag-1 autocorrelation ( $\sim 0.8$ ) of site D.

**Table 5-2** | a) Synopsis of theoretical distribution models and their moments, as well as, of CAS parameters for each variable of the multivariate simulation study. b) The upper triangle (grey cells) contains the target lag-0 cross-correlation coefficients ( $\rho_0^{i,j}$ ) between sites A-D, while the lower triangle depicts the corresponding estimated equivalent correlation coefficients ( $\tilde{\rho}_0^{i,j}$ ).

| a) Distribution/ Parameters | Theoretical                         |                         |                          |                                     | b) Lag-0 cross-correlation | Lag-0 cross-correlation |        |        |        |
|-----------------------------|-------------------------------------|-------------------------|--------------------------|-------------------------------------|----------------------------|-------------------------|--------|--------|--------|
|                             | Site A                              | Site B                  | Site C                   | Site D                              |                            | Site A                  | Site B | Site C | Site D |
|                             | $\mathcal{B}\mathcal{r}\text{-XII}$ | $\mathcal{P}\text{III}$ | $\mathcal{L}\mathcal{N}$ | $\mathcal{W}\mathcal{E}\mathcal{J}$ | Site A                     | 1                       | -0.700 | 0.750  | 0.600  |
| $a$                         | 2.5 ( $a_1$ )                       | 3                       | 0.5                      | 1.5                                 | Site B                     | -0.940                  | 1      | -0.600 | -0.700 |
| $b$                         | 1                                   | 1                       | 2                        | 10                                  | Site C                     | 0.862                   | -0.749 | 1      | 0.650  |
| $c$                         | 1.5 ( $a_2$ )                       | 10                      | -                        | -                                   | Site D                     | 0.811                   | -0.923 | 0.707  | 1      |
| Statistic                   | Theoretical                         |                         |                          |                                     |                            |                         |        |        |        |
| Mean ( $\mu$ )              | 4.76                                | 13                      | 8.37                     | 9.02                                |                            |                         |        |        |        |
| Variance ( $\sigma^2$ )     | 11.42                               | 3                       | 19.91                    | 37.56                               |                            |                         |        |        |        |
| Skewness coeff. ( $C_s$ )   | 5.01                                | 1.15                    | 1.75                     | 1.07                                |                            |                         |        |        |        |
| Kurtosis coeff. ( $C_k$ )   | -                                   | 8                       | 8.89                     | 4.39                                |                            |                         |        |        |        |
| CAS parameter, $\beta$      | 1.25                                | 1.66                    | 0                        | 0                                   |                            |                         |        |        |        |
| CAS parameter, $\kappa$     | 11.32                               | 5                       | 0.5                      | 0.2                                 |                            |                         |        |        |        |
| Hurst coeff. ( $H$ )        | 0.6                                 | 0.7                     | 0.5                      | 0.5                                 |                            |                         |        |        |        |

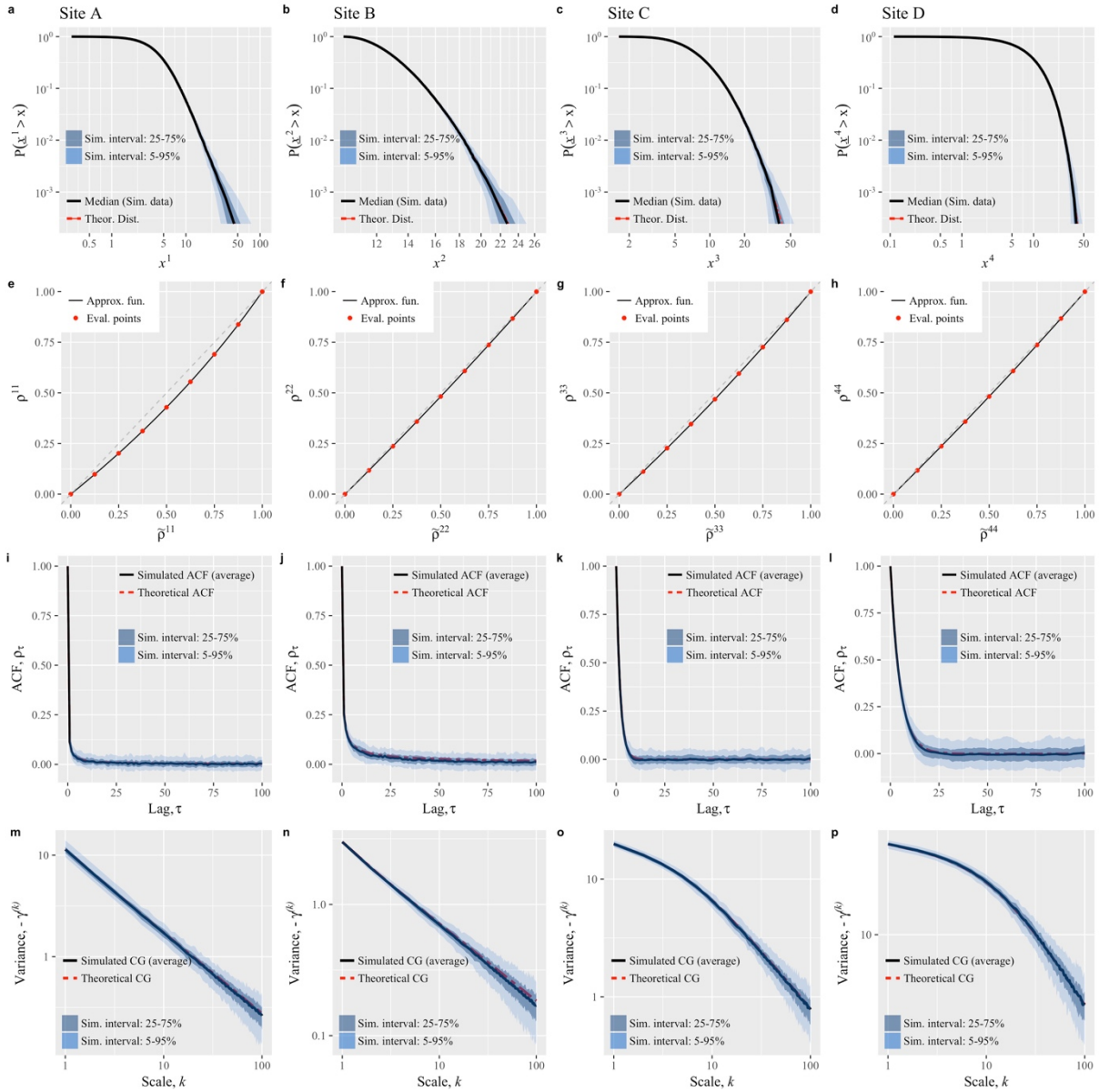
\*Distribution abbreviations:  $\mathcal{B}\mathcal{r}\text{-XII}$ : Burr type-XII ( $a_1$  = shape,  $a_2$  = shape,  $b$  = scale),  $\mathcal{P}\text{III}$ : Pearson type-III ( $a$  = shape,  $b$  = scale,  $c$  = location),  $\mathcal{L}\mathcal{N}$ : Log-Normal ( $a$  = shape,  $b$  = scale),  $\mathcal{W}\mathcal{E}\mathcal{J}$ : Weibull ( $a$  = shape,  $b$  = scale).



**Figure 5.5** | Comparison between theoretical (red dots, •) and simulated lag-1 autocorrelation and Hurst coefficient for sites A-D. Target (red dots, •) and simulated lag-0 cross-correlation coefficients for all pairs of sites A-D.

Furthermore, in **Figure 5.6a-d** we compared the empirical distribution of each realization of each site A-D with the corresponding theoretical distribution in terms of the survival function (SF), also known as complementary CDF or tail function. It is denoted by  $\bar{F}_x$  and expresses the probability of exceedance, i.e.,  $\bar{F}_x := P(x > x) = 1 - F_x$ . **Figure 5.6a-d** highlights the ability of the model to preserve the target distribution functions even in multivariate mode, since the median SF of all 100 realizations for the 4 sites is virtually identical to the associated theoretical model. Furthermore, in **Figure 5.6e-h** we depict the relationship between the equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients for each site A-D, while the preservation of the theoretical auto-dependence structure can be verified by the simulated ACFs (**Figure 5.6i-l**) and climacograms (**Figure 5.6m-p**) of the four variables that closely resemble the corresponding theoretical ones.





**Figure 5.6** | (a-d) Theoretical and simulated (SMARTA) distribution functions (using the Weibull’s plotting position) for sites A-D. (e-h) The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of sites A-D. (i-l) Theoretical and simulated ACFs for sites A-D. (m-p) Theoretical and simulated climacograms (CGs) for sites A-D. In all cases, the simulation intervals have been established using all 100 realizations.

Finally, to further explore the joint behavior of the model and the established dependence patterns we employ scatter plots. **Figure B.1** of Appendix B depicts the established dependence patterns among the variables for time lag 0 (**Figure B.1e**, i, j, m, n, o), as well as for each variable for time lag 1 (**Figure B.1a**, f, k, p). Finally, the relationship between equivalent,  $\tilde{\rho}^{i,j}$  and target  $\rho^{i,j}$  correlation coefficients is provided for every combination of sites A-D (**Figure B.1b**, c, d, g, h, l).

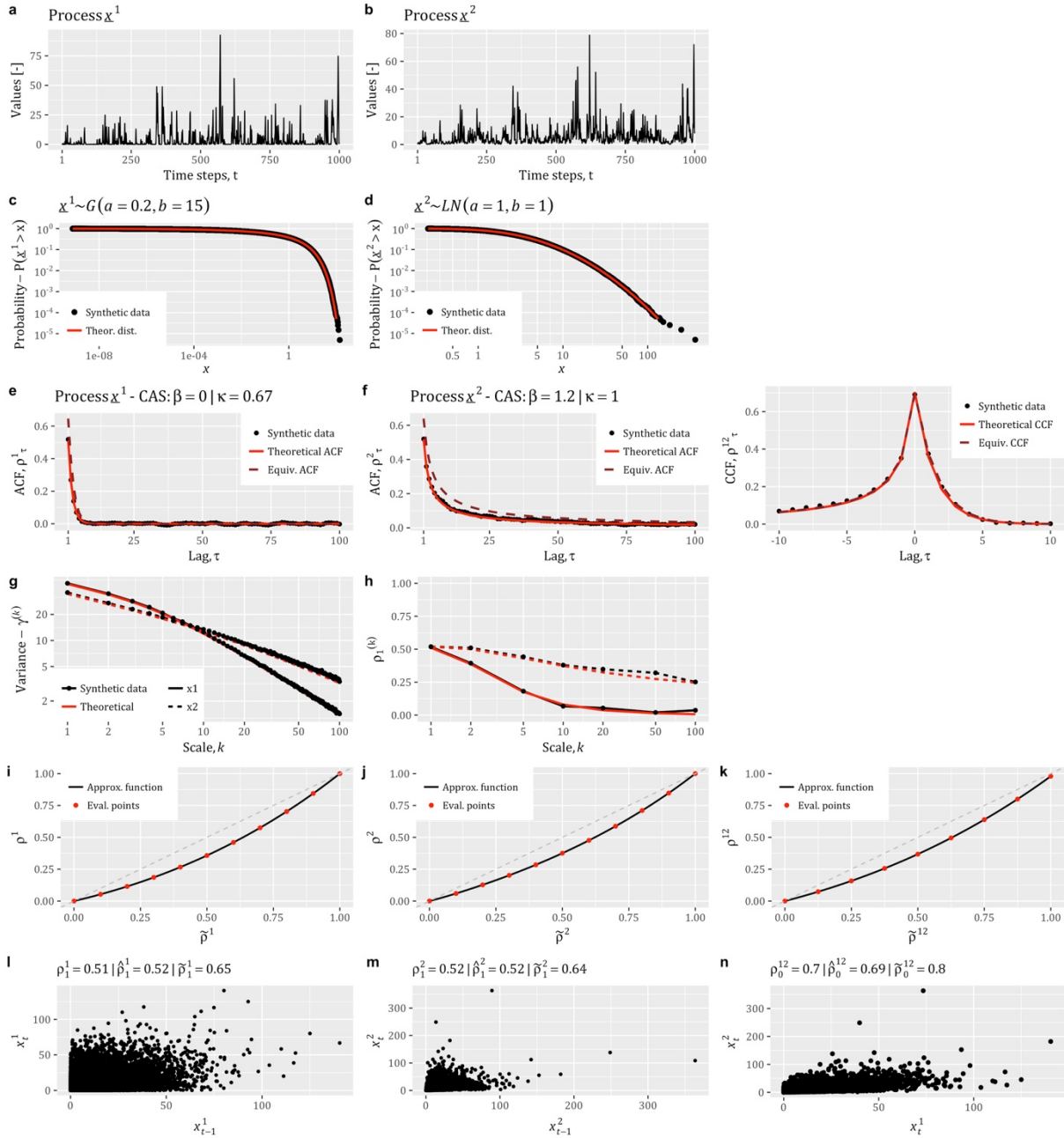
### 5.6.2 CMARTA model

To demonstrate the capabilities of CMARTA model, we setup three bivariate case studies (hereafter termed case study A, B and C) that regard the simulation of two contemporaneously cross-correlated processes,  $\underline{x}_t^1$  and  $\underline{x}_t^2$ , with continuous or discrete marginal distributions.

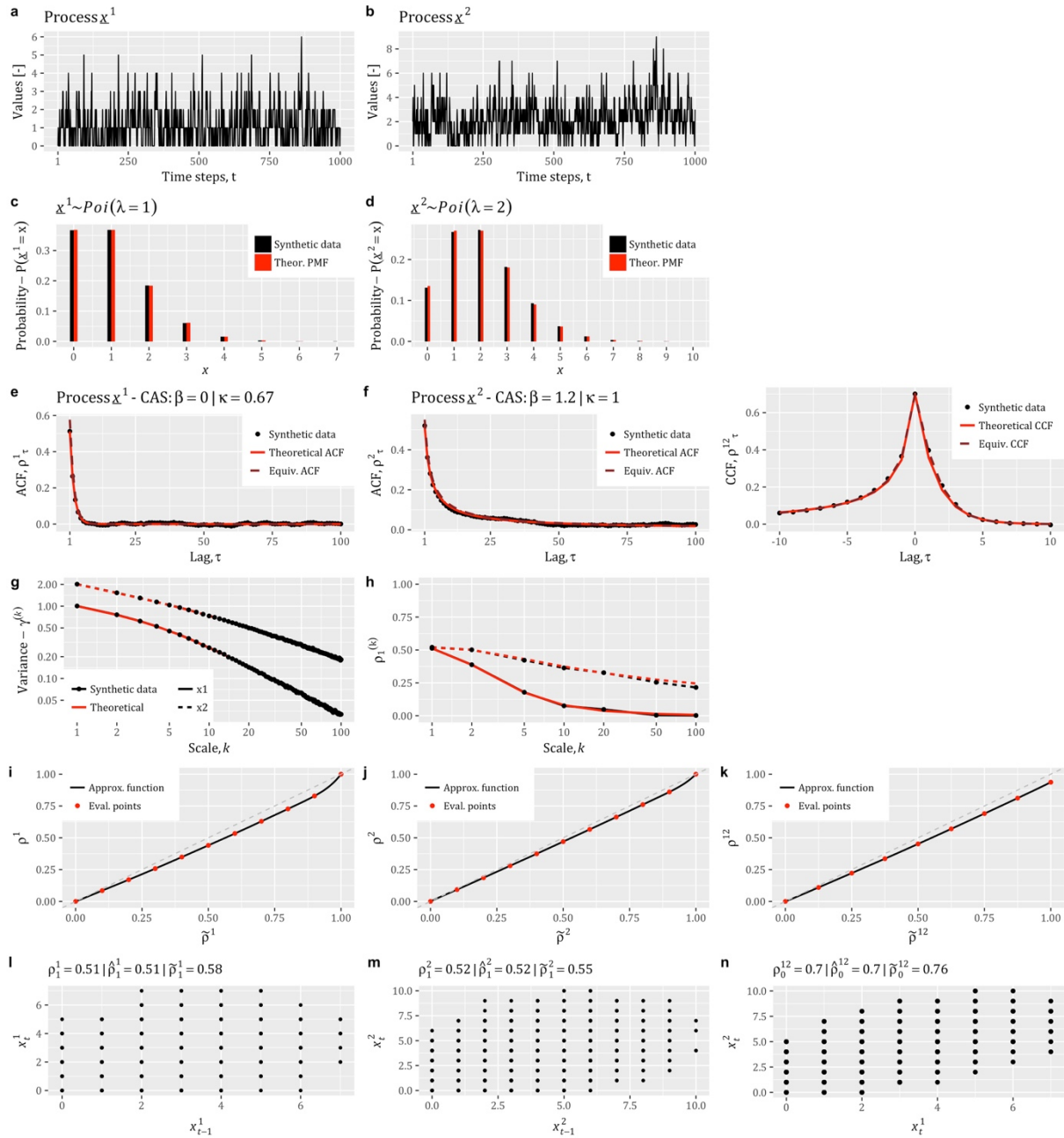
For convenience, the three cases share some common assumptions which are: 1) the order of the auxiliary Gaussian CMAR model (set to 100), 2) the length of the synthetic time series to simulate (set 100 000), 3) the target lag-0 cross-correlation coefficient, that is,  $\rho_0^{1,2} = 0.7$ , and 4) the target auto-dependence structure of each process, which is provided by CAS (i.e., Eq. (5.8)); that is,  $\underline{x}_t^1 \sim \rho_\tau^{CAS;1}(\beta = 0, \kappa = 0.67)$  and  $\underline{x}_t^2 \sim \rho_\tau^{CAS;2}(\beta = 1.2, \kappa = 1)$ . It is noted that the first process is modeled using an SRD auto-dependence structure, while the second using an LRD one. As briefly mentioned earlier (see also section 4.3.3), and since the autocorrelation structure of the processes is already specified, the use of high-order models does not introduces additional parameters, but solely controls the degree of resemblance of the target correlation structure. Particularly, by setting  $p = 100$ , the model will resemble the target CAS up to time lag 100, while for  $\tau > 100$  it will reduce according to its theoretical properties. Similarly, if we employed a higher-order model, e.g.,  $p = 1000$ , we would resemble the target CAS up to time lag  $\tau = 1000$ , without needing more parameters for the description of the autocorrelation structure. The case studies, differ in terms of the target marginal distribution of the individual processes. More specifically, in case A it is assumed that the marginals of  $\underline{x}_t^1$  and  $\underline{x}_t^2$  are continuous and are,  $\underline{x}_t^1 \sim \mathcal{G}(a = 0.2, b = 0.15)$  and  $\underline{x}_t^2 \sim \mathcal{LN}(a = 1, b = 1)$ . In case B the target distributions are regarded to be discrete, and given by the Poisson distribution. Particularly, we assume that,  $\underline{x}_t^1 \sim \text{Poi}(\lambda = 1)$  and  $\underline{x}_t^2 \sim \text{Poi}(\lambda = 2)$ . Finally, in case C the target distribution of each process was assumed to be the discrete-type Bernoulli distribution (*Bern*). Specifically,  $\underline{x}_t^1 \sim \text{Bern}(p = 0.8)$  and  $\underline{x}_t^2 \sim \text{Bern}(p = 0.75)$ .

It is noted that from a hydrometeorological processes simulation perspective, case study A can be considered as the most common simulation scenario, since it involves the simulation of processes with non-Gaussian (and skewed) continuous marginal distributions. On the other hand, cases B and C, can naturally arise when aiming to model counting (e.g., number of flood or drought events in a given year) or occurrence (i.e., binary; e.g., sequences of wet and dry transitions) processes respectively. As noted by [Serinaldi and Lombardo \[2017\]](#), within the context of a univariate binary process generator, any sequence of observations can be dichotomized to a binary one (i.e., occurrence of an event or not) by imposing appropriate rules. In a similar vein, an observed time series can also provide information regarding the frequency of certain events (e.g., times exceeding, or not, a given threshold during a certain period), hence transformed to a counting process.

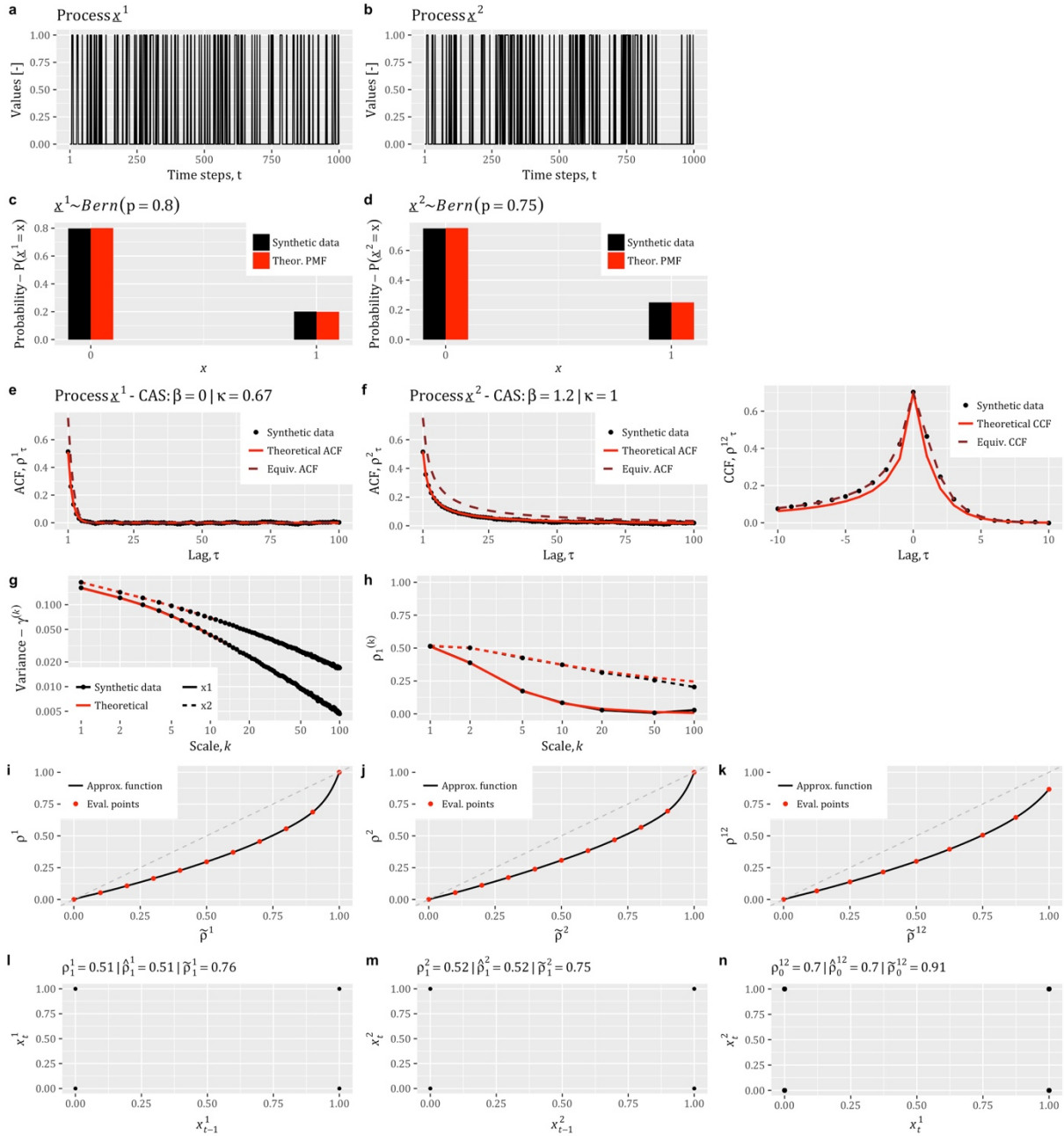
Regarding CMARTA evaluation for case studies A-C, its performance was assessed through a series of comparisons among a variety of simulated and theoretical characteristics. [Figure 5.7](#), [Figure 5.8](#), and [Figure 5.9](#) summarize the simulation results for cases A-C respectively and illustrate that the CMARTA is able to and accurately reproduce the probabilistic and stochastic structure of the target processes, regardless if its marginals are continuous or discrete. It is noted that similar results can be obtained using the SMARTA model.



**Figure 5.7** | Case A – Continuous marginal distributions. Simulated realization of process a)  $x_t^1$  and b)  $x_t^2$ . Comparison of simulated and theoretical distribution function for process c)  $x_t^1$  and d)  $x_t^2$ . Simulated, equivalent and theoretical autocorrelation function (ACF) for process e)  $x_t^1$  and f)  $x_t^2$ . g) Simulated and theoretical climacogram for process  $x_t^1$  and  $x_t^2$ . h) Simulated and theoretical lag-1 autocorrelation ( $\rho_1^{(k)}$ ) as a function of scale  $k$  for process  $x_t^1$  and  $x_t^2$ . The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of each process i)  $x_t^1$  j)  $x_t^2$ , as well as their k) interaction. Simulated dependence pattern for time lag 1 for process l)  $x_t^1$  and m)  $x_t^2$ . n) Simulated lag 0 dependence pattern among  $x_t^1$  and  $x_t^2$ .



**Figure 5.8** | Case B – Poisson marginal distributions. Simulated realization of process a)  $\underline{x}_t^1$  and b)  $\underline{x}_t^2$ . Comparison of simulated and theoretical distribution function for process c)  $\underline{x}_t^1$  and d)  $\underline{x}_t^2$ . Simulated, equivalent and theoretical autocorrelation function (ACF) for process e)  $\underline{x}_t^1$  and f)  $\underline{x}_t^2$ . g) Simulated and theoretical climacogram for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . h) Simulated and theoretical lag-1 autocorrelation ( $\rho_1^{(k)}$ ) as a function of scale  $k$  for process  $\underline{x}_t^1$  and  $\underline{x}_t^2$ . The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of each process i)  $\underline{x}_t^1$  j)  $\underline{x}_t^2$ , as well as their k) interaction. Simulated dependence pattern for time lag 1 for process l)  $\underline{x}_t^1$  and m)  $\underline{x}_t^2$ . n) Simulated lag 0 dependence pattern among  $\underline{x}_t^1$  and  $\underline{x}_t^2$ .



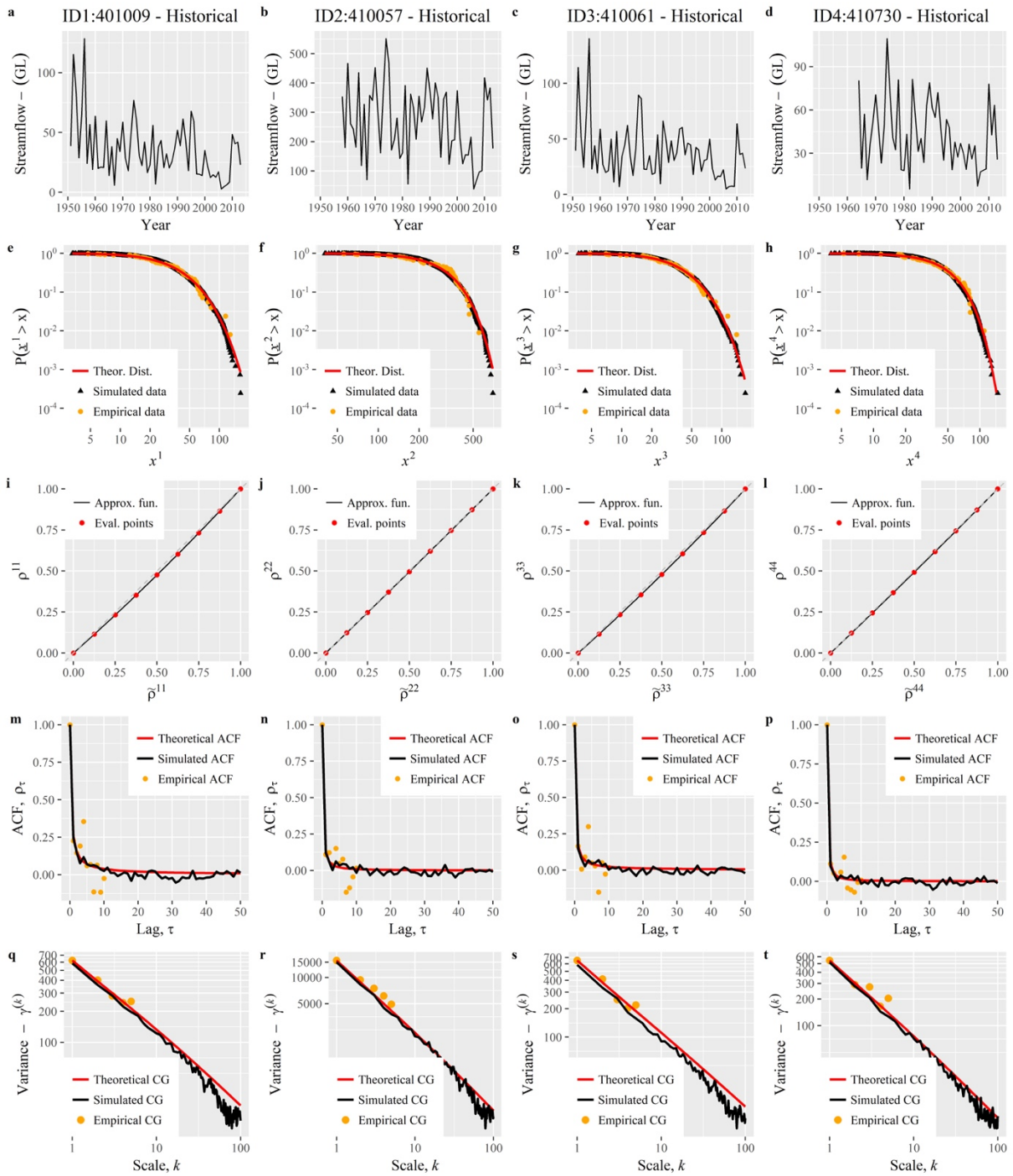
**Figure 5.9** | Case C – Bernoulli marginal distributions. Simulated realization of process a)  $x_t^1$  and b)  $x_t^2$ . Comparison of simulated and theoretical distribution function for process c)  $x_t^1$  and d)  $x_t^2$ . Simulated, equivalent and theoretical autocorrelation function (ACF) for process e)  $x_t^1$  and f)  $x_t^2$ . g) Simulated and theoretical climacogram for process  $x_t^1$  and  $x_t^2$ . h) Simulated and theoretical lag-1 autocorrelation ( $\rho_1^{(k)}$ ) as a function of scale  $k$  for process  $x_t^1$  and  $x_t^2$ . The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of each process i)  $x_t^1$  j)  $x_t^2$ , as well as their k) interaction. Simulated dependence pattern for time lag 1 for process l)  $x_t^1$  and m)  $x_t^2$ . n) Simulated lag 0 dependence pattern among  $x_t^1$  and  $x_t^2$ .

## 5.7 REAL-WORLD SIMULATION STUDIES

This section focuses on SMARTA model, since CMARTA is extensively used in the next Chapter within the context of a disaggregation-based simulation scheme. Of course, similar results could be obtained using either of the models.

### 5.7.1 Simulation of multivariate annual streamflow processes

The first real-world simulation study concerns the application of the proposed model for the stochastic simulation of annual streamflow at 4 stations in New South Wales region, Australia [Australian Government Bureau of Meteorology, 2015]. Particularly, we employed historical data (depicted in **Figure 5.10a-d**) from the following stations: Maragle Creek at Maragle (ID1: 401009), Goobarragandra River at Lacmalac (ID2: 410057), Adelong Creek at Batlow Road (ID3: 410061), Cotter River at Gingera (ID4: 410730). Hereinafter, we refer to them using their station ID, as well as model them in that order, as 4-dimensional stationary process  $\underline{x}_t = [x_t^1, x_t^2, x_t^3, x_t^4]^T$ ; (i.e.,  $i = 3$  refers to station Adelong Creek at Batlow Road with ID3: 410061). The distribution of historical data does not exhibit the typical bell shape that is often encountered in annual data, hence we use the Gamma and Weibull distributions to model them. Specifically, using the maximum likelihood estimation method we identified the following distributions,  $x_t^1 \sim \mathcal{G}(a = 2.13, b = 16.95)$ ,  $x_t^2 \sim \mathcal{WEI}(a = 2.30, b = 302.11)$ ,  $x_t^3 \sim \mathcal{WEI}(a = 2.40, b = 15.75)$  and  $x_t^4 \sim \mathcal{G}(a = 1.95, b = 48.48)$ . Furthermore, they are characterized by moderate-to-high temporal dependence and high lag-0 cross-correlation coefficients, that range from 0.83 ( $\rho_0^{1,4}$ ) to 0.93 ( $\rho_0^{2,3}$ ). Following Koutsoyiannis [2000], the parameters of CAS (i.e., Eq. (5.8) - given in vector format),  $\boldsymbol{\beta} = [0.99, 0.75, 1.13, 0.72]$  and  $\boldsymbol{\kappa} = [2.57, 4.41, 6.01, 5.07]$  were identified for each process by minimizing the mean square error (MSE) among the sample and theoretical autocorrelation coefficients. In this case study, we simulated one realization of 1 000 years using the SMARTA model (with  $q = 2^9$ ). **Figure 5.10e-h** provides, for each station, a visual comparison among the empirical, theoretical and simulated distribution. Furthermore, **Figure 5.10i-l** depicts, for each process, the relationship between the equivalent and target autocorrelation coefficients. The ability of the model to establish the target auto-dependence structures is verified by comparing, the theoretical and simulated ACF (**Figure 5.10m-p**) and corresponding climacogram (**Figure 5.10q-t**) of each process. Similarly to the previous simulation study, the model reproduced the target lag-0 cross-coefficients with high accuracy (**Figure B.2e, i, j, m, n, o**).



**Figure 5.10** | Synopsis of annual streamflow simulation study at 4 stations in New South Wales region. (a-d) Historical time series. (e-h) Empirical, simulated and theoretical distribution functions (using the Weibull's plotting position) for stations ID1-4 (i-l) The established relationships between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients given the marginal distribution of stations ID1-4. (m-p) Empirical, simulated and theoretical ACFs for stations ID1-4. (q-t) Empirical, simulated and theoretical climacograms (CGs) for stations ID1-4.

### 5.7.2 Simulation of univariate daily rainfall process

In the final case study, we employ the model for the stochastic simulation of a univariate daily rainfall process characterized by intermittency. The available data concern an observation period spanning from 1/1/1964 to 31/12/2006 (43 years) from Pavlos rain gauge located at Boeticos Kephisos river basin, Greece (**Figure 5.11a**). See also, [Efstratiadis et al. \[2014a\]](#) for further details regarding the dataset. In general, apart from *ad-hoc* techniques to handle intermittency (e.g., truncation to zero of values below a threshold), typical stochastic simulation schemes [e.g., [Serinaldi, 2009a](#); [Serinaldi and Kilsby, 2014](#); [Papalexiou, 2018](#)] rely on the use of mixed-distributions or employ two-part models, which, in a nutshell, describe precipitation processes as the product of two different processes, particularly, that of occurrence (rain or no-rain) and that of intensity [e.g., [Wilks, 1998](#); [Wilks and Wilby, 1999](#); [Brissette et al., 2007](#); [Thompson et al., 2007](#); [Khalili et al., 2009](#); [Mhanna and Bauwens, 2012](#); [Breinl et al., 2013](#); [Ailliot et al., 2015](#); [Lee, 2016, 2017](#); [Lombardo et al., 2017](#)]. Herein, we employ the former approach, that is, mixed-distributions, as it seems a convenient option [[Papalexiou, 2018](#)] given the characteristics of SMARTA and particularly its flexibility regarding the selection of the marginal distribution. An alternative option, naturally compatible with the proposed method (and Nataf-based schemes in general), would be the use of distribution functions that by construction, exhibit an atom of probability mass at zero. A characteristic example, which in the past has been used for this purpose [[Dunn, 2004](#); [Hasan and Dunn, 2011](#)], is the Tweedie distribution [[Tweedie, 1984](#); [Jorgensen, 1987](#)]. Nevertheless, in this simulation study, in order to simultaneously account for the effect of seasonality and the stationarity assumption of the model we treat each month as separate stochastic process, by varying the distribution function and autocorrelation structure on a monthly basis. Specifically, regarding the marginal distribution, we employ a discrete-continuous (i.e., mixed or zero-inflated) model (see section 4.4) whose CDF is given by,

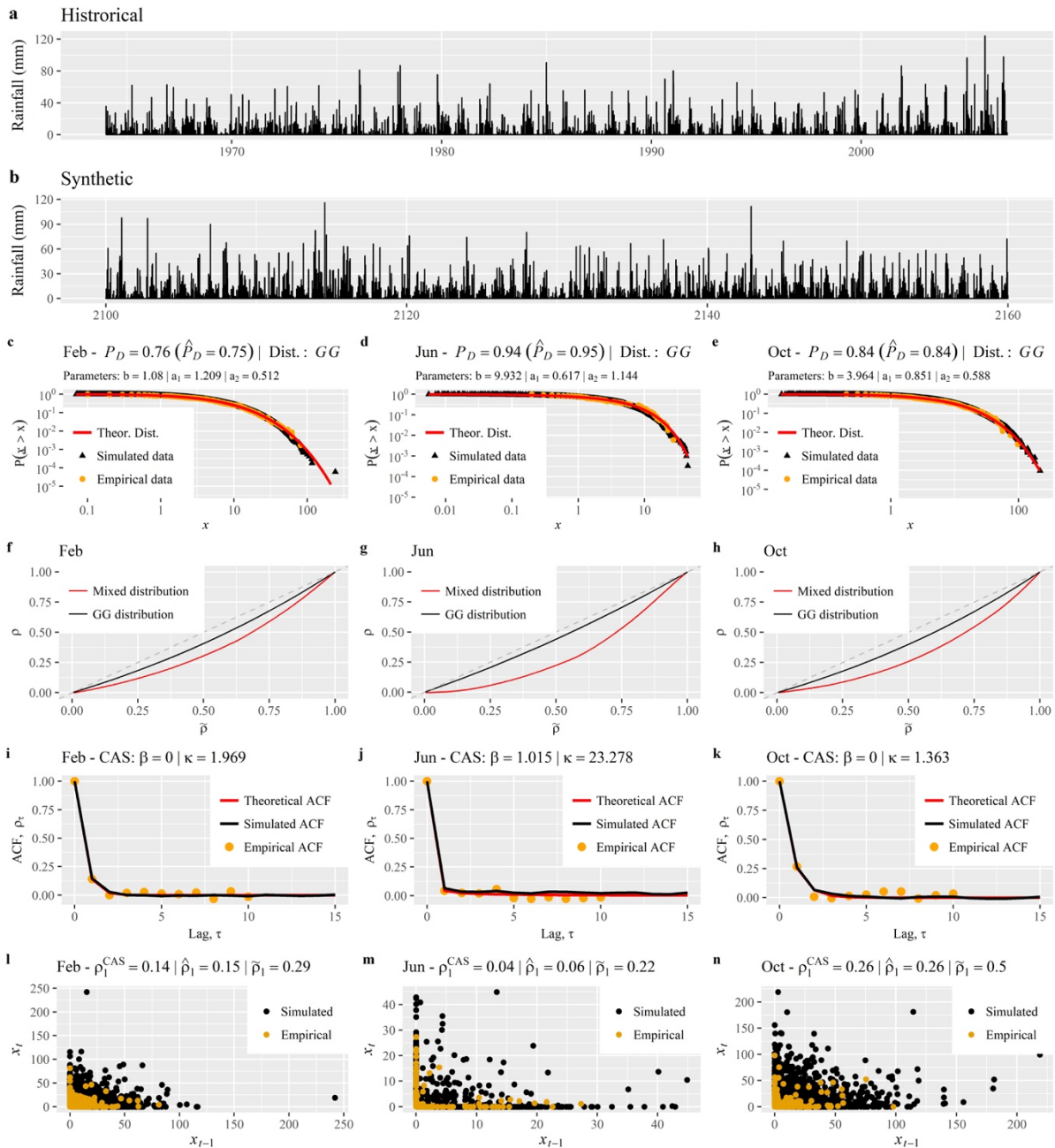
$$F_{\underline{x}}(x) = \begin{cases} p_D, & x \leq 0 \\ p_D + (1 - p_D)G_{\underline{x}}(x), & x > 0 \end{cases} \quad (5.43)$$

where,  $p_D$  denotes the probability of a dry interval (abbreviated as probability dry), i.e.,  $p_D := P(\underline{x} \leq x_D)$  and  $G_{\underline{x}}$  stands for the distribution of amounts greater than the threshold  $x_D$ , i.e.,  $G_{\underline{x}} := F_{\underline{x}|\underline{x}>x_D} = P(\underline{x} \leq x | \underline{x} > x_D)$ . We remind the reader that the solely requirement of the algorithm of section 4.5.1, that is used to establish the relationship between equivalent ( $\tilde{\rho}$ ) and target ( $\rho$ ) correlation coefficients, is the ICDF (see Eq. (4.46) in section 4.4). Nevertheless, after the specification of the threshold  $x_D$ , the empirical probability dry,  $p_D$ , can be directly obtained from the available data by counting the number of dry occurrences and dividing it with the total number of observed data. Regarding,  $G_{\underline{x}}$ , it is obtained by selecting and fitting a theoretical distribution to the amount data above  $x_D$ . In this demonstration, we set  $x_D := 0$ , and for the description of the positive daily precipitation amounts of all months, we employ the generalized gamma ( $\mathcal{GG}$ ) distribution [[Stacy, 1962](#)], which has been proved particularly capable for the task at hand [[Papalexiou and Koutsoyiannis, 2016](#); [Chen et al., 2017](#); [Papalexiou, 2018](#)]. Of course, depending on the case, the  $\mathcal{GG}$  could be replaced with other distribution functions. Back in our case, the parameters of the  $\mathcal{GG}$  distribution were identified using a fitting approach based on L-moments [[Hosking, 1990](#)]; specifically the one proposed by [Papalexiou and Koutsoyiannis \[2016\]](#). The PDF of  $\mathcal{GG}$  distribution is given by,



$$f_{\mathcal{GG}}(x; a_1, a_2, b) = \frac{a_2}{b\Gamma(a_1/a_2)} \left(\frac{x}{b}\right)^{a_1-1} \exp\left(-\left(\frac{x}{b}\right)^{a_2}\right), \quad x > 0 \quad (5.44)$$

where  $\Gamma(\cdot)$  denotes the gamma function, while,  $a_1 > 0, a_2 > 0$  are parameters that control the shape of the distribution and  $b > 0$  is a scale parameter. The interested reader is referred to the above works for further details regarding the  $\mathcal{GG}$  distribution and the associated fitting method. For instance, for the marginal characteristics of October's daily rainfall, we estimated,  $p_D = 0.84$ , while the parameters of  $\mathcal{GG}$  were found  $b = 3.96, a_1 = 0.851$  and  $a_2 = 0.588$ . Furthermore, regarding the description of the auto-dependence structure of the process we employed CAS and estimated its parameters on a monthly basis (e.g., for October it we identified,  $\beta = 0$  and  $\kappa = 1.36$ ) by minimizing the MSE among the sample and theoretical autocorrelation coefficients. Finally, we generated 1 000 years (i.e., 365 000 days) of synthetic data (**Figure 5.11b** depicts a random window of 60 years) and performed a similar analysis with the previous cases studies; which is summarized in **Figure 5.11**, where we depict the results of three characteristic months, i.e., February, June and October (the results are similar for the other months – see Appendix B, **Figure B.3 - Figure B.6**). Particularly, panels (c)-(e) illustrate the capability of the model to reproduce the target distributions (in terms of the SF) of positive precipitation amounts ( $p_D$  is explicitly preserved since it is embedded in the employed mixed-distribution model), while, panels (f)-(h) depicts the relationship of equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients for both  $\mathcal{GG}$  and mixed-distribution models. It is observed that the non-linearity of this relationship increases from  $\mathcal{GG}$  to mixed distribution due to the fact that the latter is zero-inflated. Furthermore, panels (i)-(k) depict the accurate resemblance of the target autocorrelation structure (i.e., CAS), while, panels (l)-(n) provide a comparison of empirical and simulated scatter for time lag 1, which seems to be in agreement with the historical pattern. Finally, preliminary analysis (not shown herein) indicated that the model has the potential to approximate some of the empirical statistics (in terms of L-moments) across coarser time scales, even though they are not explicitly modelled by it. This observation should not be interpreted as a general conclusion, rather as a direction for further investigation. We remark that the literature offers several well-established techniques with proven results, specifically designed for this purpose, i.e., to address scaling and intermittency, such as disaggregation [e.g., *Kossieris et al., 2016; Lombardo et al., 2017*] and multi-fractal methods, based on cascade models [*Tessier et al., 1996; e.g., Deidda et al., 1999; Kantelhardt et al., 2006*]. These methods, by design, aim to simultaneously resemble the process at multiple aggregation levels, employing scaling relationships for high order moments (often greater than second). In our view, an interesting topic of future research would be a comparison among these simulation techniques with Nataf-based methods for the reproduction of the multi-scale behavior that characterizes hydrometeorological processes. Similar works, yet involving alternative simulation schemes, are those of *Lombardo et al. [2012]* and *Pui et al. [2012]*.



**Figure 5.11** | Synopsis of daily rainfall simulation at Pavlos' station. a) Historical time series. b) Synthetic time series; randomly selected window of 60 years. Empirical, simulated and theoretical distribution function of positive precipitation amounts for c) February, d) June and e) October (using the Weibull's plotting position); the title of each plot provides the parameters of the  $GG$  distribution, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry. The established relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients for the mixed and  $GG$  distribution for f) February, g) June and h) October. Empirical, simulated and theoretical ACF for i) February, j) June and k) October; the title of each plot depicts the parameters of CAS. Empirical and simulated dependence pattern for time lag 1 for l) February, m) June and n) October; the title of each plot depicts the lag-1, target ( $\rho_1^{CAS}$ ), simulated ( $\hat{\rho}_1$ ), and equivalent ( $\tilde{\rho}_1$ ) autocorrelation coefficients.

## 5.8 SUMMARY

This Chapter introduces two novel versatile stochastic models, termed SMARTA and CMARTA, with solid theoretical background and proven capability of addressing important hydrometeorological simulation problems. A prominent characteristic of the models is their ability to simulate univariate and multivariate stationary processes with any autocorrelation structure and marginal distribution, provided that the former is feasible and the latter have finite variance. Their central idea relies on the use of an appropriately parameterized (expressed through *equivalent* correlation coefficients) auxiliary Gaussian process which after its mapping to the actual domain results in a process with the desired stochastic structure and marginal distribution.

Briefly, the proposed approach is built upon three major elements, that is, a) auxiliary linear stochastic models (i.e., the SMA scheme of [Koutsoyiannis \[2000\]](#) and the CMAR model) in the Gaussian domain, b) theoretical autocorrelation structures (e.g., CAS), that allows the parsimonious description of both SRD and LRD processes, and c) the rationale of NDM [[Nataf, 1962](#)], and the associated mapping procedure, that provide the theoretical basis of the method and in turn allows the identification of the *equivalent* correlation coefficients; hence determine the parameters of the auxiliary model.

Overall, the proposed methodology is parameter parsimonious and exhibits a series of virtues, as demonstrated through several hypothetical and two real-world simulation studies. Among them:

- a) The unambiguous advantage of explicitly simulating any-range dependent (SRD or LRD) stationary processes with arbitrary distributions (that may be continuous, discrete or mixed-type), using a single simulation scheme.
- b) Its ability to simulate univariate and multivariate processes that exhibit contemporaneous cross-correlations. The generation of time series at multiple locations, or of individual correlated processes, is often the case in hydrological studies, making SMARTA and CMARTA particularly useful methods for such tasks.
- c) The possible incorporation of novel advances in statistical science in stochastic simulation; such as new distributions and robust fitting methods (e.g., L-moments). In addition, regarding distributions of hydrometeorological processes, both models (as any Nataf-based model; see section 4.3.1, as well as Chapter 6 and 7) can take advantage of years of research in statistical analysis of hydrometeorological variables, since it can incorporate any distribution function whose variance exists.
- d) The ability of the model to explicitly avoid the generation of negative values, which is a shortcoming of many linear stochastic models. This is due to the direct use of the distribution function(s) within the generation mechanism of the model. If the used distribution is defined in the positive real line, then all the generated values will be within those bounds (i.e., positive).

Typical, but not limited, applications of the proposed models entail the simulation of stationary processes at time scales not affected by cyclostationary correlation structures (e.g., monthly scale). For instance, given the wide range of admissible correlation structures and distributions, the models could be applied for the generation of synthetic time series of various hydrometeorological processes, such as, precipitation, streamflow and temperature, at annual and fine time scales (e.g., daily), which are characterized by stationarity.

Next, in Chapter 6, the particularly interesting case of cyclostationarity is discussed, and a novel non-Gaussian Nataf-based stochastic model is being proposed [Tsoukalas et al., 2017a, 2018e]. Ongoing research aims in an enhanced stochastic simulation scheme that will combine (using disaggregation techniques) both stationary (e.g., SMARTA and CMARTA) and cyclostationary Nataf-based models (next chapter); thus providing an even more flexible and versatile simulation method for synthetic time series generation (see Chapter 7).

## SIMULATION OF CYCLOSTATIONARY STOCHASTIC PROCESSES WITH ARBITRARY MARGINAL DISTRIBUTIONS \*

---

### PREAMBLE

This Chapter presents a novel model, termed Stochastic Periodic AutoRegressive To Anything (SPARTA), for the simulation of cyclostationary processes (univariate and multivariate) with arbitrary marginal distributions. SPARTA offers an alternative and novel approach which allows the explicit representation of each process and season of interest with any distribution model, while simultaneously establishes dependence patterns that cannot be fully captured by the typical linear stochastic schemes. Cornerstone of the proposed approach is the Nataf joint-distribution model, which is related with the Gaussian copula, combined with Gaussian periodic autoregressive (PAR) processes. Theoretical and practical benefits of the proposed method, contrasted to outcomes from widely-used stochastic models, are demonstrated by means of real-world as well as hypothetical monthly simulation examples involving both univariate and multivariate time series.

The organization of this Chapter is as follows: Section 6.1 introduces the problem of generating cyclostationary process with emphasis on the reproduction of marginal distributions. The rationale and computational procedure of SPARTA are described in the next three sections, where section 6.2 summarizes the overall methodology, section 6.3 describes auxiliary Gaussian PAR model, while section 6.4 described the generation procedure of SPARTA in a step-by-step manner. In section 6.5 we evaluate the proposed method by means of three case studies, involving real-world and hypothetical simulations. The case studies also involve a comparison with the widely used implicit Periodic AutoRegressive (PAR) scheme. Finally, the key conclusions and perspectives of this research are outlined in section 6.6.

---

\* Based on:

Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2018e), Stochastic Periodic Autoregressive to Anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions, *Water Resour. Res.*, 54(1), 161–185, doi:10.1002/2017WR021394.

Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2017a), Stochastic simulation of periodic processes with arbitrary marginal distributions, in 15th International Conference on Environmental Science and Technology, CEST 2017., Rhodes, Greece.

## 6.1 INTRODUCTION

The generation of synthetic time series following specific, typically skewed, distribution functions becomes even more challenging when aiming to simulate hydrometeorological processes at time scales finer than annual, that are dominated by periodicity. Characteristic examples are the monthly processes of precipitation and river flow discharge, which exhibit strong seasonal variations in both their marginal and joint properties. In that case, the stochastic model should account for all facets of cyclostationarity, involving not only, the stochastic structure of the underlying processes but also their distribution, which may be seasonally-varying and non-Gaussian (see section 2.2). As detailed in section 2.3, typical stochastic models in hydrology (e.g., implicit linear stochastic schemes, point-process models, resampling schemes, and disaggregation models) traditionally aim at reproducing the empirically-derived statistical characteristics of the observed data rather than any specific distribution model that attempts to describe the usually non-Gaussian behavior of the associated processes. Notable exceptions are copula-based methods, which however are subject to high-computational requirements and complex generation mechanisms (see section 2.3). Further to this, some of the available schemes are not designed, hence capable, for the simulation of such processes (e.g., point-process models and two-part models).

In order to address the aforementioned shortcomings, this Chapter presents an *explicit* method, called Stochastic Periodic AutoRegressive To Anything (SPARTA) model, which offers a generalized procedure with solid theoretical background for the generation of cyclostationary processes from *a priori* defined distribution functions that are seasonally-varying. The proposed method builds upon the so-called Nataf joint-distribution model [NDM; Nataf, 1962], which is generic mapping procedure, and extends the AutoRegressive To Anything (ARTA) model, introduced by Cario and Nelson [1996] that represents stationary processes with arbitrary marginal distributions and autocorrelation structure. Initially, ARTA was formulated as univariate and later extended for multivariate simulations [Biller and Nelson, 2003]. Both versions involve the simulation of stationary processes, but they have not been extended to account for cyclostationarity, which is a *sine qua non* for hydrological processes. Beyond this, it noted that SPARTA is able to establish dependence patterns that cannot be fully captured by the typical linear stochastic schemes (see Chapter 3, as well as section 6.5.1 and 6.5.3). Briefly, the proposed approach involves the simulation of an auxiliary process from the Periodic AutoRegressive (PAR) family, in the *normal* domain (i.e., Gaussian), which allows accounting for cyclostationarity, and then its mapping to the *real* domain, via the desired inverse cumulative distribution functions (ICDFs).

## 6.2 SPARTA AT A GLANCE

SPARTA aims at simulating periodic processes from any given marginal distribution and a given stochastic structure, typically (but not exclusively) expressed in terms of first order autocorrelations and lag zero cross-correlations. Its fundamental advantage is the explicit preservation of the theoretical marginal distributions of the processes, in contrast to existing linear stochastic approaches that preserve the marginal statistics (not the distributions themselves) up to a specific order, typically the third one (i.e., mean, standard deviation, skewness).

More specifically: Let  $\underline{x}_{s,t} = [x_{s,t}^1, \dots, x_{s,t}^m]^T$  be a  $m$ -dimensional vector of cyclostationary stochastic processes to simulate, where  $s = 1, \dots, S, 1, \dots, S, \dots$  denotes the season (e.g., month),  $S$  the total number of seasons, and  $t = 1, \dots, T$  denotes the aggregated time scale (e.g.,

year). This process can also be written as  $\underline{x}_{s,n}$ , where  $n \in \mathbb{Z}^>$ , denotes the time index. In this form, the season  $s$  is recovered by,  $s = n \bmod(S)$ , while when  $n \bmod(S) = 0$ ,  $s = S$ . The period  $t$  is obtained by,  $t = 1 + (n - s)/S$ . For convenience, the first formulation will be employed in the following paragraphs. Each element of  $\underline{x}_{s,t}$  is symbolized  $x_{s,t}^i$ , where  $i = 1, \dots, m$  denotes an individual random process, and  $x_{s,t}^i$  denotes its realization. Herein, index  $i$  will be also referred to as *location* or *site*, without necessarily implying spatial reference. Let also  $\rho_{s,s-\tau}^{i,j} := \text{Corr}[x_{s,t}^i, x_{s-\tau,t}^j]$  be the Pearson coefficient of correlation among processes  $i$  and  $j$ , for season  $s$  and time lag  $\tau$ . For instance, when  $j = i$  and  $\tau \neq 0$ , the quantity  $\rho$  represents the season-to-season correlation of the process for lag  $\tau$ , while for  $j \neq i$  and  $\tau = 0$ ,  $\rho$  represents the cross-correlation between  $i$  and  $j$ , for zero time lag. Furthermore, when the superscripts or subscripts of  $\rho$  are identical (i.e., when  $j = i$  or  $\tau = 0$ ) we may omit repeating them for convenience (e.g.,  $\rho_{s,s-\tau}^{i,i}$  may be written as  $\rho_{s,s-\tau}^i$  and  $\rho_{s,s}^{i,j}$  as  $\rho_s^{i,j}$ ).

For each process at each season  $s$  and each location  $i$ , we assign a specific distribution function,  $F_{x_s^i} := P(x_s^i \leq x)$ , and also assign target coefficients of correlation,  $\rho_{s,s-\tau}^{i,j}$ , to preserve within the proposed generation scheme.

The key idea of SPARTA employs the concept of NDM and the associated methods, and establishes the process  $\underline{x}_{s,t}$  through an auxiliary Gaussian cyclostationary process  $\underline{z}_{s,t} = [z_{s,t}^1, \dots, z_{s,t}^m]^T$  with  $z_{s,t}^i \sim \mathcal{N}(0,1)$  and *equivalent* correlation coefficients  $\tilde{\rho}_{s,s-\tau}^{i,j} := \text{Corr}[z_{s,t}^i, z_{s-\tau,t}^j]$ . The  $\underline{z}_{s,t}$  process is generated from a standard Normal (i.e., Gaussian) Periodic AutoRegressive process (symbolized PAR-N), with such parameters that their mapping via the corresponding inverse marginal distributions (ICDFs) results into processes with the target marginal distributions and the target correlation structure, i.e.,

$$\underline{x}_{s,t}^i = F_{x_s^i}^{-1}(\Phi(z_{s,t}^i)) \quad (6.1)$$

where  $\Phi(\cdot)$  is the CDF of the standard Gaussian distribution and  $F_{x_s^i}^{-1}(\cdot)$  denotes the ICDFs of the target distributions of process  $i$  at season  $s$ . This mapping function, ensures the representation of *any* distribution across seasons and processes.

However, as thoroughly discussed in section 4.3, the main challenge, encountered in all Nataf-based models (including SPARTA), is the identification of proper parameters for the auxiliary process in the *normal* domain that reproduce the desired stochastic structure, after applying the mapping procedure. This arises from the fact that the Pearson correlation coefficient, which is used to describe all kinds of dependencies within linear stochastic models (including PAR), cannot be preserved when applying a non-linear monotonic transformation, such as the ICDF. In particular, Eq. (6.1) results into underestimation of target correlations,  $\rho_{s,s-\tau}^{i,j}$ , when they are directly applied to the auxiliary processes. The origin of this shortcoming is the fact that the Pearson's correlation coefficient (in contrast to rank correlation statistics) is invariant only under linear transformations [Embrechts et al., 1999 p. 7], while for any other transformation, the correlation coefficients should be properly adjusted. Section 4.6 mentions some early works in stochastic hydrology that were aware of this issue and attempted to provide analytical or empirical solutions to this problem, for specific distributions (e.g., Log-Normal).

Therefore it is essential to identify the equivalent ( $\tilde{\rho}_{s,s-\tau}^{i,j}$ ) correlation coefficients which should be used within the parameters estimation procedure of the auxiliary PAR model, that result

into the desired target  $(\rho_{s,s-\tau}^{i,j})$  correlations. By building upon the theoretical background of NDM, and assuming that the marginal distributions  $F_{\underline{x}_s^i}$  for all  $i = 1, \dots, m$  and  $s = 1, \dots, S$ , have been specified, the relationship between the equivalent  $(\tilde{\rho}_{s,s-\tau}^{i,j})$  and target  $(\rho_{s,s-\tau}^{i,j})$  correlations is given by (see section 4.3.1; [Tsoukalas et al., 2017a, 2018e]),

$$\rho_{s,s-\tau}^{i,j} = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_{\underline{x}_s^i}^{-1}(\Phi(z_s^i)) F_{\underline{x}_{s-\tau}^j}^{-1}(\Phi(z_{s-\tau}^j)) \varphi_2(z_s^i, z_{s-\tau}^j, \tilde{\rho}_{s,s-\tau}^{i,j}) dz_s^i dz_{s-\tau}^j - E[\underline{x}_s^i] E[\underline{x}_{s-\tau}^j]}{\sqrt{\text{Var}[\underline{x}_s^i] \text{Var}[\underline{x}_{s-\tau}^j]}} \quad (6.2)$$

where  $\varphi_2(z_s^i, z_{s-\tau}^j, \tilde{\rho}_{s,s-\tau}^{i,j})$  is the bivariate standard normal probability density function and  $E[\underline{x}_s^i]$ ,  $E[\underline{x}_{s-\tau}^j]$  and  $\text{Var}[\underline{x}_s^i]$ ,  $\text{Var}[\underline{x}_{s-\tau}^j]$  denote the mean and variance of  $\underline{x}_s^i$  and  $\underline{x}_{s-\tau}^j$  respectively which are known from the corresponding distributions  $F_{\underline{x}_s^i}$  and  $F_{\underline{x}_{s-\tau}^j}$  and have to be finite. For convenience, the above equation is abbreviated as,

$$\rho_{s,s-\tau}^{i,j} = \mathcal{F}(\tilde{\rho}_{s,s-\tau}^{i,j} | F_{\underline{x}_s^i}, F_{\underline{x}_{s-\tau}^j}) \quad (6.3)$$

The relationship of Eq. (6.3) can be established (in a pair-wise basis) through the hybrid method of section 4.5.1, and subsequently should be inverted, i.e.,  $\tilde{\rho}_{s,s-\tau}^{i,j} = \mathcal{F}^{-1}(\rho_{s,s-\tau}^{i,j} | F_{\underline{x}_s^i}, F_{\underline{x}_{s-\tau}^j})$ , in order to identify the equivalent coefficients,  $\tilde{\rho}_{s,s-\tau}^{i,j}$ , to be used within the PAR-N generation procedure.

### 6.3 THE AUXILIARY GAUSSIAN PAR MODEL

As mentioned above, the generation procedure of SPARTA requires the synthesis of an auxiliary process  $\underline{z}_{s,t}$ , which is then mapped to the actual one, i.e.,  $\underline{x}_{s,t}$ . This process has to be cyclostationary (since the target process is also cyclostationary) and normal. These premises are fulfilled by standard periodic autoregressive models with normally-distributed noise (PAR-N) of any order [e.g., Salas and Pegram, 1977; Salas et al., 1985; Salas, 1993].

Although any stochastic scheme from the PAR-N family may be applicable, we pay attention to the PAR(1) process, in order to keep things simple and parsimonious, thus providing an easy to follow narrative. In addition, it is argued that the assumption of a first-order model is well-justified for most of practical applications in hydrology [Efstratiadis et al., 2014a]. Nevertheless, higher-order models may be cumbersome, because the empirical estimation of joint statistics from historical samples is subject to major uncertainty, usually resulting to ill-posed conditions (e.g., due to inconsistent autocorrelation structures), which in turn leads to substantial defects within parameter estimation.

With respect to cross-correlations, the multivariate PAR(1) model, in its full formulation, preserves both the lag zero and lag one dependencies. However, as Koutsoyiannis and Manetas [1996] have shown, for reasons of parsimony it is sufficient using the contemporaneous PAR(1) [Salas, 1993 p. 19.31], which does not explicitly accounts for lag-one cross-correlations within parameter estimation. This is also advocated by an older study of Pegram and James [1972]. For instance, in a four-variable problem with 12 seasons, the full PAR(1) model requires the



specification of 264 parameters to describe the dependencies among the variables, while the contemporaneous one entails 120.

It is reminded that in order to employ the multivariate contemporaneous PAR(1)-N within SPARTA, it is essential to provide the equivalent lag-1 month-to-month correlations (i.e., autocorrelations),  $\tilde{\rho}_{s,s-1}^i$ , for each process  $i$  and season  $s$ , as well as the equivalent zero-lag cross-correlations,  $\tilde{\rho}_s^{i,j}$ , for each pair of processes  $i$  and  $j$  and season  $s$ . Hence in order to emphasize on the use of equivalent correlation coefficients within the parameter estimation procedure of the PAR model the tilde notation will be employed.

### 6.3.1 Multivariate contemporaneous PAR(1) model

Keeping the same notation for the auxiliary and actual processes, the multivariate PAR(1) reads (for convenience, the index of the period  $t$  is omitted):

$$\underline{\mathbf{z}}_s = \tilde{\mathbf{A}}_s \underline{\mathbf{z}}_{s-1} + \tilde{\mathbf{B}}_s \underline{\mathbf{w}}_s \quad (6.4)$$

where  $\underline{\mathbf{z}}_s = [\underline{z}_s^1, \dots, \underline{z}_s^m]^\top$  is a vector of  $m$  stochastic processes in season  $s$ ,  $\tilde{\mathbf{A}}_s, \tilde{\mathbf{B}}_s$  are  $m \times m$  parameter matrices that depend on season  $s$ , and  $\underline{\mathbf{w}}_s = [\underline{w}_s^1, \dots, \underline{w}_s^m]^\top$  is a vector of mutually independent random variables. By definition, the random process  $\underline{\mathbf{z}}_s$  is Gaussian, provided that  $\underline{\mathbf{w}}_s$  is generated from the standard normal distribution, i.e.,  $\underline{w}_s^i \sim \mathcal{N}(0, 1)$ .

For each season  $s$ , the parameter matrix  $\tilde{\mathbf{A}}_s$  is diagonal and contains the equivalent lag-1 month-to-month correlations,  $\tilde{\rho}_{s,s-1}^i$ , i.e.,

$$\tilde{\mathbf{A}}_s = \text{diag}(\tilde{\rho}_{s,s-1}^1, \dots, \tilde{\rho}_{s,s-1}^m) \quad (6.5)$$

On the other hand, parameter matrices  $\tilde{\mathbf{B}}_s$  are calculated by,  $\tilde{\mathbf{B}}_s \tilde{\mathbf{B}}_s^\top = \tilde{\mathbf{G}}_s$  where  $\tilde{\mathbf{G}}_s := \tilde{\mathbf{C}}_s - \tilde{\mathbf{A}}_s \tilde{\mathbf{C}}_{s-1} \tilde{\mathbf{A}}_s^\top$  and  $\tilde{\mathbf{C}}_s$  is a symmetric  $m \times m$  matrix that contains the equivalent lag-zero cross-correlations,  $\tilde{\rho}_s^{i,j}$ , i.e.,

$$\tilde{\mathbf{C}}_s = \begin{bmatrix} 1 & \dots & \tilde{\rho}_s^{1,m} \\ \vdots & \ddots & \vdots \\ \tilde{\rho}_s^{m,1} & \dots & 1 \end{bmatrix}$$

Furthermore, as discussed in *Koutsoyiannis [2001]*, the lagged correlation matrices  $\tilde{\mathbf{C}}_{sr} := \text{Corr}[\underline{\mathbf{z}}_s, \underline{\mathbf{z}}_r]$  of the PAR(1)-N model can be estimated, on the basis of the parameter matrices  $\tilde{\mathbf{A}}_s$  and the lag-zero cross-correlation matrices  $\tilde{\mathbf{C}}_s$ , for any time lag  $(s - r)$ , by,

$$\tilde{\mathbf{C}}_{sr} = \text{Corr}[\underline{\mathbf{z}}_s, \underline{\mathbf{z}}_r] = \tilde{\mathbf{A}}_s \tilde{\mathbf{A}}_{s-1} \dots \tilde{\mathbf{A}}_{r+1} \tilde{\mathbf{C}}_r, \quad s > r \quad (6.6)$$

As mentioned earlier (e.g., section 4.1.3), the estimation of the parameter matrix  $\tilde{\mathbf{B}}_s$  (which is often assumed to be lower triangular) requires the formulation of a decomposition problem (i.e., finding the square root of  $\tilde{\mathbf{G}}_s$ ), typically resolved using standard matrix decomposition techniques (e.g., Cholesky or singular value decomposition [e.g., *Johnson, 1987*]), when  $\tilde{\mathbf{G}}_s$  is positive definite, or otherwise, approximated via optimization techniques [e.g., *Koutsoyiannis, 1999; Higham, 2002*]. In particular, *Koutsoyiannis [1999]* has developed an optimization-based approach, paying attention on the preservation of skewness, which is a well-known trouble of multivariate stochastic models, asking for generating skewed white noise [e.g., *Todini, 1980*].

A great advantage of SPARTA approach is the assumption of normality within the auxiliary process, which substantially simplifies the optimization problem within decomposing non-positive definite matrices. More precisely, the empirical penalty term considered by [Koutsoyiannis \[1999\]](#), in order to prohibit the generation of highly-skewed white noise, which introduces significant complexity to the optimization procedure [[Efstratiadis et al., 2014a](#)], is neglected, thus resulting to a *reduced* objective function that only contains a distance term to minimize. Even in case of non-positive definite correlation matrices, where the desired stochastic characteristics are not explicitly preserved by the auxiliary model, the *reduced* optimization approach ensures a very good approximation, with minimal computational burden.

### 6.3.2 Univariate PAR(1) model

The univariate model can easily be derived from the above equations. Since  $m = 1$ ,  $\tilde{\mathbf{A}}_s = \tilde{\rho}_{s,s-1}^1$  and  $\tilde{\mathbf{C}}_s = 1$ , thus  $\tilde{\mathbf{B}}_s \tilde{\mathbf{B}}_s^T = 1 - \tilde{\rho}_{s,s-1}^1 \tilde{\rho}_{s,s-1}^1$ , which leads to  $\tilde{\mathbf{B}}_s = \sqrt{1 - \tilde{\rho}_{s,s-1}^1{}^2}$ . Hence, by substituting in Eq. (6.4) and removing the redundant indices we read:

$$\underline{z}_s = \tilde{\rho}_{s,s-1} \underline{z}_{s-1} + \sqrt{1 - \tilde{\rho}_{s,s-1}^2} \underline{w}_s \quad (6.7)$$

where  $\underline{w}_s$  are i.i.d. white noise with  $\mathcal{N}(0, 1)$ . We remark that since  $i = 1$  the superscript of  $\tilde{\rho}$  has been omitted for simplicity.

## 6.4 GENERATION PROCEDURE OF SPARTA MODEL

Summarizing, the implementation of SPARTA comprises five steps:

**Step 1.** For each variable  $i$  and each season  $s$ , specify a suitable target marginal distribution,  $F_{\underline{x}_s^i}$ , and also identify the dependencies to be preserved in time and space, as well as the target values of the associated coefficients of correlation,  $\rho_{s,s-\tau}^{i,j}$ .

**Step 2.** On the basis of the desirable dependencies to preserve (in terms of auto- and cross-correlations), identify the suitable auxiliary model from the PAR-N family.

**Step 3.** Determine the equivalent coefficients of correlation,  $\tilde{\rho}_{s,s-\tau}^{i,j}$ , for all pairs of variables that are required by the auxiliary model (e.g., using the algorithm of section 4.5).

**Step 4.** Estimate the parameters of the auxiliary PAR-N model, on the basis of equivalent correlations, and run the model to generate the auxiliary Gaussian synthetic time series of  $\underline{z}_{s,t}$ .

**Step 5.** Map the auxiliary process  $\underline{z}_{s,t}$  to the actual domain using their ICDFs, i.e., through Eq. (6.1), to obtain  $\underline{x}_{s,t}$ .

It is noted that, in contrast to classical stochastic approaches (see section 2.3.1 and 4.3.7), which imply the use of a specific statistical model for the noise, Nataf-based methods allow to employ pre-specified distribution models, in order to describe the probabilistic and stochastic structure of the modelled processes themselves and not of the noise.

## 6.5 CASE STUDIES

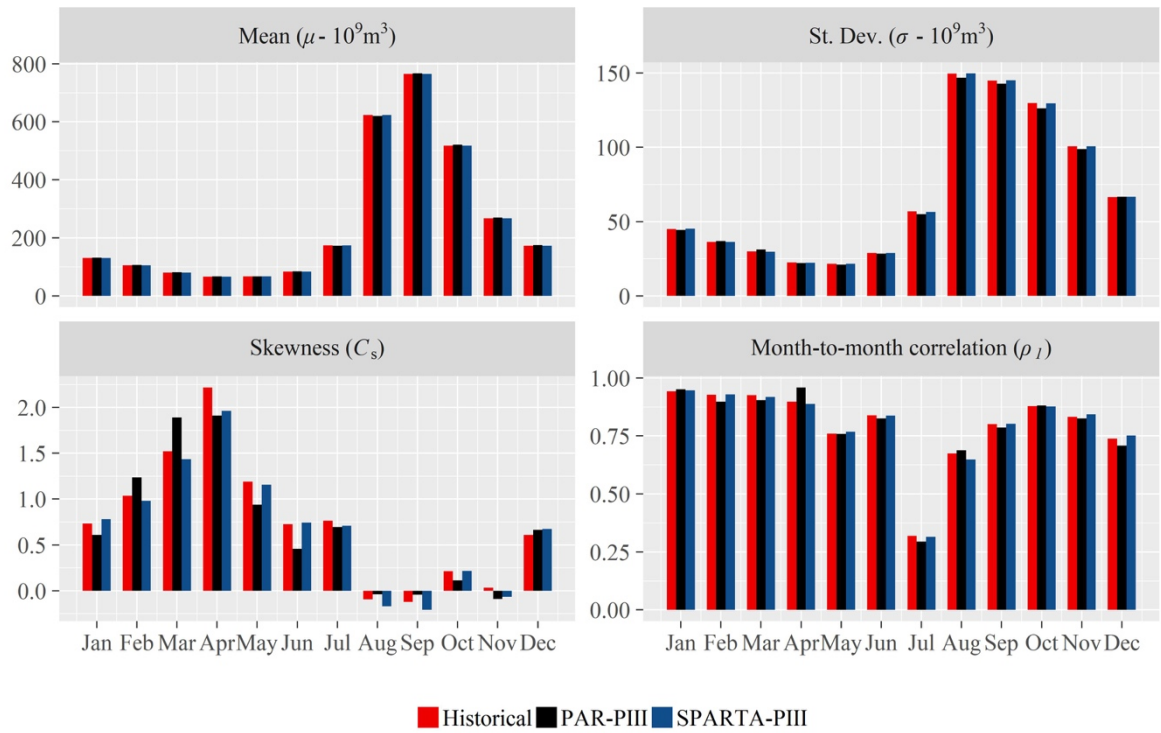
### 6.5.1 Univariate simulation with common distribution models

The first case study involves the simulation of monthly flow of Nile River at Aswan dam, based on a historical dataset from March 1870 to December 1945 [Hipel and McLeod, 1994]. The flows are characterized by strong seasonality and high correlations across all subsequent months (Figure 6.1). In order to demonstrate the performance of SPARTA against the classic implicit PAR model, we compare the outcomes of a stochastic simulation scenario of 2 000 years length, which has been used several times in the past for providing synthetic flows [e.g., Koutsoyiannis et al., 2008]. The implicit PAR(1) model is typically coupled with Pearson type-III distribution for white noise generation (referred to as PAR- $\mathcal{P}$ III model). Hence, in order to conduct a fair and meaningful evaluation, within SPARTA we also set this distribution as target one for all months (referred to as SPARTA- $\mathcal{P}$ III model). We remind that SPARTA explicitly accounts for the marginal distribution of each season, while PAR- $\mathcal{P}$ III, similarly to most linear stochastic models (see section 2.3.1), attempts to resemble the statistical characteristics (e.g., mean, variance and skewness) via implicitly representing the marginal distributions into the innovation term. The multivariate formulation of PAR- $\mathcal{P}$ III of order 1 is given in Appendix C.1.

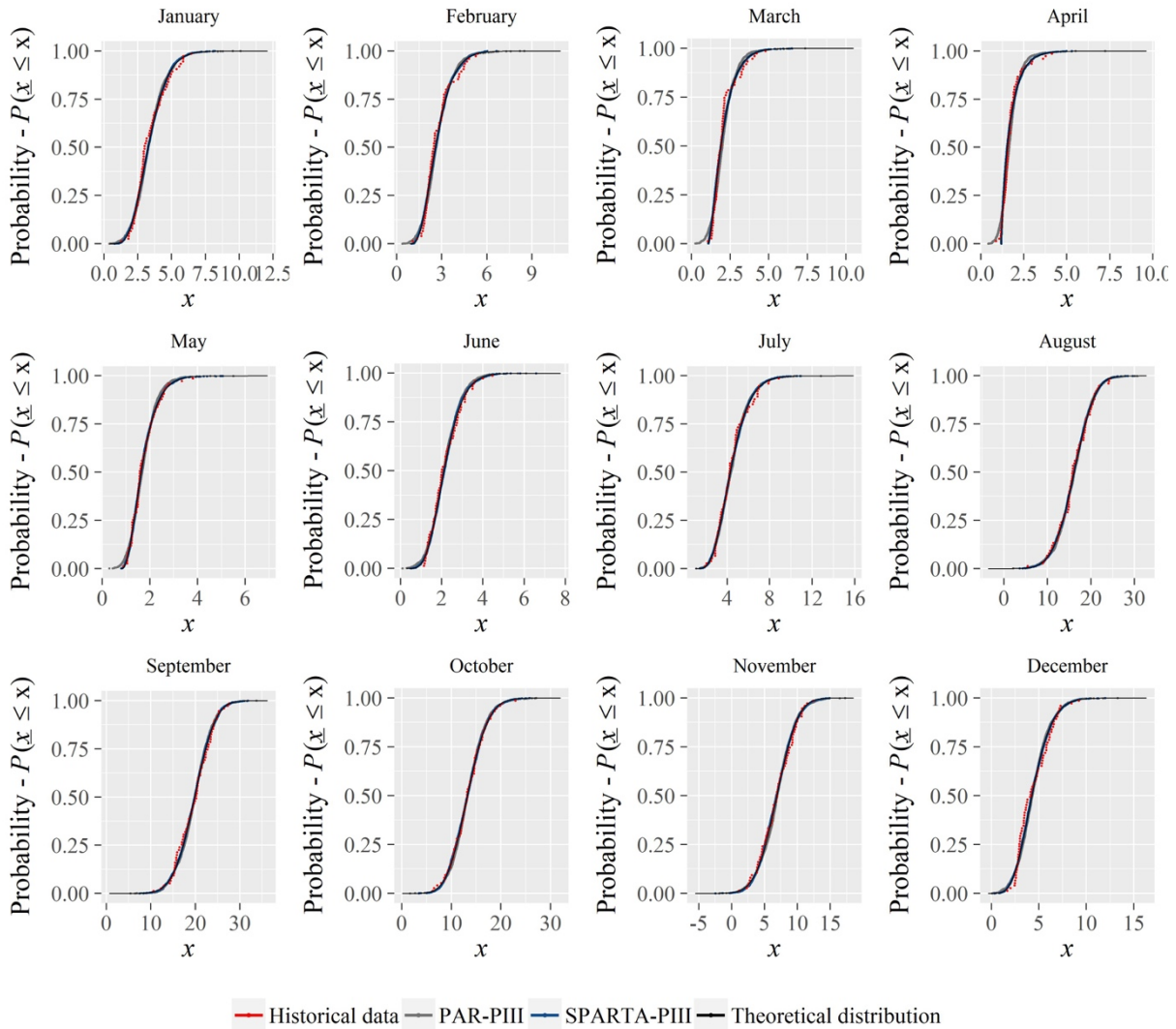
It is remarked that due to the use of Pearson type-III distribution, which allows for negative location parameters, the two models can produce negative values that would not be acceptable in a real-world hydrological study. A typical way to address this inconsistency within both models is the artificial truncation of all synthetic values to zero, which would yet introduce bias to the stochastic structure of the synthetic processes. However, among the two models, SPARTA also offers a much more rigorous alternative, since the data are generated via the corresponding ICDFs. This property enables fitting another positively bounded distribution model (e.g., Gamma, Log-Normal, etc.) to the observed data that explicitly prohibits the generation of negative values.

The two models are evaluated through visual inspection of simulated against observed values of their monthly statistical characteristics, in terms of calculated values of mean,  $\mu$ , standard deviation,  $\sigma$ , skewness coefficient,  $C_s$ , and lag-1 month-to-month correlation,  $\rho_1$  (Figure 6.1), as well as in terms of their monthly marginal distributions (Figure 6.2). It is noted that these statistics were calculated after truncation of negative values. Except for the trivial case of means and standard deviations, which are perfectly reproduced by both models, for the skewness and month-to-month correlations, only SPARTA- $\mathcal{P}$ III ensures full consistency with the target values across all seasons. In addition, SPARTA- $\mathcal{P}$ III fits perfectly the target theoretical distribution models, which is a direct outcome of employing the inverse mapping, while PAR- $\mathcal{P}$ III occasionally deviates from the target distributions, and particularly their tails (e.g., in February, March, April and May).

CYCLOSTATIONARY PROCESSES

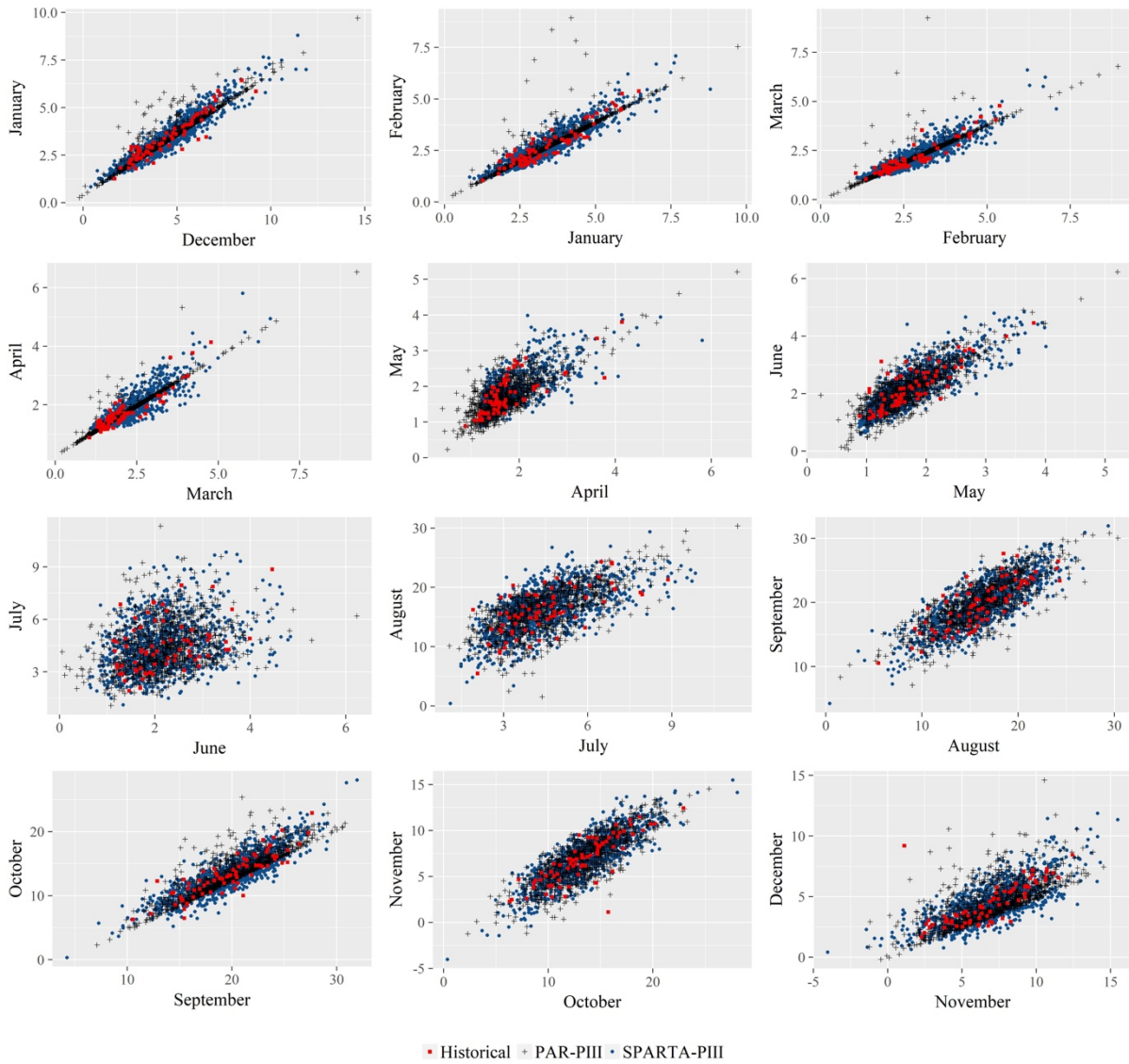


**Figure 6.1** | Comparison of key statistics ( $\mu$ ,  $\sigma$ ,  $C_s$  and  $\rho_1$ ) between historical and simulated flow data of Nile River (PAR and SPARTA).



**Figure 6.2** | Comparison between simulated flow data ( $10^9 \text{ m}^3$ ), through PAR-P-III and SPARTA-P-III, empirical and theoretical cumulative distribution functions (Weibull's plotting position). Simulated negative values are also included to avoid the distortion of the established CDFs.

To further highlight the advantages of SPARTA over PAR-P-III, we also investigate the derived dependence forms, by focusing on the scatter plots of the 12 pairs of adjacent monthly data sets (**Figure 6.3**). Interestingly, PAR-P-III, although it preserves quite satisfactory the key statistical characteristics, including the observed coefficients of correlation, it fails to capture the full extent of the observed patterns, in contrast to SPARTA-P-III, which generates well-spread data pairs which are in compliance with the observations. In particular, in the scatter plots of pairs December – January, January – February, February – March and March – April, it is shown that PAR-P-III not only fails to capture the dependence patterns of the historical data, but also seems fails to produce synthetic pairs out of a lower boundary. Therefore, the synthetic dependencies are not in good agreement with the observed ones, although the correlation coefficients themselves are reproduced with high accuracy. For further details regarding this behavior, as well as its origin, see Chapter 3.



**Figure 6.3** | Month-to-month scatter plots of historical and simulated flow data ( $10^9 \text{ m}^3$ ), through PAR- $\mathcal{P}$ III and SPARTA- $\mathcal{P}$ III. Simulated negative values are also included to avoid the distortion of the established dependence patterns.

### 6.5.2 Toy simulation with seasonally-varying distribution models

The second case study involves the simulation of a hypothetical seasonal process  $\{\underline{x}_{s,t}\}$ , with different marginal distribution per season (for convenience, 12 seasons are considered). The target distribution models and the associated parameters across seasons are given in **Table 6-1**. In addition, we assume the target lag-1 (i.e., season-to-season) correlation coefficients equal to  $\boldsymbol{\rho} = [\rho_{12,1}, \rho_{1,2}, \dots, \rho_{s,s-1}, \dots, \rho_{10,11}, \rho_{11,12}] = [0.93, 0.90, 0.76, 0.84, 0.32, 0.67, 0.80, 0.88, 0.83, 0.74, 0.94, 0.93]$ . Using SPARTA we generated  $1\,000 \times 12 = 12\,000$  synthetic values of  $\underline{x}_{s,t}$  and compared their statistical characteristics against the target ones. We remark that in contrast to the previous case study, we do not compare against another linear stochastic model (e.g., PAR- $\mathcal{P}$ III), given that we have specified different statistical distributions across seasons, which cannot be represented by such models.

**Table 6-1** | Theoretical distributions and associated parameters of hypothetical process across seasons, as well as MLE estimation of simulated data.

| Season                      | 1                | 2               | 3             | 4             | 5              | 6               | 7               | 8              | 9               | 10               | 11              | 12            |
|-----------------------------|------------------|-----------------|---------------|---------------|----------------|-----------------|-----------------|----------------|-----------------|------------------|-----------------|---------------|
| Distribution/<br>Parameters | $\mathcal{P}III$ | $\mathcal{EXP}$ | $\mathcal{G}$ | $\mathcal{N}$ | $\mathcal{LN}$ | $\mathcal{WEI}$ | $\mathcal{WEI}$ | $\mathcal{LN}$ | $\mathcal{EXP}$ | $\mathcal{P}III$ | $\mathcal{WEI}$ | $\mathcal{G}$ |
| Theoretical Values          |                  |                 |               |               |                |                 |                 |                |                 |                  |                 |               |
| $a$                         | 1.7              | 0.015           | 10            | 85            | 0.3            | 4.5             | 6               | 0.25           | 0.003           | 11               | 3               | 9             |
| $b$                         | 10               | -               | 0.15          | 30            | 5              | 680             | 820             | 6              | -               | 19               | 155             | 0.2           |
| $c$                         | 40               | -               | -             | -             | -              | -               | -               | -              | -               | -50              | -               | -             |
| Simulated Values            |                  |                 |               |               |                |                 |                 |                |                 |                  |                 |               |
| $a$                         | 1.72             | 0.015           | 10.01         | 85            | 0.29           | 4.47            | 5.99            | 0.25           | 0.003           | 9.12             | 2.97            | 9.09          |
| $b$                         | 9.88             | -               | 0.15          | 29.98         | 5              | 680.03          | 819.91          | 6              | -               | 20.98            | 154.90          | 0.20          |
| $c$                         | 39.94            | -               | -             | -             | -              | -               | -               | -              | -               | -51.39           | -               | -             |

\*Distribution abbreviations:  $\mathcal{P}III$ : Pearson type-III ( $a$  = shape,  $b$  = scale,  $c$  = location),  $\mathcal{EXP}$ : Exponential ( $a$  = 1/scale),

$\mathcal{G}$ : Gamma ( $a$  = shape,  $b$  = 1/scale),  $\mathcal{N}$ : Normal ( $a$  = mean,  $b$  = st. dev.),  $\mathcal{LN}$ : Log-Normal ( $a$  = shape (log mean),  $b$  = scale (log st. dev.)),  $\mathcal{WEI}$ : Weibull ( $a$  = shape,  $b$  = scale)

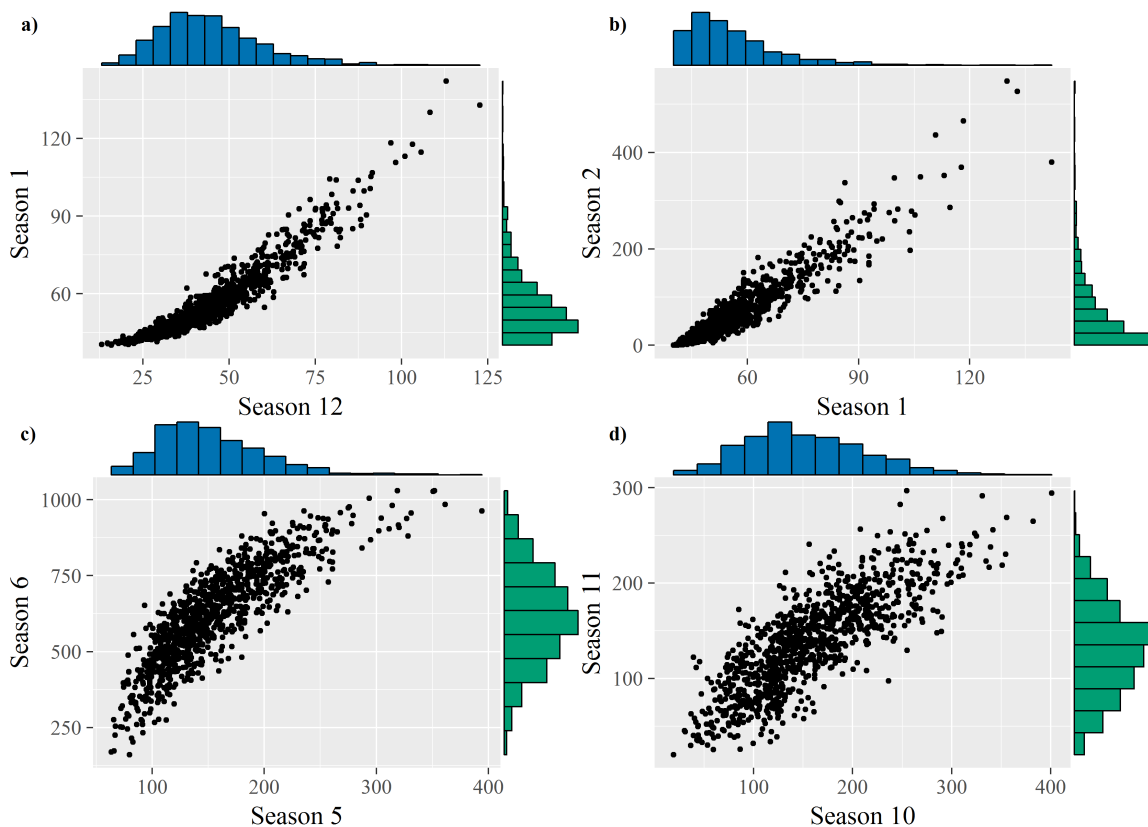
The theoretical and simulated values of the key statistical characteristics of the modelled process are illustrated in **Table 6-2**. The former were calculated through the corresponding theoretical equations of each distribution. As shown, SPARTA is very efficient, since it reproduces all key statistics, including the kurtosis coefficient,  $C_k$ . Furthermore, SPARTA preserves the parameters of the target marginal distributions (**Table 6-1**, upper part), which are estimated through the MLE method. Actually, as shown in **Table 6-1** (lower part), there is close agreement between the target and simulated parameter values for all seasons. This is also visually confirmed by plotting the associated CDFs (**Figure 6.5**), as the discrepancies between the theoretical and empirical distributions are almost indistinguishable. It is noted that the distributions employed for season 4 and 10 allowed the generation of negative values since we assigned to the former a Gaussian one (which is unbounded) and in the latter a Pearson Type-III with location parameter  $c = -50$  which coincides with its theoretical lower bound (given that  $b > 0$ ). All other distributions are defined in the positive real axis, hence they don't allow the generation of negative values.

Furthermore, the stochastic structure of the hypothetical process, by means of season-to-season correlations,  $\rho_1$ , is reproduced, despite the fact that it exhibits significant variability, also comprising some very high  $\rho_1$  values. In order to shed further light on the seasonal dependence patterns, we provide scatter plots combined with histograms for four adjacent seasons, from which it becomes clear that SPARTA can reproduce a plethora of marginal distributions and simultaneously account for dependence patterns of different complexity (**Figure 6.4**).

**Table 6-2** | Simulated and theoretical values of key statistical characteristics of hypothetical process.

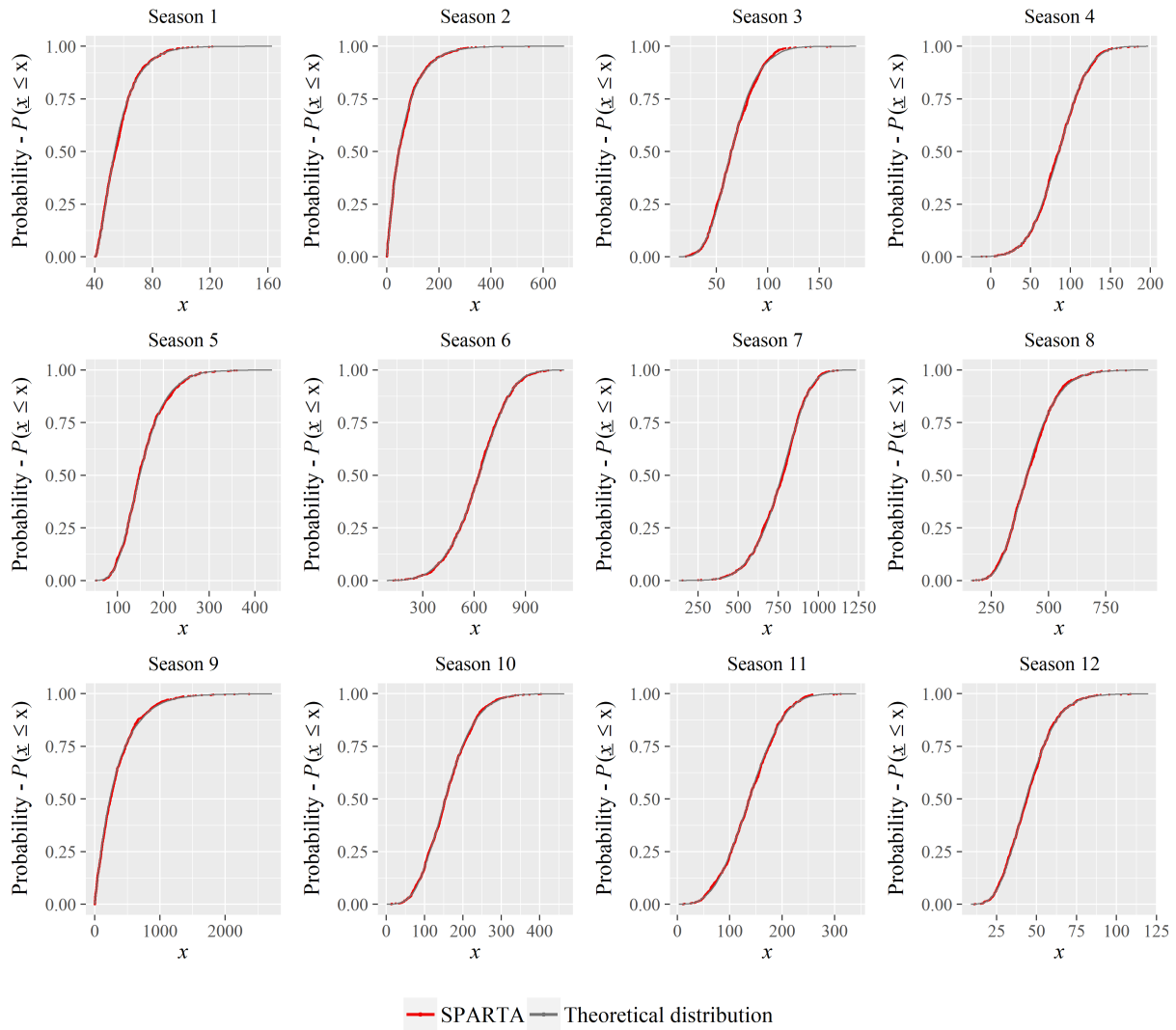
| Season/ Statistic         | 1     | 2     | 3     | 4     | 5      | 6      | 7      | 8      | 9      | 10     | 11     | 12    |
|---------------------------|-------|-------|-------|-------|--------|--------|--------|--------|--------|--------|--------|-------|
| $\mu$ (Theor.)            | 57.00 | 66.67 | 66.67 | 85.00 | 155.24 | 620.55 | 760.72 | 416.23 | 333.33 | 159.00 | 138.41 | 45.00 |
| $\mu$ (Sim.)              | 56.99 | 66.56 | 66.67 | 85.00 | 155.27 | 620.53 | 760.81 | 416.34 | 333.23 | 159.01 | 138.37 | 45.00 |
| $\sigma$ (Theor.)         | 13.03 | 66.67 | 21.08 | 30.00 | 47.64  | 156.45 | 147.40 | 105.70 | 333.33 | 63.02  | 50.30  | 15.00 |
| $\sigma$ (Sim.)           | 13.26 | 66.96 | 21.20 | 30.00 | 48.18  | 156.02 | 147.18 | 107.38 | 335.69 | 63.80  | 50.24  | 15.14 |
| $C_s$ (Theor.)            | 1.53  | 2.00  | 0.63  | 0.00  | 0.97   | -0.17  | -0.37  | 0.88   | 2.00   | 0.60   | 0.16   | 0.66  |
| $C_s$ (Sim.)              | 1.75  | 1.98  | 0.72  | -0.04 | 1.09   | -0.13  | -0.39  | 0.94   | 1.89   | 0.75   | 0.27   | 0.82  |
| $C_k$ (Theor.)            | 6.53  | 9.00  | 3.60  | 3.00  | 4.99   | 2.80   | 3.03   | 4.06   | 9.00   | 3.54   | 2.72   | 3.66  |
| $C_k$ (Sim.)              | 7.62  | 8.01  | 3.84  | 2.98  | 5.20   | 2.88   | 3.20   | 4.46   | 7.32   | 3.85   | 3.05   | 4.20  |
| $\rho_1$ (Theor.)         | 0.93  | 0.90  | 0.76  | 0.84  | 0.32   | 0.67   | 0.80   | 0.88   | 0.83   | 0.74   | 0.94   | 0.93  |
| $\rho_1$ (Sim.)           | 0.94  | 0.90  | 0.76  | 0.82  | 0.31   | 0.66   | 0.80   | 0.87   | 0.85   | 0.77   | 0.95   | 0.93  |
| $\tilde{\rho}_1$ (Equiv.) | 0.95  | 0.91  | 0.80  | 0.85  | 0.32   | 0.70   | 0.80   | 0.90   | 0.88   | 0.78   | 0.96   | 0.94  |

\*Table abbreviations: Theor: Theoretical value, Sim: Simulated value, Equiv: Equivalent value.



**Figure 6.4** | Scatter plots with histograms for a) season 12 vs. 1 b) season 1 vs. 2, c) season 5 vs. 6, and d) season 10 vs. 11.





**Figure 6.5** | Comparison between simulated (SPARTA) and theoretical cumulative distribution functions (Weibull plotting position) of hypothetical process. Simulated negative values (season 5 and 10) are also included to avoid the distortion of the established CDFs.

### 6.5.3 Multivariate simulation

The third case study involves the simultaneous generation of monthly runoff and rainfall data at two major reservoirs of the water supply system of Athens, i.e., Evinos and Mornos (details about the system are provided by [Koutsoyiannis et al. \[2003a\]](#)). The historical data cover a 29-year period (Oct/1979 – Sep/2008), which is marginally adequate for estimating up to third moment statistics with acceptable accuracy. For convenience, herein we will refer to Evinos runoff and rainfall as *sites A* and *B*, respectively, and to Mornos runoff and rainfall as *sites C* and *D*, respectively (here term *site* denotes a specific hydrological process at a specific location).

In this problem we employed the multivariate version of SPARTA and compared against the contemporaneous PAR(1) model with Pearson type-III white noise, again, referred as PAR-PIII model (Appendix C.1). Similarly to the case study of section 6.5.1, in the context of specifying the underlying marginal distributions of SPARTA, and in order to ensure fair comparisons, we decided fitting the Pearson type-III model at all sites and for all months, and estimating its parameters via the method of moments. Under this premise, the generating scheme will be next referred to as SPARTA-PIII. Although we remark, that in an operational,

*real-world study* one could take advantage of SPARTA model flexibility and select appropriate distributions models that are positively bounded, thus directly surpass the problem of negative values generation (see also the previous sections).

The performance of both models was assessed in a monthly basis, by contrasting the statistical characteristics of historical data that should be theoretically preserved by the corresponding generating schemes (i.e., monthly means, standard deviations, and skewness coefficients, lag-1 correlations across months, and zero-lag cross-correlations between all sites) against the simulated ones.

It is well-known that while the theoretical equations of any stochastic model are built in order to explicitly reproduce a specific set of statistical characteristics, this preservation is only ensured for very long (theoretically infinite) simulation horizons [Efstratiadis et al., 2014a]. If we consider relatively small horizons and repeat the simulation many times, the smaller the length of the synthetic sample, the larger is expected to be the variability of the simulated against the theoretical values of these characteristics. In this context, the stochastic model that ensures the minimum variability will be recognized as the most robust, since its performance will be the less sensitive against the simulation length. In this context, we employed two experiments, the first one by employing a single simulation of 500 000 years length, and the second one by running each model 500 times, to obtain independent synthetic samples of 1 000 years length. This Monte Carlo approach allowed for evaluating the uncertainty of the simulated statistical characteristics (after truncation of negative values to zero), which is depicted by means of box-plots (Figure 6.6 to Figure 6.10).

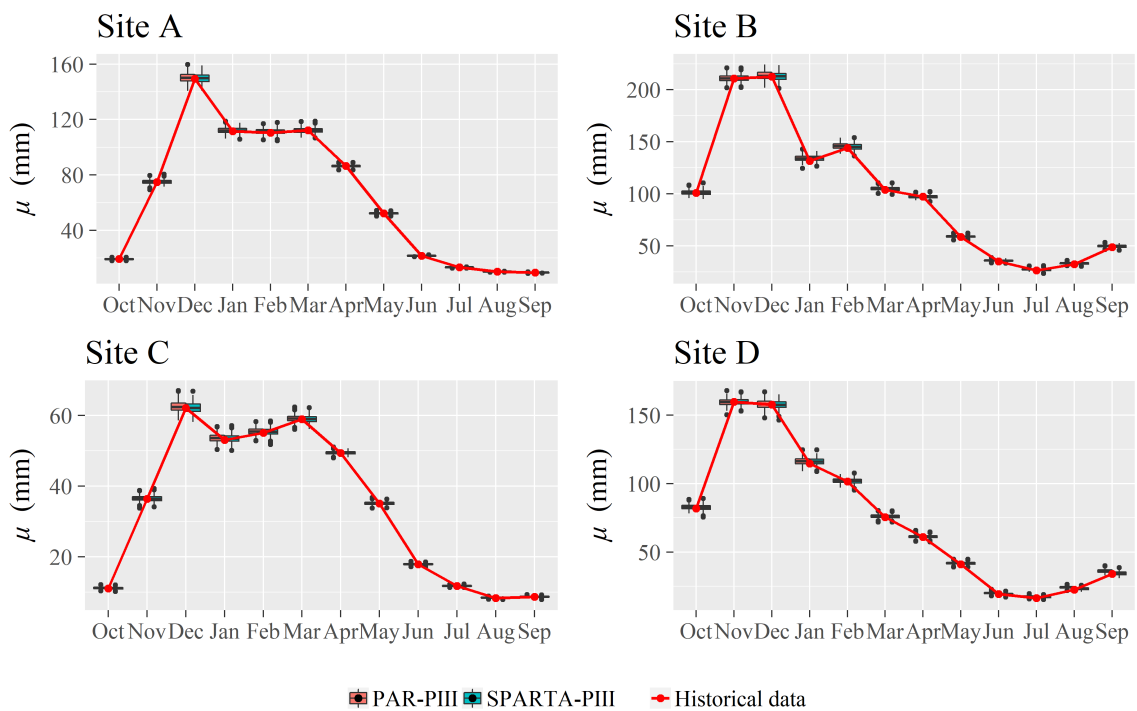
As shown in Appendix C (Figure C.1 to Figure C.5), the estimated statistical characteristics from the large (i.e., 500 000 years) synthetic sample perfectly agree with the historical ones, thus confirming the solid theoretical background of SPARTA- $\mathcal{P}_{III}$ . As expected, PAR- $\mathcal{P}_{III}$  also ensures perfect fitting of the simulated to the observed statistics, except for skewness, which are slightly underestimated. Probably, this systematic deviation is due to the simplified method employed for covariance matrix decompositions (namely, the Cholesky technique). The superiority of SPARTA- $\mathcal{P}_{III}$  against PAR- $\mathcal{P}_{III}$  is further revealed when evaluating the fitting of synthetic data to the theoretical distribution that has been adopted in this simulation experiment, i.e., Pearson type III. The  $\mathcal{P}_{III}$  distribution is mathematically defined through Eq. (3.5) comprising three parameters, i.e., shape,  $a$ , scale,  $b$ , and location,  $c$ , which have been estimated for each site and each month with the method of moments (Table C.1). It is shown that the estimated parameter values originated by SPARTA- $\mathcal{P}_{III}$  are very close to the theoretical ones, thus the desirable distributions are accurately reproduced. On the other hand, there are several cases where the PAR-derived parameters, and consequently the derived distributions, oscillate significantly from the theoretical model. This becomes even more evident when expressing these deviations in terms of root mean square error, per site and parameter. As shown in Table C.2, this error is up to three times larger than the error induced by SPARTA- $\mathcal{P}_{III}$ .

With respect to the second (i.e., Monte Carlo) experiment, from Figure 6.6 and Figure 6.7 it is shown that both SPARTA- $\mathcal{P}_{III}$  and PAR- $\mathcal{P}_{III}$  are able to reproduce the observed monthly means and standard deviations, respectively, since their variability is generally low across all sites and seasons.

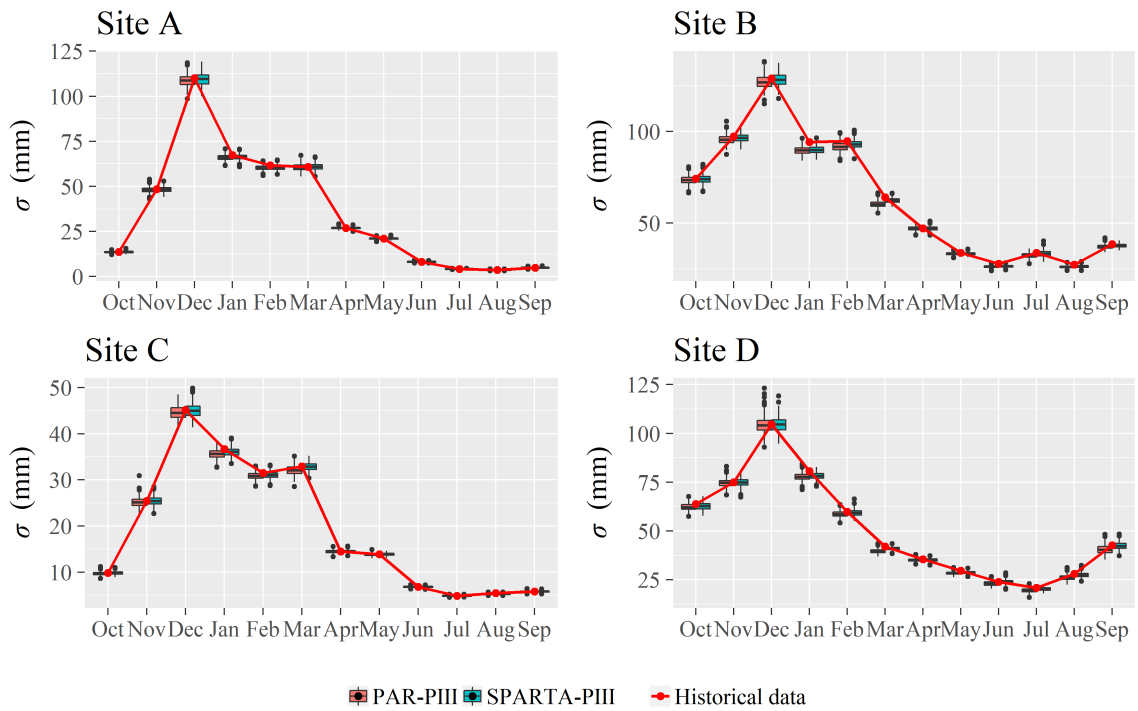
Regarding the reproduction of monthly coefficients of skewness (Figure 6.8), it seems that SPARTA- $\mathcal{P}_{III}$  slightly outperforms PAR- $\mathcal{P}_{III}$  in terms of statistical uncertainty, as indicated by the narrower box-plots that are provided in several cases (e.g., October, March, August and

September for site A, October, November and March for site B, November, December and March for site C, and March, August and September for site D). Finally, in terms of lag-1 month-to-month and lag-0 cross-correlations, both schemes ensure robustness, as illustrated in **Figure 6.9** and **Figure 6.10**, respectively.

As already highlighted, a great advantage of SPARTA over linear stochastic schemes, such as PAR- $\mathcal{P}$ III, is its ability to reproduce realistic dependence patterns, in compliance to the observed ones (see also Chapter 3). This is also empirically confirmed in the current case study, which aims to reproduce both temporal and spatial dependencies (i.e., dependencies between different processes). A characteristic example is given in **Figure 6.11**, illustrating the scatter plots of historical and simulated runoff values of at Evinos (site A) and Mornos (site C), for months January and February, from the long-term experiment (i.e., 500 000 years). It becomes now even more clear that the SPARTA- $\mathcal{P}$ III generation scheme provides reasonably-distributed data, while the synthetic data by PAR- $\mathcal{P}$ III are again bounded within a specific range, which is far from truthful and does not capture the full extent of the observed scatter (notice the incompatibility between the synthetic series of PAR- $\mathcal{P}$ III and the historical data in **Figure 6.11**).



**Figure 6.6** | Comparison of monthly mean values,  $\mu$ , of historical and synthetic data.



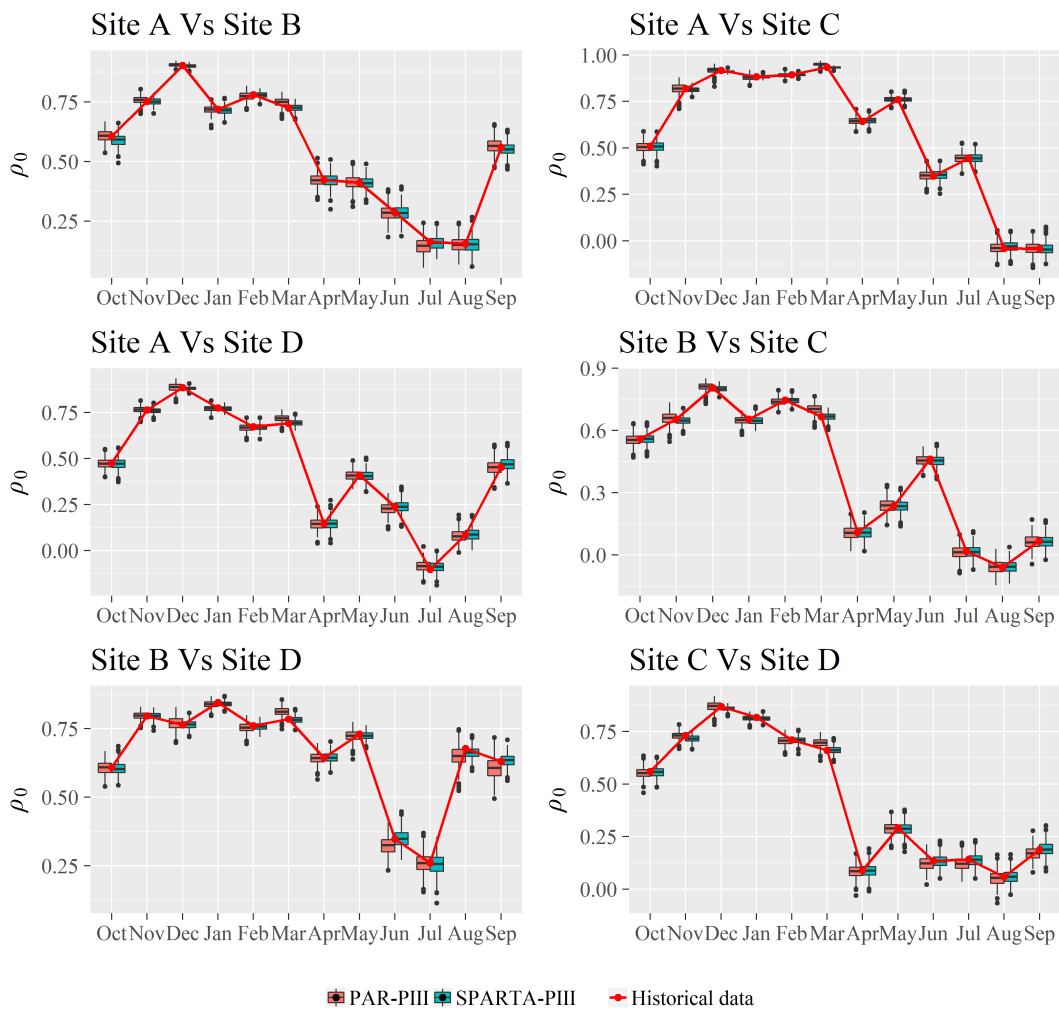
**Figure 6.7** | Comparison of monthly standard deviation values,  $\sigma$ , of historical and synthetic data.



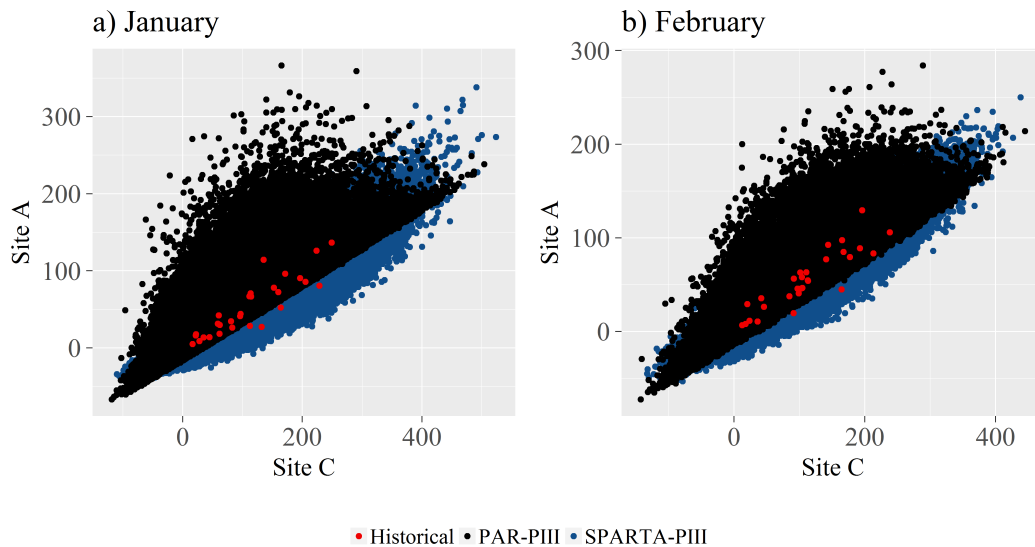
**Figure 6.8** | Comparison of monthly skewness coefficients,  $C_s$ , of historical and synthetic data.



**Figure 6.9** | Comparison of month-to-month lag-1 correlations,  $\rho_1$ , of historical and synthetic data.



**Figure 6.10** | Comparison of monthly lag-0 cross-correlations,  $\rho_0$ , between sites of historical and synthetic data.



**Figure 6.11** | Scatter plots of 500 000 synthetic data for sites A and C, representing monthly runoff (mm) processes at Evinos and Mornos reservoirs, respectively, for (a) January and (b) February. Simulated negative values are also included to avoid the distortion of the established dependence patterns.

## 6.6 SUMMARY

This Chapter presents a novel approach, termed SPARTA, for the explicit stochastic simulation of univariate and multivariate cyclostationary (i.e., periodic) processes with arbitrary marginal distributions. SPARTA uses an auxiliary Gaussian PAR process with properly identified parameters, such as after its mapping to the actual domain through the ICDFs, it results to a process with the target correlation structure and *a priori* specified marginal distributions. Since the temporal and spatial dependencies are typically expressed by means of Pearson correlation coefficients, we focus on the identification of equivalent correlation coefficients of the auxiliary processes to be used in the Gaussian domain, in order to attain the target correlations in the actual domain. In this context, we use the Nataf joint distribution model, originated from statistical sciences for the generation of correlated random variables with prescribed distributions (see also Chapter 4).

Further to the advantage of simulating cyclostationary processes with arbitrary marginal distributions, the proposed approach is also flexible in implementing any distribution fitting method, offered by recent advances in statistical sciences. This flexibility also offers the capability of explicitly ensuring the generation of non-negative values within simulations, through selecting appropriate distributions that are positively bounded. This important property, which is not offered by most of known stochastic schemes used in hydrology, is attributed to the use of the ICDF; if the employed distributions are positively bounded, the generated values will be by definition non-negative.

The advantages of SPARTA in practice, i.e., in the context of generating monthly synthetic data, have been illustrated through three stochastic simulation studies, emphasizing different aspects of the proposed methodology. Furthermore, in two out of three studies, SPARTA has been contrasted to the well-established linear stochastic model PAR-PIII, i.e., PAR(1) with Pearson type-III white noise. The major outcomes of our analyses are:

- Both models reproduced almost perfectly the essential statistical characteristics of the simulated processes up to second order (means, standard deviations, lag-1 month-to-month correlations (i.e., autocorrelations), zero-lag cross-correlations).
- SPARTA was also able to preserve with high accuracy the third order statistics, expressed in terms of skewness coefficients, while in several cases PAR-PIII provided quite underestimated skewness, which varied significantly across independently generated synthetic samples.
- SPARTA was able not only to preserve the theoretical statistical characteristics of the observed data but also the parameters of the prescribed marginal distributions, which is in fact the primary goal of simulation.
- SPARTA produced dependence structures in time and space that are in agreement with the observed patterns, while, in some cases, PAR-PIII provided rather irregular scatter patterns that were fragmented out of the observed ranges.

To this end, it is argued, that SPARTA is a convenient way to simulate cyclostationary processes, either univariate or multivariate, yet it should not be regarded as a *panacea* for all kind of simulation problems, since it inherits the characteristics of the auxiliary process from the periodic autoregressive family. In this context, it cannot preserve the statistical characteristics at aggregated time scales, e.g., annual, including long-range dependence (Hurst phenomenon).

For this reason, the Chapter 7 regards the integration of SPARTA within a multi-scale stochastic simulation framework (by coupling multiple Nataf-based models; see Chapter 5), allowing us to reproduce the desirable distribution and desirable correlation structures at multiple time scales, and also reproduce the peculiarities of different scales. As shown in the literature, an effective and efficient way to address this is through disaggregation techniques. For instance, the coupling procedures formalized by *Koutsoyiannis and Manetas [1996]* and *Koutsoyiannis [2001]*, which has been successfully implemented within advanced simulation schemes [e.g., *Efstratiadis et al., 2014a*; *Kossieris et al., 2016*; *Tsoukalas et al., 2018c*], can be easily aligned with SPARTA and other Nataf-based models to ensure statistical consistency across scales.

## BUILDING A PUZZLE FOR MULTI-TEMPORAL STOCHASTIC SIMULATION

\*

---

### PREAMBLE

The generation of hydrometeorological time series that exhibit a given probabilistic and stochastic behavior across multiple temporal levels, traditionally expressed in terms of specific statistical characteristics of the observed data, is a crucial task for risk-based water resources studies, and simultaneously a puzzle for the community of stochastics. The main challenge stems from the fact that the reproduction of a specific behavior at a certain temporal level does not imply the reproduction of the desirable behavior at any other level of aggregation. In this respect, we first introduce a pairwise coupling of Nataf-based stochastic models within a disaggregation scheme, and next we propose their puzzle-type configuration to provide a generic stochastic simulation framework for multivariate processes exhibiting any distribution and any correlation structure. Within case studies we demonstrate two characteristic configurations, i.e., a three-level one, operating at daily, monthly and annual basis, and a two-level one to disaggregate daily to hourly data. The first configuration is applied to generate correlated daily rainfall and runoff data at the river basin of Achelous, Western Greece, which preserves the stochastic behavior of the two processes at the three temporal levels. The second configuration disaggregates daily rainfall, obtained from a meteorological station at Germany, to hourly. The two studies reveal the ability of the proposed framework to represent the peculiar behavior of hydrometeorological processes at multiple temporal resolutions, as well as its flexibility on formulating generic simulation schemes.

This Chapter is organized as follows: Section 7.1 reviews the literature and presents the objectives of this chapter. Section 7.2 describes the disaggregation-based coupling approach, designed to maintain consistency across pairwise scales. Section 7.3 presents a generic and modular stochastic simulation framework that enables the development of various multi-scale schemes (i.e., configurations). Section 7.4 demonstrates a three-level configuration through two case studies, which highlight the capabilities of the framework to simulate a wide range of processes (multivariate) exhibiting intermittency, different distribution functions and correlation structures across multiple time scales. Section 7.5 entails a simpler configuration and aims at synthesizing hourly rainfall data from a given (i.e., observed) daily record, thus illustrating the efficiency of the method against challenging disaggregation problems. Finally, section 7.6 summarizes the overall modelling framework and discusses potential applications.

---

\* Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2018b), Building a puzzle to solve a riddle: a new approach to multi-temporal stochastic simulation, *J. Hydrol.*, doi:(in review).



## 7.1 INTRODUCTION

Today, most of water resources studies employ Monte Carlo simulations, by running deterministic models that are driven by synthetic inputs, which are typically generated by stochastic models. In this context, key requirement for extracting statistically consistent outcomes is the concise representation of the probabilistic behavior and stochastic structure of the input hydrometeorological processes (e.g., rainfall, runoff, temperature). It is well-known that these exhibit a significantly complex regime, the most prominent aspects of which are non-Gaussianity, intermittency, auto- and cross-dependence, as well as periodicity [Moran, 1970; Salas et al., 1980; Koutsoyiannis, 2005b]. All above peculiarities dictate the specifications of a good simulation model (see also of section 2.2).

During more than a half century, the need for *good* synthetic data generators, to be used within risk-aware decision-making frameworks for design, assessment and operation of water resource systems (see section 1.2) has triggered numerous researchers for developing a plethora of stochastic approaches and associated modelling tools. These can be primarily classified into two broad categories, i.e., single-scale and multi-scale. The former ensure the reproduction of a set of statistical and stochastic properties at a unique time scale of interest, i.e., the time interval of simulation, while the latter attempt to simultaneously represent the desirable properties of the simulated data, as well as the properties of the aggregated data at coarser temporal scales.

The numerous single-scale simulation schemes that have been developed so far can be further distinguished into (see the review of section 2.3): 1) linear stochastic models, also known as time series generators [e.g., Thomas and Fiering, 1962; Matalas, 1967; Matalas and Wallis, 1976; Salas et al., 1980; Bras and Rodríguez-Iturbe, 1985; Koutsoyiannis, 1999, 2000]; 2) point process models [e.g., Bo et al., 1994; Onof et al., 2000; Kilsby et al., 2007; Burton et al., 2008; Evin and Favre, 2008; Kaczmarek et al., 2014]; 3) two-part models, i.e. product models of occurrence and amount that are represented as discrete and continuous processes, respectively [e.g., Todorovic and Woolhiser, 1975; Katz, 1977; Richardson and Wright, 1984; Wilks, 1998; Khalili et al., 2009; Srikanthan and Pegram, 2009; Baigorria and Jones, 2010; Breinl et al., 2013, 2015; Ailliot et al., 2015; Lee, 2016]; 4) resampling methods [e.g., Lall and Sharma, 1996; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001; Wójcik and Buishand, 2003; Clark et al., 2004; Mehrotra et al., 2006; Mehrotra and Sharma, 2007; Salas and Lee, 2010]; and 5) copula-based models [e.g., Bárdossy and Pegram, 2009; Serinaldi, 2009a; Hao and Singh, 2011, 2013; Lee and Salas, 2011; Chen et al., 2015; Jeong and Lee, 2015; Lee, 2017].

By design, single-scale simulation models attempt to reproduce the desirable statistical and stochastic behavior within the synthetic data at the scale of simulation, yet they provide limited control to the properties of the same process, when aggregated at higher (coarser) time scales. It is well-known that the reproduction of the probabilistic and stochastic behavior of a process, expressed either in terms of a distribution function or a set of statistical properties, at a certain time scale does not ensure the reproduction of the associated characteristics of the aggregated process at any other time scale.

The necessity for the hereto referred to as *multi-scale consistency* has been early recognized by the hydrological community, through the pioneering work by Harms and Campbell [1967]. Actually, from the first steps of Monte Carlo approaches in water resources it has been accepted that that the outcomes of stochastic analyses are associated with the overall statistical and stochastic behavior of the input hydrometeorological processes, which may extend far beyond the time interval of the underlying (deterministic) simulation model [see, Klemeš, 1981;

[Koutsoyiannis, 2005b](#)]. For instance, the design and operation of large reservoir systems that employ overyear regulation, which are typically modelled in monthly intervals, is strongly dictated by the probabilistic and stochastic properties of the aggregated inflows, at the annual and even over-annual scales. Similarly, the outputs of continuous flood simulation models, driven by fine-time (e.g., hourly) rainfall series, are substantially affected by the sequence of accumulated rainfall, as the runoff production strongly depends on the antecedent soil moisture conditions. In this respect, multi-scale consistency in stochastic simulation can be regarded as an operational *sine qua non*.

Furthermore, multi-scale consistency is directly linked with the so-called issue of low-frequency variability or over-depression (i.e., the deficiency to reproduce the process' variance at higher time scales), which is encountered in many popular daily *weather-generation* models [e.g., [Wilks, 1998](#); [Katz and Parlange, 1998](#); [Wilks and Wilby, 1999](#); [Mehrotra et al., 2006](#); [Brissette et al., 2007](#); [Srikanthan and Pegram, 2009](#); [Khalili et al., 2009](#); [Serinaldi, 2009a](#); [Baigorria and Jones, 2010](#); [Mhanna and Bauwens, 2012](#); [Breinl et al., 2013, 2015](#); [Lee, 2017](#)].

Multi-scale simulation schemes, with the exception of few specifically designed models [e.g., [Rodriguez-Iturbe et al., 1987](#); [Langousis and Koutsoyiannis, 2006](#)], is typically build upon the disaggregation paradigm. Essential element of disaggregation is the *additive property*, which enables the generation of multi-scale consistent time series via the transfer of information among different temporal scales. This implies that the sum of the generated variables at the lower level (e.g., monthly) at any period should add to the corresponding value at the higher level (e.g., annual), which is assumed known, either from observed or synthetic (simulated) data. This property distinguishes disaggregation from downscaling [e.g., [Wilks and Wilby, 1999](#); [Cannon, 2008](#); [Lombardo et al., 2012](#)], which focus on generating lower level time series that statistically resemble the properties of higher level ones, and not necessarily honor the additive constraint.

As already mentioned, the beginning of the quest (at least in hydrological domain) for multi-scale simulation models can be attributed to [Harms and Campbell \[1967\]](#), who developed a two-level version of the classical stochastic model by [Thomas and Fiering \[1962\]](#) that preserves some key statistical properties of the observed data at both the annual and monthly scale. Little later, the interest on such methods reinforced with the theoretical research on disaggregation by [Valencia and Schakke \[1973\]](#) and [Mejia and Rousselle \[1976\]](#). However, the proposed methods were fully general only for normally distributed variables, thus limiting their applicability to a relatively narrow range of processes and scales.

Next generation approaches offered multi-scale schemes that utilized the notion of the so-called *adjusting procedures* [[Harms and Campbell, 1967](#); [Stedinger and Vogel, 1984](#); [Grygier and Stedinger, 1988](#); [Koutsoyiannis and Manetas, 1996](#); [Koutsoyiannis, 2001](#)]. These aimed in coupling single-scale simulation models of any type, operating independently at different time scales. The rationale is generating low-level synthetic data as auxiliary information, and next adjusted them to the known higher-level values, by using relatively simple algebraic transformations, such as the partial sums at the low level equal the values of the higher level. [Koutsoyiannis and Manetas \[1996\]](#) and [Koutsoyiannis \[2001\]](#) investigated several adjusting procedures, and also standardized the concept of *repetitive sampling* (kind of Monte Carlo approach), to ensure that the partial sums are close to the given values. This can be regarded as an informal method of conditional sampling, that can significantly improve the efficiency of such schemes [see also, [Glasbey et al., 1995](#)].

Adjusting procedures of varying complexity have been implemented within a number of disaggregation-based schemes, in order to couple single-scale simulation models (such as the ones described above) across various time scales. In particular, they were used within linear stochastic models [e.g., *Koutsoyiannis et al.*, 2003b; *Segond et al.*, 2006; *Lombardo et al.*, 2012; *Efstratiadis et al.*, 2014a; *Allard and Bourotte*, 2015; *Tsoukalas et al.*, 2018c], point processes [e.g., *Glasbey et al.*, 1995; *Koutsoyiannis and Onof*, 2001; *Onof et al.*, 2005; *Kossieris et al.*, 2016], two-part models [e.g., *Shao et al.*, 2016; *Evin et al.*, 2018], resampling methods [e.g., *Lee et al.*, 2010] and copula-based models [e.g., *Gyasi-Agyei*, 2011; *Gyasi-Agyei and Melching*, 2012]. It is highlighted that the overall simulation capabilities of adjusting-based schemes are determined by the underlying simulation models, which consist the core data generation mechanism.

In addition, several modern schemes for establishing multi-scale consistency are built upon the concepts of scaling and multifractality [e.g., *Tessier et al.*, 1996; *Kantelhardt et al.*, 2006; *Veneziano et al.*, 2006]. Typically, these employ multiplicative random cascade models [*Gupta and Waymire*, 1990, 1993] to generate multi-scale consistent (in terms of typically high-order moments) realizations [e.g., *Menabde et al.*, 1997; *Olsson*, 1998; *Deidda et al.*, 1999; *Molnar and Burlando*, 2005; *Rupp et al.*, 2009; *Müller and Haberlandt*, 2015, 2018]. Recent works by, *Liczner et al.* [2011], *Lombardo et al.* [2012] and *Pui et al.* [2012] provide comparative studies involving such models, as well as alternative downscaling or disaggregation methods.

Besides the vast effort made so far, the quest for full generality and full consistency across multiple temporal scales still remains a *puzzle*. Recently, *Tsoukalas et al.* [2018e] highlighted that many of widespread schemes, including linear stochastic models with non-Gaussian innovations, point-process models and resampling techniques, emphasize on the reproduction of a specific set of summary statistical characteristics, which arguably cannot capture the full behavior of a random process. As also shown, under some common conditions these may lead to bounded dependence patterns, which are not realistic (see Chapter 3; [*Tsoukalas et al.*, 2018a]). On the other hand, two-part and copula-based models are actually able to explicitly account for the distributional properties of simulated processes, yet they are mainly designed to represent specific correlation structures. For instance, two-part models often neglect temporal dependencies, while copula-based schemes typically account for temporal dependencies spanning over only few time lags.

In this Chapter, our focus is not on disaggregation *per se*, rather than we employ the flexibility provided by the concepts of repetitive sampling and adjusting procedures to link individual stochastic models, in order to represent the varying regime of hydrometeorological processes across multiple temporal scales. Our emphasis is to shift from the classical paradigm of resembling a process in terms of few summary statistics (in particular, moments up to third order and low order correlation coefficients), to the explicit representation of its marginal and stochastic properties, in terms of distribution functions and theoretical correlation structures, respectively. This is accomplished by building upon a recently introduced (in hydrology) class of stochastic models, the so-called *Nataf-based* [*Tsoukalas et al.*, 2017a, 2018e, 2018d]. These, through the mapping of an auxiliary Gaussian process (Gp) (Chapter 4-6), are able to simulate multivariate, stationary and cyclostationary processes with any marginal distributions and any correlation structures. These properties allow for characterizing Nataf-based models as *good* single-scale stochastic simulators, and thus appropriate data generators within multi-scale adjusting-based schemes. Taking advantage of the above concepts, we propose a scale-free disaggregation approach for pairwise coupling of Nataf-based models, next referred to as *Nataf-based Disaggregation to Anything* (NDA). Eventually, a chain configuration NDA allows

for developing *puzzle-type*, i.e., modular, simulation schemes that ensure consistent simulations across any sequence of temporal scales.

## 7.2 ADDRESSING MULTI-SCALE CONSISTENCY

Provided that the theoretical background of Nataf-based stochastic models has been extensively discussed (Chapter 4-6), and in order to avoid repetition, this section focuses on addressing the problem of multi-scale consistency. Nataf-based models, although fulfill the requirements of a *good* stochastic model, i.e., the explicit reproduction of any distribution and any correlation structure, do not account for multi-scale consistency. Since the problem is independent of the generation procedure and the time scale of simulation, we first provide a global overview and then propose a generic solution for Nataf-based models, herein referred to as Nataf-based Disaggregation To Anything (NDA).

### 7.2.1 Problem description

Let us begin from the univariate case, denoting by  $\{\underline{\omega}_n\}_{n \in \mathbb{Z}^>}$  a discrete-time, stationary or cyclostationary (the season indicator  $s$  is omitted for simplicity), stochastic process at time scale  $k = 1$ , where  $n$  is a time index. Let also define the aggregated process  $\underline{\omega}_l^{(k)}$  at a higher time scale  $k \in \mathbb{Z}^{\geq 2}$ , obtained by:

$$\underline{\omega}_l^{(k)} = \sum_{n=(l-1)k+1}^{kl} \underline{\omega}_n \quad (7.1)$$

where  $l$  is the time index of the aggregated process. Alternatively (e.g., if  $\underline{\omega}_n$  refers to an instantaneous quantity), we can define the averaged process, also denoted by  $\underline{\omega}_l^{(k)}$ , by,  $\underline{\omega}_l^{(k)} = \sum_{n=(l-1)k+1}^{kl} \underline{\omega}_n / k$ . Apparently, the properties of  $\{\underline{\omega}_n\}$  at scale  $k = 1$  are related with those of the aggregated (or averaged) process at a higher time scale  $k \in \mathbb{Z}^{\geq 2}$ .

Herein, without loss of generality, we focus on the aggregated case. To simplify, we first remark that the operations implied by Eq. (7.1), can be viewed as a sum of  $k$  RVs. Thus, if we were interested on the distribution of  $\{\underline{\omega}_l^{(k)}\}$ , it would be identical to solve an aggregated distribution problem. If the process  $\{\underline{\omega}_n\}$  is stationary at  $k = 1$ , then at any higher scale  $k$  we would have the sum of  $k$  identical RVs. On the other hand, if  $\{\underline{\omega}_n\}$  is cyclostationary at the lower scale  $k = 1$ , at any higher scale  $k$  we would have the sum of  $k$  non-identical RVs (their marginal and dependence properties depend on the season  $s = 1, \dots, S$ , implied by the time index  $n$ ; see section 4.3.1).

Arguably, the problem of identifying the distribution of  $\underline{\omega}_l^{(k)}$  at  $k > 1$  is particularly challenging, since there is not a general method (without resorting to simulation) to identify the distribution of the sum of  $k$  RVs, especially, in the presence of dependence, which is typical for hydrometeorological processes. Furthermore, apart from some low order moments (i.e., mean, variance, autocovariance and autocorrelation), higher order moments of the aggregated process are also particularly difficult to estimate, either analytically or theoretically. Analogously, it is also challenging to specify a process  $\{\underline{\omega}_n\}$  that has the desirable (for this time scale) marginal and stochastic properties, when it is aggregated at a higher scale  $k > 1$ .

The problem becomes even harder when multiple processes are involved, in the context of multivariate simulation problems. Let  $\underline{\xi}_t = [\underline{\xi}_t^1, \dots, \underline{\xi}_t^m]^\top$  and  $\underline{\omega}_n = [\underline{\omega}_n^1, \dots, \underline{\omega}_n^m]^\top$  be two  $m$ -dimensional vectors of two discrete-time processes  $\underline{\xi}_t^i$  and  $\underline{\omega}_n^i$ , indexed using  $t \in \mathbb{Z}^>$  and  $n \in \mathbb{Z}^>$ , respectively. Furthermore, let assume that  $\underline{\xi}_t^i$  and  $\underline{\omega}_n^i$  represent the same process at two different temporal scales, higher and lower, respectively, with time units denoted by  $\delta_{\underline{\xi}}$  and  $\delta_{\underline{\omega}}$ , respectively (i.e.,  $\delta_{\underline{\xi}} > \delta_{\underline{\omega}}$ ).

Similarly to Eq. (7.1), when  $k^* := k = \delta_{\underline{\xi}}/\delta_{\underline{\omega}}$  (e.g., 1 year/1 month = 12, or 1 month/1 hour =  $28 \times 24$ ,  $30 \times 24$ ,  $31 \times 24$ ; depending on the number of days of the month), we obtain an aggregated process at the same temporal level of  $\underline{\xi}_t^i$ , i.e.,

$$\underline{\xi}_t^i := \underline{\omega}_l^{i:(k^*)} = \sum_{n=(l-1)k^*+1}^{lk^*} \underline{\omega}_n^i, (l = t) \quad (7.2)$$

Evidently, when  $\underline{\omega}_n^i$  is simulated without reference to the higher-level process  $\underline{\xi}_t^i$ , then  $\underline{\xi}_t^i \neq \underline{\xi}_t^i$ . Hence, for each process  $i = 1, \dots, m$ , our target is to generate a  $k^*$ -dimensional random sequence,  $\underline{\omega}_{t:(k^*)}^i = [\underline{\omega}_{(t-1)k^*+1}^i, \dots, \underline{\omega}_{tk^*}^i]$ , of the low-level process ( $k = 1$ ), with the desirable properties, which honors the equality,  $\underline{\xi}_t^i = \underline{\xi}_t^i$ , when aggregated to the time scale  $k^*$ . The multivariate formulation of the problem is written as:

$$\underline{\Omega}_{t:(k^*)} = [\underline{\omega}_{t:(k^*)}^1, \dots, \underline{\omega}_{t:(k^*)}^m]^\top = \begin{bmatrix} \underline{\omega}_{(t-1)k^*+1}^1 & \cdots & \underline{\omega}_{tk^*}^1 \\ \vdots & \ddots & \vdots \\ \underline{\omega}_{(t-1)k^*+1}^m & \cdots & \underline{\omega}_{tk^*}^m \end{bmatrix}, \text{ and} \quad (7.3)$$

$$\underline{\xi}_t = [\underline{\xi}_t^1, \dots, \underline{\xi}_t^m]^\top = \begin{bmatrix} \sum_{n=(t-1)k^*+1}^{tk^*} \underline{\omega}_n^1, \dots, \sum_{n=(t-1)k^*+1}^{tk^*} \underline{\omega}_n^m \end{bmatrix}^\top$$

### 7.2.2 The NDA approach: Step-by-step implementation

In order to address the problem, we develop the Nataf-based Disaggregation To Anything (NDA) approach, which combines Nataf-based models, considered as data generation mechanisms, with a coupling procedure that encompasses the notions of repetitive sampling and adjusting procedures. These two key notions are thoroughly discussed by [Koutsoyiannis and Manetas \[1996\]](#).

The NDA procedure starts from a given realization,  $\underline{\xi}_t$ , of a process  $\underline{\xi}_t$ , at a specific time scale, aiming to produce a consistent realization,  $\underline{\omega}_n$ , at a lower scale. The given realization  $\underline{\xi}_t$  is known either from observations or already generated by another model (deterministic or stochastic). In the second case, if a Nataf-based model is employed, the synthesized higher-level realization would have the desirable marginal distributions and correlation structure, hence the problem would reduce to generating a lower-level realization with the target properties, which when aggregated to the higher-level honors the additive property. Fulfilling both conditions allows preserving the properties of the process at both temporal levels, given that the realization at the higher level is kept as is.

Therefore, given the realization  $\xi_t$ , and assuming a *temporary* Nataf-based lower-level process, denoted by  $\tilde{\omega}_n$ , with properties identical to those of the target process  $\omega_n$  (i.e.,  $\tilde{\omega}_n = \omega_n$ ), the following steps are applied for all time indices  $t$ .

- 1) Using a Nataf-based model (Chapter 4-6), generate  $N_{\xi/\omega}$  temporary realizations  $\tilde{\omega}_n$  of the lower level process  $\tilde{\omega}_n$ , of length  $k^*$ , thus obtaining  $N_{\xi/\omega}$  sets of matrices  $\{\tilde{\Omega}_{t;(k^*)}(\nu); \nu = 1, \dots, N_{\xi/\omega}\}$ .
- 2) For each of the  $N_{\xi/\omega}$  matrices  $\tilde{\Omega}_{t;(k^*)}$ , estimate the corresponding vector  $\check{\xi}_t$  and obtain a set of vectors  $\{\check{\xi}_t(\nu); \nu = 1, \dots, N_{\xi/\omega}\}$ .
- 3) Calculate the difference between  $\check{\xi}_t(\nu)$  and the known  $\xi_t$  using a distance metric,  $e_t(\nu) = D(\check{\xi}_t(\nu), \xi_t)$ . See also Eq. (7.4).
- 4) Formulate the set  $\{e_t(\nu); \nu = 1, \dots, N_{\xi/\omega}\}$  and select the realization  $\tilde{\Omega}_{t;(k^*)}(\nu)$  with the minimum value of  $e_t(\nu)$ , hereafter denoted  $\Omega'_{t;(k^*)}$  (the breve notation has been omitted for simplicity). Under this premise, by aggregating  $\Omega'_{t;(k^*)}$  to time scale  $k^*$ , thus obtaining the corresponding sum  $\check{\xi}'_t$ , its difference with the target values of  $\xi_t$  will be the minimum over the simulated set.
- 5) Produce the final values of  $\Omega_{t;(k^*)}$  by adjusting the remaining difference between  $\check{\xi}'_t$  and  $\xi_t$ , by employing a specific adjusting procedure. Herein we employ the proportional adjusting procedure of Eq. (7.5).

We remark that since we employ Nataf-based models, in order to ensure a proper sequential generation procedure, it is essential to maintain an archive of the realizations generated by the auxiliary Gaussian process (Gp) model. These are needed to condition the generation mechanism on the required number of previous values. For instance, if we employ CMARTA( $p$ ) for generating the temporary realizations  $\tilde{\omega}_n$ ,  $p$  previous values of the auxiliary Gaussian realization are needed to condition the generation of  $\tilde{\omega}_{n+1}$ .

### 7.2.3 Computational details

For convenience, within repetitive sampling (step 3), we employ as distance metric the following quantity, also used by *Koutsoyiannis and Manetas [1996]*:

$$e_t = D(\check{\xi}_t, \xi_t) = \frac{1}{m} \sum_{i=1}^m |\xi_t^i - \check{\xi}_t^i| / \text{Var}[\xi_t^i] \quad (7.4)$$

On the other hand, all available adjusting procedures (APs) that are found in the literature [see, *Harms and Campbell, 1967; Grygier and Stedinger, 1988; Koutsoyiannis, 2001*] are compatible with the proposed approach. Here we employ the so-called proportional AP that can be implemented independently for each  $\omega'_{t;(k^*)}{}^i$  and reads as follows:

$$\omega_{t;(k^*)}^i = \omega'_{t;(k^*)}{}^i \xi_t^i / \check{\xi}_t^i \quad (7.5)$$

Apart from its simplicity, key advantage of this AP is the preservation of the sign of each realization  $\omega'_{t;(k^*)}{}^i$ . For instance, in case of rainfall, where the underlying Nataf-based model is combined with a mixed-type distribution to represent intermittency, the proportional

adjustment not only prohibits the generation of negative rainfall values but also preserves the sequence of zero and non-zero values, as explicitly foreseen by the auxiliary Nataf model.

A final technical issue involves the termination criteria for repetitive sampling. Here, we consider that the iterative procedure terminates when reaching a maximum number of allowable iterations,  $N_{\xi/\omega}$ . An alternative option would imply the use of a convergence criterion, by means of a similarity metric between  $\check{\xi}'_t$  and  $\xi_t$ . Nevertheless, the stopping criteria should be carefully assigned, since they control both the accuracy and computational efficiency of NDA, which are inherently conflicting. In our examples, we set  $N_{\xi/\omega} = 250$  to 350, which was heuristically identified as a fair conciliation, even for multivariate problems involving up to five individual processes.

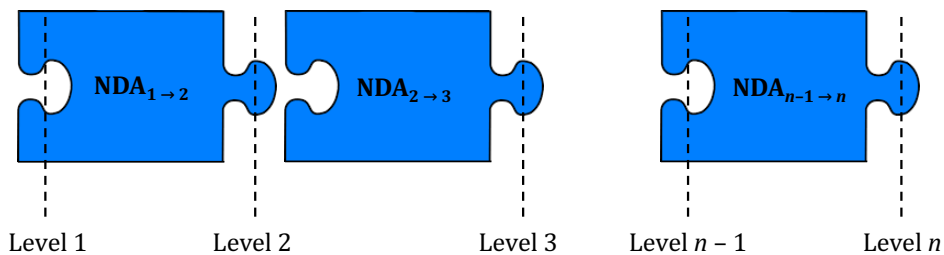
We remark that in contrast to other disaggregation schemes, where repetitive sampling had an optional role [cf. *Koutsoyiannis and Manetas, 1996*], in our approach its role is pivotal, since it allows the preservation of the advantages of Nataf-based models, and hence generate lower-level realizations with the target probabilistic and stochastic properties.

### 7.3 MODULAR FRAMEWORK FOR DEVELOPING MULTI-TEMPORAL SIMULATION SCHEMES

#### 7.3.1 Multi-temporal stochastic simulation as a puzzle

As already discussed, there does not exist a general, *bottom-up* solution to the problem of multi-scale consistency, by means of a generation procedure that provides consistent synthetic data at a time scale of interest, and simultaneously captures the scale-varying stochastic-probabilistic behavior of the aggregated process at higher time scales. In a practical context, the generally accepted requirement for a *good* stochastic model is to reproduce the desirable probabilistic and dependence properties across specific temporal scales that have operational interest. Typically, these follow the standard resolutions of hydrometeorological time series, i.e., annual, monthly, daily, hourly, etc.

In this context, we propose a *puzzle-type* implementation of NDA, to address multi-scale simulation problems of any complexity. Essentially, this can be done by coupling, in a pairwise manner, multiple Nataf-based models, which operate independently of each other. Thereby, one can establish a modular, top-down approach, starting from the first level, which corresponds to the highest time scale of interest, and subsequently moving to next levels, until reaching the lowest scale, which is dictated by the simulation problem at hand. As shown in **Figure 7.1**, each individual coupling of subsequent scales through NDA can be considered as the *pieces* of a puzzle. The generic design of NDA ensures flexibility regarding the combination of temporal scales, while at the same time, the robustness of the underlying Nataf-based approach ensures the preservation of the desirable process properties.



**Figure 7.1** | The stochastic simulation framework as a puzzle, involving a chain implementation of individual NDA *pieces*.

For demonstration, we next present a typical configuration of this puzzle, by means of a three-level scheme for annual to daily simulation, which is of significant interest for a wide range of operational hydrological problems. In section 7.4, we explore the capacities of this configuration, in the context of a real-world case study (an additional one is presented in Appendix D), involving the generation of synthetic daily rainfall and runoff series. Moreover, in section 7.5, we present another useful configuration, this time for handling a classical disaggregation problem, i.e., the generation of hourly rainfall from a given daily time series.

### 7.3.2 Three-level configuration for annual to daily simulations

In this configuration we couple three Nataf-based models, shown in Table 7-1, to provide a multivariate three-level simulation scheme. This modular scheme (i.e., puzzle) aims to preserve the probabilistic and dependence properties of typical hydrometeorological processes at the annual, monthly, and daily scales. From the models of Table 7-1, SMARTA and CMARTA are designed for stationary processes, while SPARTA for cyclostationary ones (i.e., accounting for the season-to-season correlations). A common characteristic of the three models is the direct reproduction of lag-0 cross-correlations coefficient among multiple contemporaneous processes. It is stressed that, regardless the choice of the auxiliary Gp model, in order to generate realizations with the equivalent correlation structure, the model parameters have to be estimated using the equivalent correlation coefficients.

**Table 7-1** | Summary of employed Nataf-based models ( $p$  and  $q$ , denote the order of the model).

| Auxiliary Gp model | Associated Nataf-based model | Type            | References |
|--------------------|------------------------------|-----------------|------------|
| SMA( $q$ )         | SMARTA( $q$ )                | Stationary      | Chapter 5  |
| CMAR( $p$ )        | CMARTA( $p$ )                | Stationary      | Chapter 5  |
| PAR( $p$ )         | SPARTA( $p$ )                | Cyclostationary | Chapter 6  |

**Abbreviations:** SMA (Symmetric Moving Average), CMAR (Contemporaneous Multivariate AutoRegressive), PAR (Periodic AutoRegressive), SMARTA (Symmetric Moving Average nearLy To Anything), CMARTA (Contemporaneous Multivariate AutoRegressive nearLy To Anything), SPARTA (Stochastic Periodic AutoRegressive To Anything).

To elaborate on the devised configuration, let us first introduce some notation regarding the main assumptions and specifications. Let  $\underline{y}_t = [\underline{y}_t^1, \dots, \underline{y}_t^m]^T$  be a vector of  $m$  stationary stochastic process at the annual time scale (where  $t \in T_y$  denotes the time index, i.e., year, over the set  $T_y$ ). In the context of this configuration we model the annual processes using SMARTA, in order to preserve:

- the distribution function of  $\underline{y}_t^i$ , i.e.,  $F_{\underline{y}^i}(\underline{y})$ ;
- its autocorrelation structure,  $\rho_{\underline{y}^i; \tau}^i = \text{Corr}[\underline{y}_t^i, \underline{y}_{t+\tau}^i]$ ;
- the lag-0 cross-correlations among processes  $\underline{y}_t^i$  and  $\underline{y}_t^j$ , i.e.,  $\rho_{\underline{y}^i, \underline{y}^j}^{i,j} = \text{Corr}[\underline{y}_t^i, \underline{y}_t^j]$ .

On the other hand, the standard hypothesis for the monthly time scale is cyclostationarity. Let the monthly process be represented by a  $m$  dimensional vector  $\underline{x}_{s,n} = [\underline{x}_{s,n}^1, \dots, \underline{x}_{s,n}^m]^T$ , where  $s$  ( $= 1, \dots, 12, 1, \dots, 12, \dots$ ) denotes the month and  $n \in T_x$  is the time index. The index  $t$  of the annual process (i.e., the year) may be recovered by  $t = 1 + (n - s)/12$ . For monthly simulation we employ SPARTA in order to resemble:

- the seasonally-varying marginal distribution of  $\underline{x}_{s,n}^i$ , i.e.,  $F_{\underline{x}^i}(x) = F_{\underline{x}_{s+12}^i}(x)$ ;



- the lag-1 month-to-month correlation coefficients  $\rho_{\underline{x}_s, \underline{x}_{s-1}}^i = \text{Corr}[\underline{x}_s^i, \underline{x}_{s-1}^i]$ ;
- the lag-0 cross-correlations among processes  $\underline{x}_s^i$  and  $\underline{x}_s^j$  for each season  $s$ , i.e.,  $\rho_{\underline{x}_s}^{i,j} = \text{Corr}[\underline{x}_s^i, \underline{x}_s^j]$ .

Finally, the hydrometeorological processes at sub-monthly time scales (e.g., daily) are typically regarded to be cyclically stationary within in each month  $s$ . In this respect, let  $\underline{w}_{s;d} = [\underline{w}_{s;d}^1, \dots, \underline{w}_{s;d}^m]^T$  be a  $m$ -dimensional vector of stationary processes at month  $s$ , where  $d \in T_{\underline{w}_s}$ , denotes the time index. We remark that in this case,  $k^* = \delta_{\underline{x}_s} / \delta_{\underline{w}_s}$ , where  $\delta_{\underline{x}_s}$  and  $\delta_{\underline{w}_s}$  denote the time units of  $\underline{x}_{s,n}^i$  and  $\underline{w}_{s;d}^i$  respectively. For instance, if  $\underline{w}_{s;d}^i$  represents the process of month  $s$ , at the daily temporal level,  $k^* = D_s$ , where  $D_s$  stands for the days of a month  $s$  (i.e., 28, 30 or 31, excluding leap years; similarly, if  $\underline{w}_{s;d}^i$  denotes an hourly process, then  $k^* = D_s \times 24$ ). Nonetheless, for the simulation of daily temporal level, we employ CMARTA model, and aim to reproduce:

- the seasonally varying marginal distribution of  $\underline{w}_{s;n}^i$ , i.e.,  $F_{\underline{w}_s}(w) = F_{\underline{w}_{s;n}}(w)$ ;
- the within-month autocorrelation structure  $\rho_{\underline{w}_s, \tau}^i = \text{Corr}[\underline{w}_{s;d}^i, \underline{w}_{s;d+\tau}^i]$ ;
- the lag-0 cross-correlation coefficients among processes  $\underline{w}_s^i$  and  $\underline{w}_s^j$  for each season  $s$ , i.e.,  $\rho_{\underline{w}_s}^{i,j} = \text{Corr}[\underline{w}_s^i, \underline{w}_s^j]$ .

Provided that the parameters of the individual models have been identified (see Chapter 4 for a general overview, as well as Chapter 5 and 6 for a model-specific description), the simulation procedure starts with generating a realization of the annual process, using the SMARTA model, and subsequently, moves to the monthly and daily level, through the NDA approach. The overall procedure can be organized as follows:

#### **Generation of annual synthetic time series**

Using SMARTA synthesize a  $m$ -dimensional realization of the annual process  $\underline{y}_t$  with  $t = 1, \dots, T$ , where  $T$  denotes the desired length of the time series. The synthesized realization is represented by a  $m \times T$  matrix  $\mathbf{Y}$ , i.e.,

$$\mathbf{Y} = \begin{bmatrix} y_1^1 & \cdots & y_T^1 \\ \vdots & \ddots & \vdots \\ y_1^m & \cdots & y_T^m \end{bmatrix}$$

#### **Generation and adjustment of monthly synthetic time series**

By construction, the realization  $\underline{y}_t$ , fulfils the specifications of the annual level, hence the next step is to generate  $T$  realizations of the monthly process  $\underline{x}_{s,n}$ , each of length 12 (i.e., equal to the number of months) in a way that they reproduce the specifications implied for the monthly time scale, and additionally when aggregated to the annual temporal level they honor the additive property, i.e., each  $y_t^i = \sum_{n=(t-1)12+1}^{12t} x_{s,n}^i$ . Therefore, for each year  $t = 1, \dots, T$ , we employ NDA with SPARTA model as generation mechanism (by setting  $\underline{\xi}_t = \underline{y}_t$  and  $\underline{\omega}_n = \underline{x}_{s,n}$ ), and obtain  $T$  matrices  $\mathbf{X}_1, \dots, \mathbf{X}_t, \dots, \mathbf{X}_T$ , which contain the final adjusted monthly realizations. Each matrix has the form:

$$\mathbf{X}_t = \begin{bmatrix} x_{1,(t-1)12+1}^1 & \cdots & x_{12,t12}^1 \\ \vdots & \ddots & \vdots \\ x_{1,(t-1)12+1}^m & \cdots & x_{12,t12}^m \end{bmatrix}$$

Finally, the matrices are concatenated in  $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_t, \dots, \mathbf{X}_T]$ .

### Generation and adjustment of daily synthetic time series

For the disaggregation of monthly to the daily temporal level, and given the previous matrix organization, it is convenient to refer to the obtained, adjusted, monthly realization with reference in season  $s$  and year  $t$  (not time index  $n$ ), i.e.,  $\mathbf{x}_{s,t}$ , where  $s = 1, \dots, 12$  and  $t = 1, \dots, T$ . For instance, in this notation,  $\mathbf{x}_{3,2}$ , refers to the third month of the second year. At this point we have at our disposal, a realization at the monthly level of length  $12 \times T$ , and seek to generate an equal number of realizations of the daily time scale, each one with length  $D_s$ . Similarly, to the previous level and for the same reasons, we wish the realizations of  $\underline{\mathbf{w}}_{s;d}$  to resemble the specifications of the sub-monthly time scale, and fulfil the additive property, i.e.,  $x_{s,t}^i = \sum_{d=(t-1)D_s+1}^{D_s t} w_{s;d}^i$ . In this vein, for each month  $s = 1, \dots, 12$  and year  $t = 1, \dots, T$ , employ NDA using CMARTA for data generation (by setting  $\xi_t = \mathbf{x}_{s,t}$  and  $\omega_n = \mathbf{w}_{s;d}$ ), and obtain  $12 \times T$  matrices  $\mathbf{W}_{s,t}$ , which contain the final adjusted daily realizations, i.e.,

$$\mathbf{W}_{s,t} = \begin{bmatrix} w_{s;(t-1)D_s+1}^1 & \cdots & w_{s;tD_s}^1 \\ \vdots & \ddots & \vdots \\ w_{s;(t-1)D_s+1}^m & \cdots & w_{s;tD_s}^m \end{bmatrix}$$

Finally, the matrices are concatenated in  $\mathbf{W} = [\mathbf{W}_{1,1}, \dots, \mathbf{W}_{12,1}, \dots, \mathbf{W}_{1,T}, \dots, \mathbf{W}_{12,T}]$ , which contains the complete sequence of the daily realization.

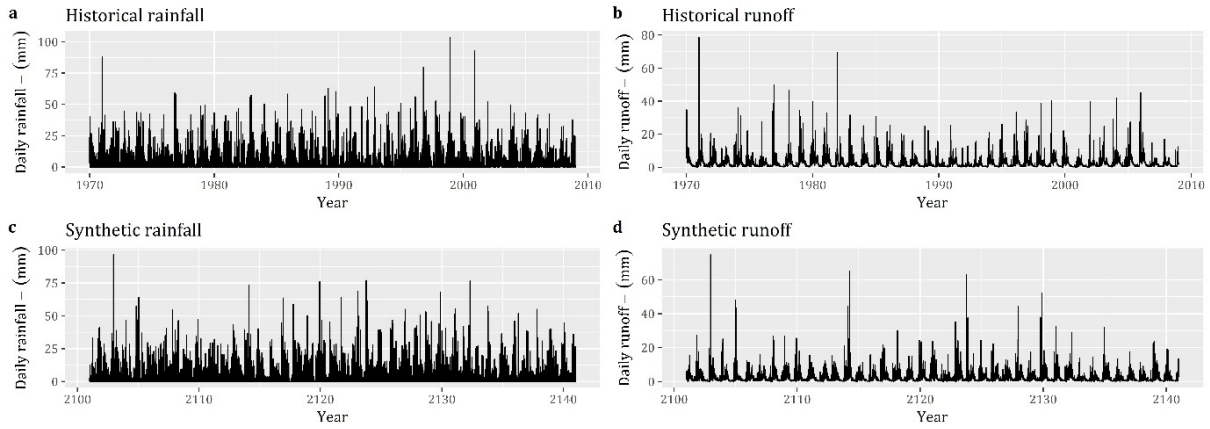
## 7.4 CASE STUDY A: MULTI-TEMPORAL SIMULATION OF DAILY PROCESSES

To assess the performance of the aforementioned three-level configuration scheme, we employed two case studies, one that regards the synthesis of contemporaneous daily rainfall-runoff series at a single location, and another that concerns the generation of intermittent daily rainfall at four locations (presented in Appendix D). In both cases, the evaluation of the model is performed at multiple time scales, by aggregating the generated time series and comparing the empirical, simulated and theoretical (i.e., target) marginal and joint characteristics. Herein we shall describe only the first case study, since the results are similar in both cases.

This case study regards the contemporaneous synthesis of daily rainfall and runoff data, at the river basin of Achelous, Western Greece, upstream of Kremasta dam (Figure 7.2a-b). We note that the runoff series of this dataset has been employed in Tsoukalas et al. [2018a] (section 3.3), to demonstrate the so-called *envelope behavior* of the AR(1) model when combined white noise from Pearson type-III distribution (i.e., Thomas-Fiering approach). Herein the same dataset is employed to demonstrate the three-level configuration scheme, for the synthesis of long daily rainfall-runoff time series (2 000 years; Figure 7.2c-d). It is noted that the units have been converted to mm (from m<sup>3</sup>/s) for convenience in aggregation/disaggregation operations.

Regarding the model parameterization, we employed a theoretical autocorrelation model, i.e., Cauchy-type (CAS; Eq. (5.8)) for describing the auto-dependence structure of the processes, at the annual and daily time scales. It is noted that at daily scale, the parameters of CAS were varied on a monthly basis. Furthermore, the target distribution functions were varied according to the time scale of simulation, the season and the type of processes (i.e., runoff or rainfall). In all cases, the parameters of the distribution functions have been identified on the basis of historical data, using the L-moments method. Particularly, in the case of runoff, we modeled the data using either the three-parameter Log-Normal ( $\mathcal{LN}$ ; Eq. (4.48)), the Generalized

Gamma ( $\mathcal{G}\mathcal{G}$ ; Eq. (5.44)) or the Burr type-XII ( $\mathcal{B}r\mathcal{X}II$ ; Eq. (5.41)) distribution. On the hand, for the daily rainfall process, which is characterized by intermittent behavior, we employed the zero-inflated distribution model of Eq. (4.45), using for the continuous component one of the above distributions.

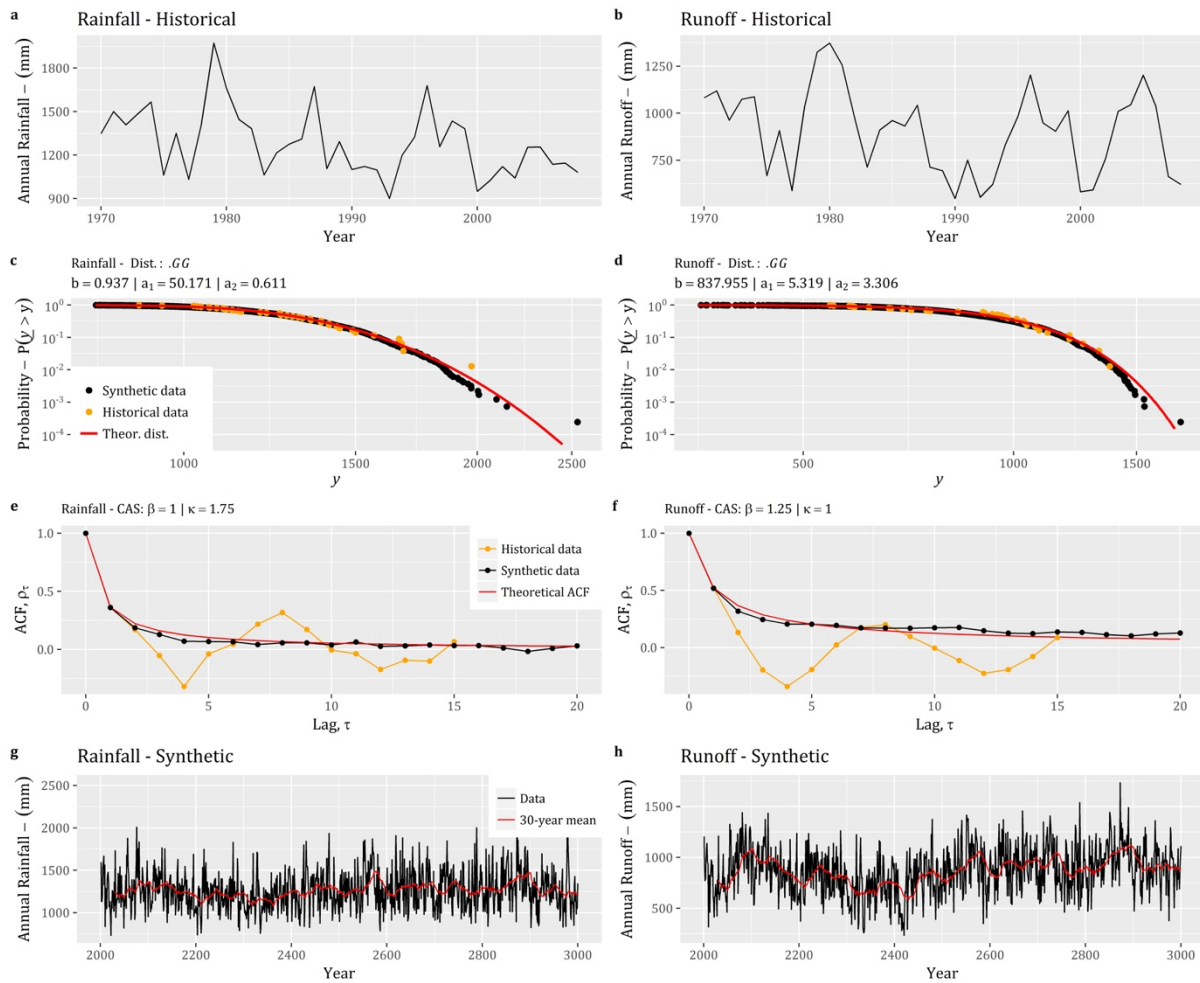


**Figure 7.2** | a-b) Historical daily rainfall-runoff time series (1 January 1970 to 31 December 2008). c-d) Synthetically generated time series (randomly selected window of 40 years).

Starting from the annual temporal level, **Figure 7.3**, summarizes the ability of the highest-level model to preserve both the target distribution function and the autocorrelation structure of each process. Furthermore, the model resembled the lag-0 cross-correlation among the two processes with high accuracy (the historical and simulated values are 0.813 and 0.815 respectively). It is noted that the parameters of CAS have been manually fine-tuned in order to increase the *degree* of annual long-range dependence and stress-test the capabilities of the associated simulation scheme.

**Figure 7.4-Figure 7.5** provides a quick outlook of the results obtained at the monthly time scale, preserving with high accuracy, the empirical L-moments, the seasonality, expressed by means of month-to-month correlation coefficients, as well as the lag-0 cross-correlations.

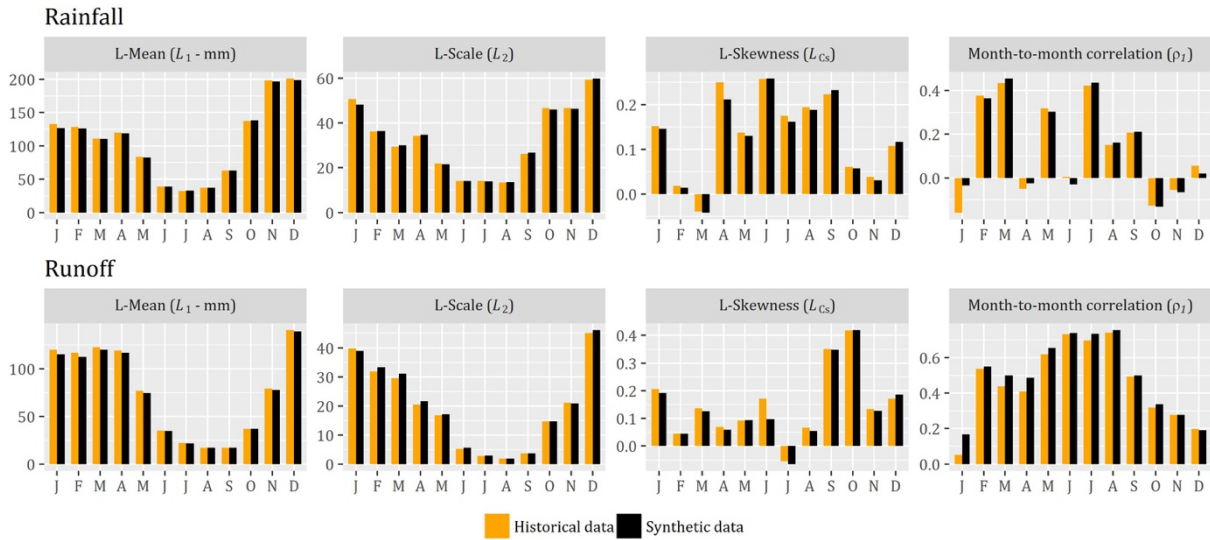
Beyond summary statistics, a more challenging test is the reproduction of the monthly target marginal distributions. **Figure 7.6-Figure 7.7**, compare the empirical distribution of the historical and synthetic data with the target theoretical model (the fitted distribution, as well as its parameters are shown in the title of each sub-plot). In all cases, the model resembled the target distribution with notable accuracy.



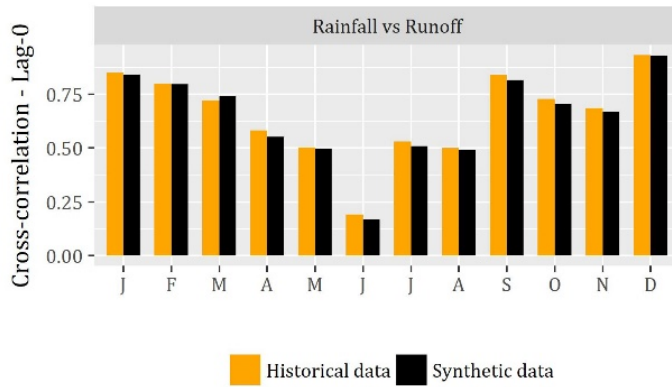
**Figure 7.3** | Rainfall-runoff series: (a-b) Historical annual time series. (c-d) Empirical, simulated and theoretical distribution functions (using the Weibull’s plotting position). (e-f) Empirical, simulated and theoretical ACFs. (g-h) Synthetic annual time series (randomly selected window of 1 000 years).

The previous figures, illustrate the ability of the integrated model, to generate cyclostationary realizations that are also consistent with the specifications of the annual temporal level. As an additional diagnostic, and to test the model for *envelope behavior* we employed scatter plots, and depicted the established dependence patterns. An example is given in **Figure 7.8**, which depicts the lag-1 month-to-month dependence patterns of runoff series. The scheme is relieved from the aforementioned behavior, yet more interestingly, it was found capable of creating a variety of dependence forms, which are also in accordance with the historical ones. The results obtained for other time scales (or rainfall) are similar, hence not shown herein.

7.4 CASE STUDY A: MULTI-TEMPORAL SIMULATION OF DAILY PROCESSES

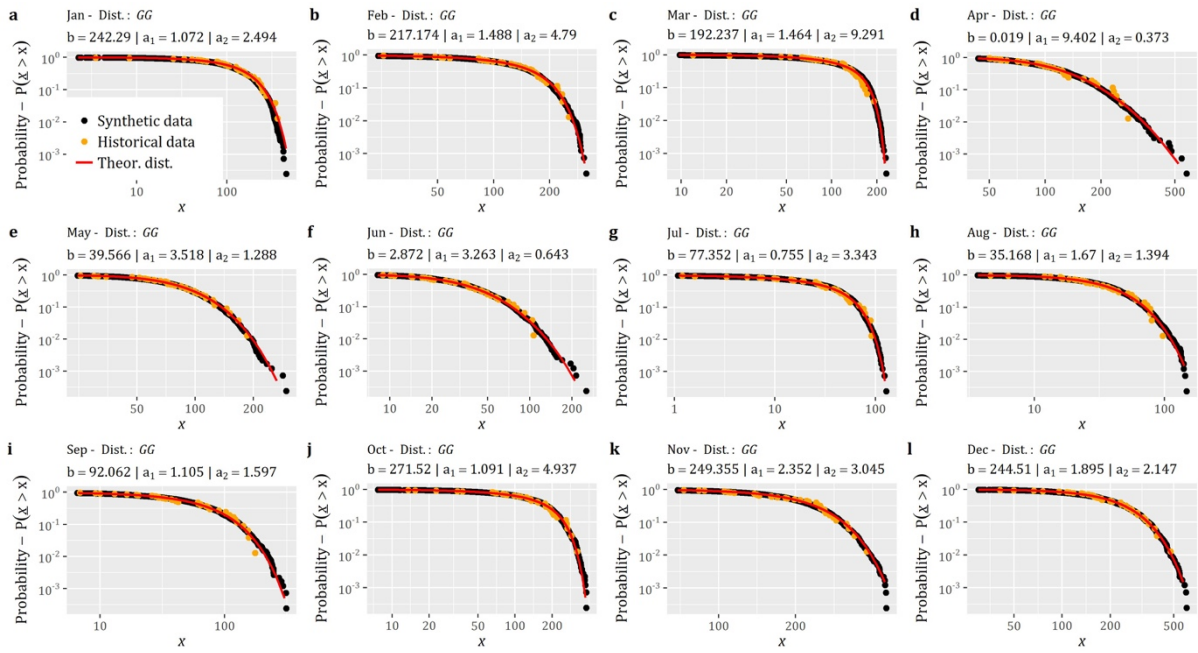


**Figure 7.4** | Comparison of monthly empirical and simulated L-Mean, L-Scale and L-Skewness, as well as historical and simulated lag-1 month-to-month correlations.

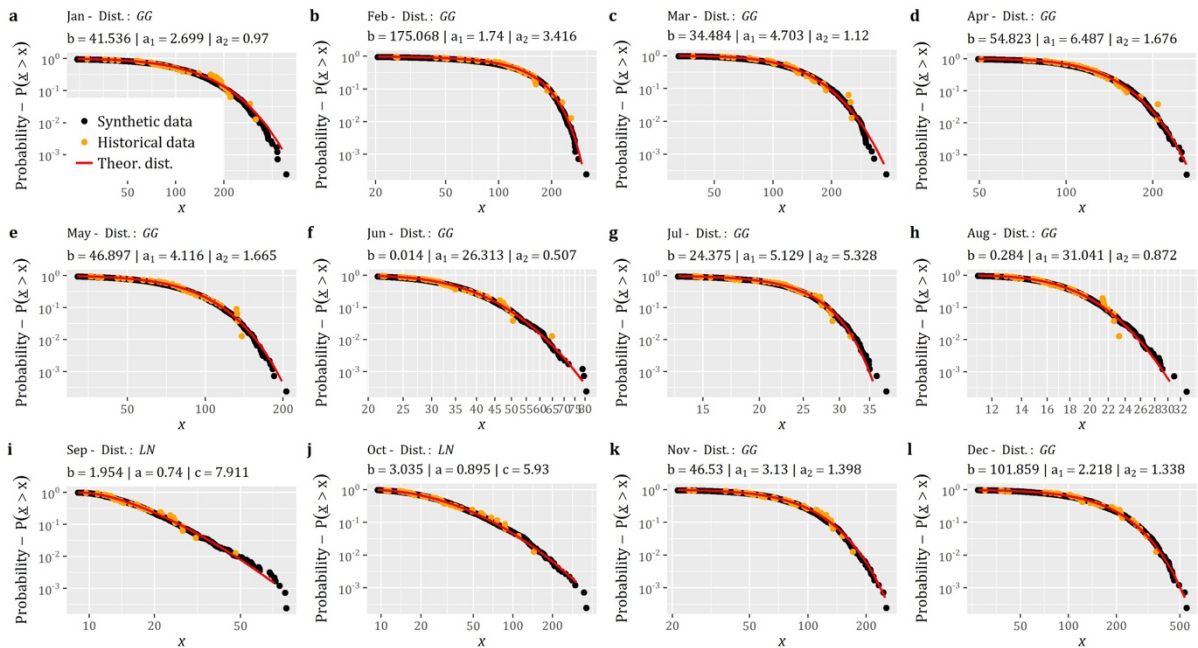


**Figure 7.5** | Comparison of monthly historical and simulated lag-0 cross-correlations.

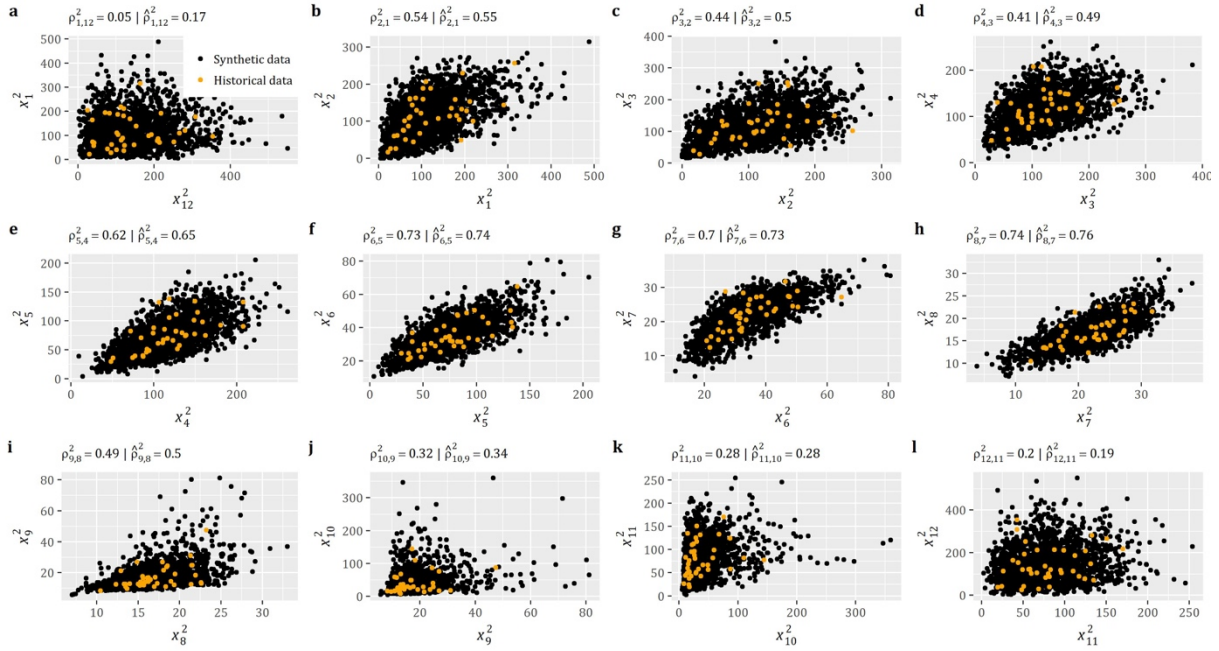
# THE PUZZLE OF MULTI-TEMPORAL STOCHASTIC SIMULATION



**Figure 7.6** | Monthly rainfall - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.



**Figure 7.7** | Monthly runoff - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.



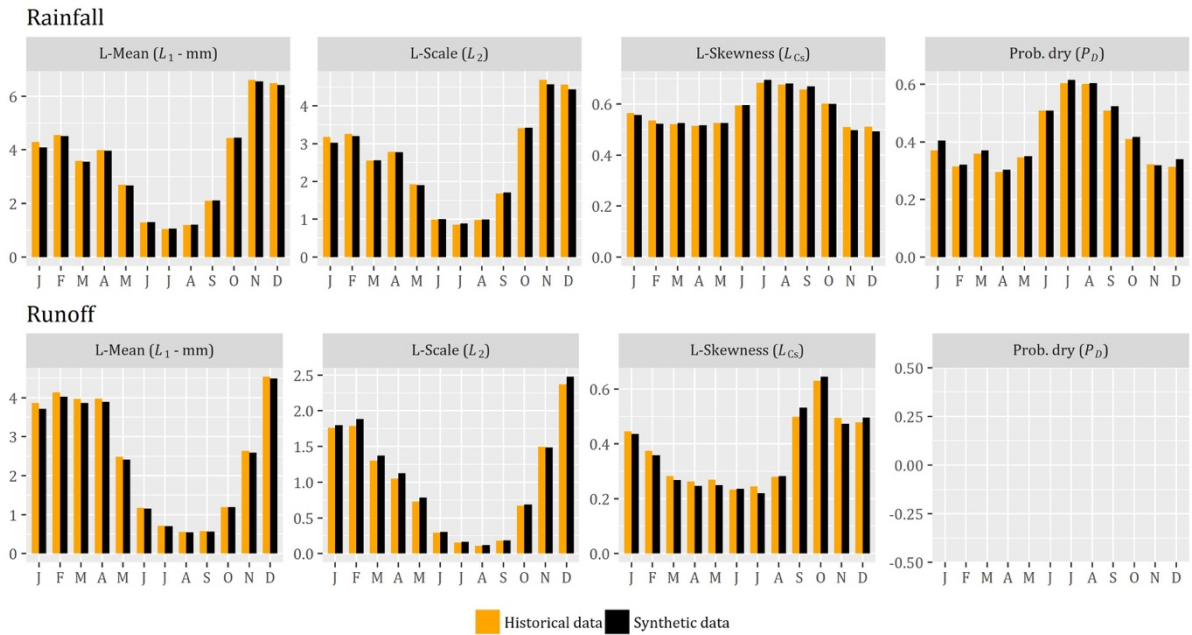
**Figure 7.8** | Monthly runoff (mm) - month-to-month scatter plots of historical and simulated series. The title of each subplot provides the lag-1 month-to-month target ( $\rho_{s,s-1}$ ) and simulated ( $\hat{\rho}_{s,s-1}$ ) correlation coefficients.

Regarding the lowest level of simulation, that is the daily time scale, the comparison among summary statistics of **Figure 7.9** and **Figure 7.10** and, as well as the empirical, simulated and theoretical distribution functions depicted in **Figure 7.11**-**Figure 7.12**, underline the ability of the model to generate consistent realizations with the higher levels, and also preserve the target distribution functions of the daily process, which at this time scale, are characterized by considerably heavier tails. Notice that for daily runoff, and for the months, February to May, we selected the *BrXII* model, which is a heavy-tailed distribution with power-type tail. Recall, that the  $r^{th}$ -moment of the *BrXII* exist only if  $a_1 a_2 < r$ . Remarkably, the scheme accurately simulated even February's daily runoff, which is characterized by  $a_1 a_2 < 2.90$ ; implying that it only has finite mean and variance.

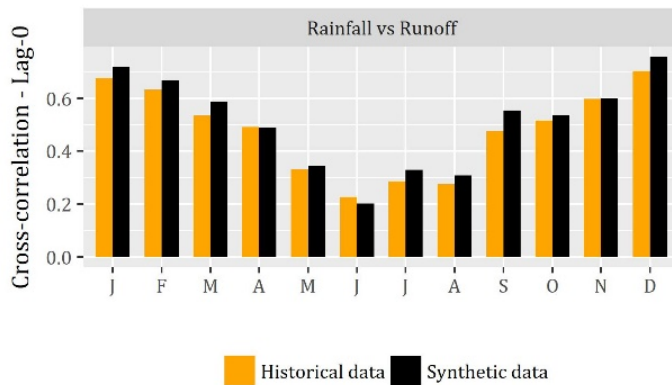
Furthermore to this, **Figure 7.13**-**Figure 7.14** depict a monthly-based comparison of the empirical, simulated and theoretical autocorrelation function (ACF) of the daily process, which in most cases deviates from the typical AR(1) ACF, that most daily stochastic models are capable of simulating. Inspection of this figure, reveals that the integrated model can resemble the theoretical auto-dependence structure with high precision. This result stems from the combination, within NDA, of two modelling components; the CMARTA and the use of theoretical autocorrelation structures (i.e., CAS).

Further the above analysis, which concerned the three characteristic time scales of simulation, in order to investigate the performance of the model at the intermediate time scales between daily and monthly, we aggregated, on a monthly basis, the generated daily series at several scales  $k \in \{2, \dots, D_s\}$  and estimated the L-scale ( $L_2^{(k)}$ ), L-Skewness ( $L_{CS}^{(k)}$ ) coefficients, as well as probability dry ( $P_D^{(k)}$ ) at scale  $k$ . The latter analysis is presented in Appendix **D.1**. It is remarked that the intermediate time scales (i.e.,  $k \neq \{1, D_s\}$ ) are not explicitly modelled neither by the three-level scheme or NDA, hence the arguably good results can be attributed to

the accurate simulation of the process at daily and monthly time scales. Similar results were obtained for the case studies of Appendix D.2 and section 7.5.



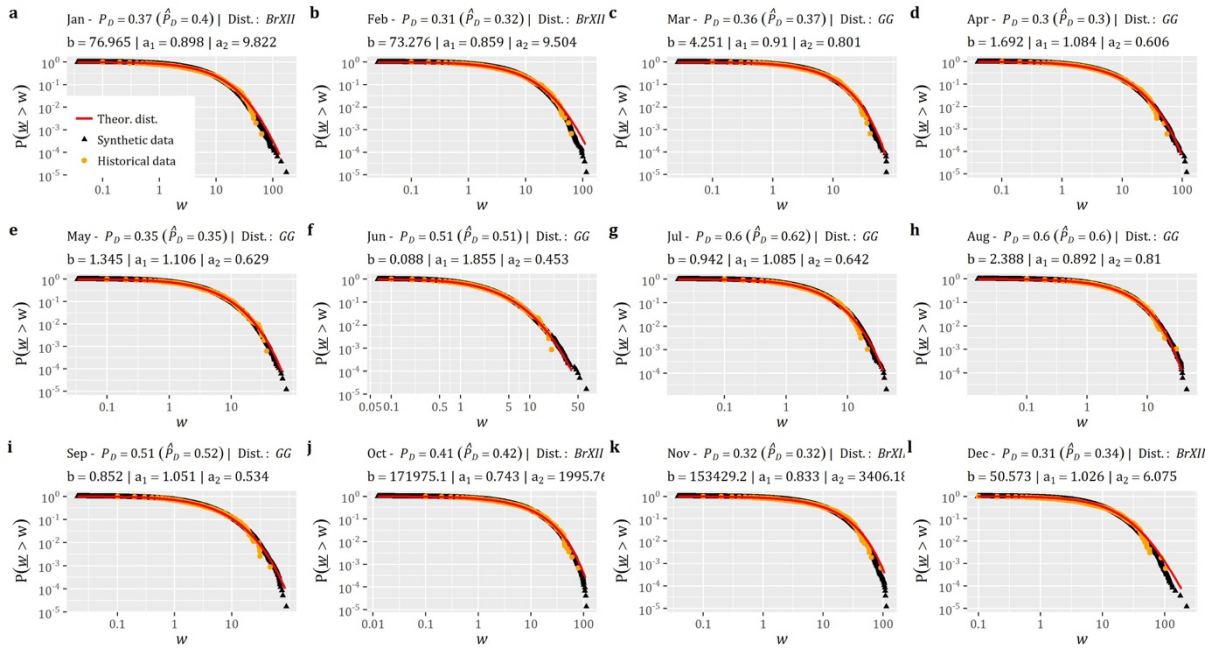
**Figure 7.9** | Comparison of daily empirical and simulated L-Mean, L-Scale, L-Skewness, as well as probability dry.



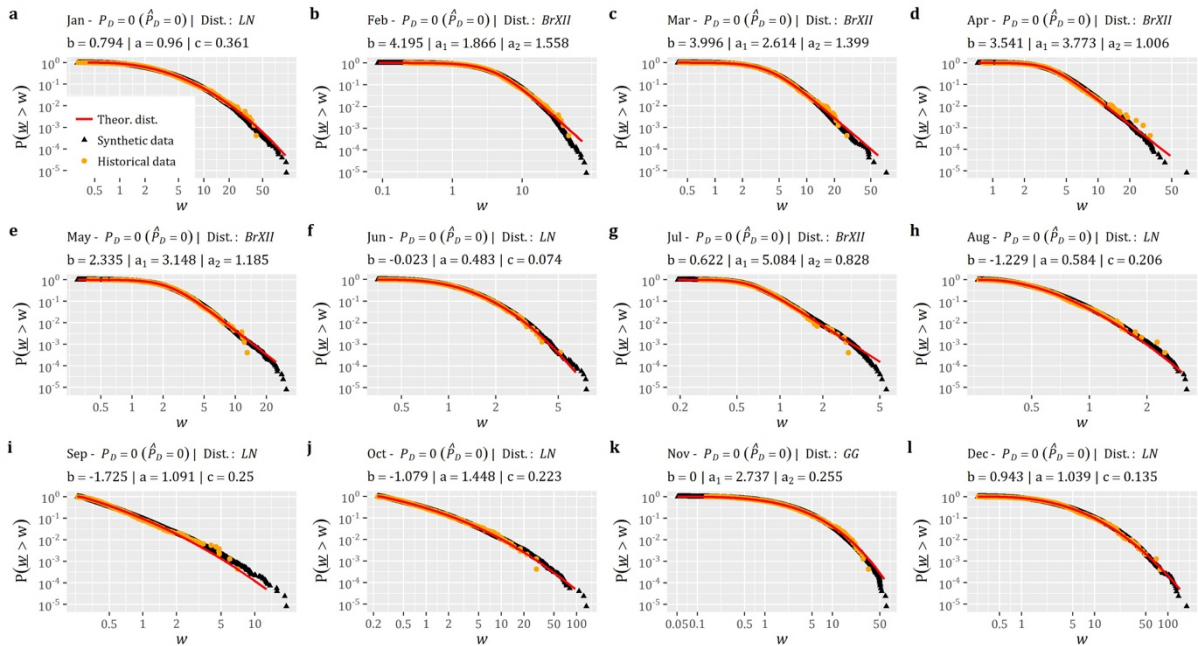
**Figure 7.10** | Comparison of daily historical and simulated lag-0 cross-correlations.



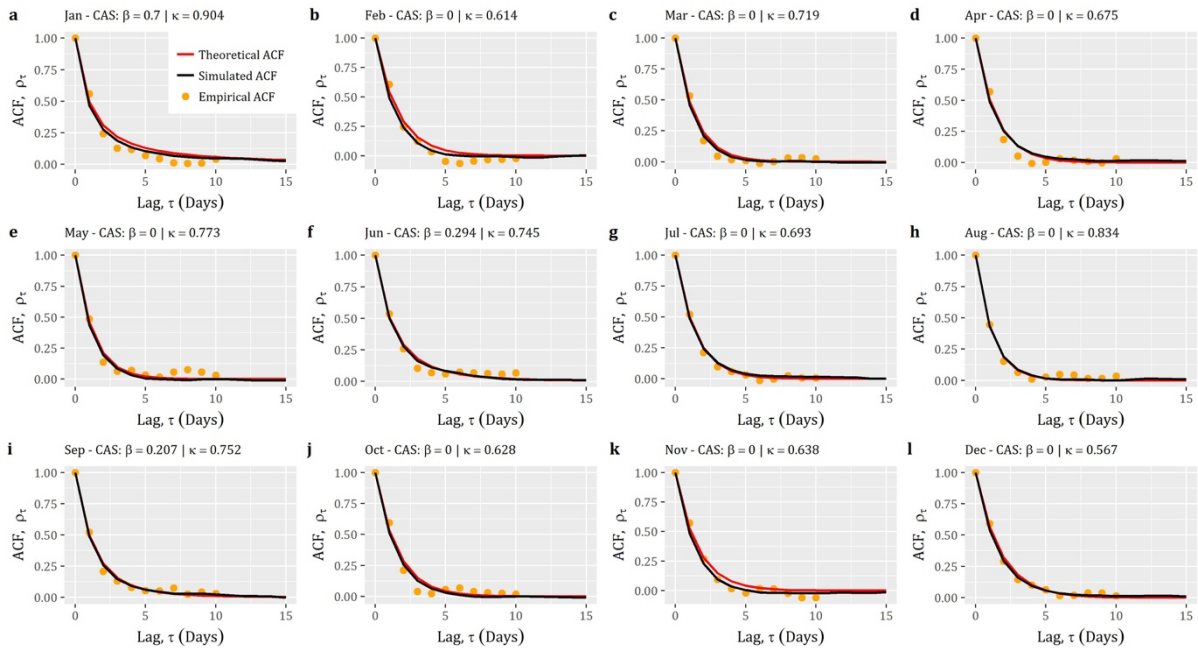
## 7.4 CASE STUDY A: MULTI-TEMPORAL SIMULATION OF DAILY PROCESSES



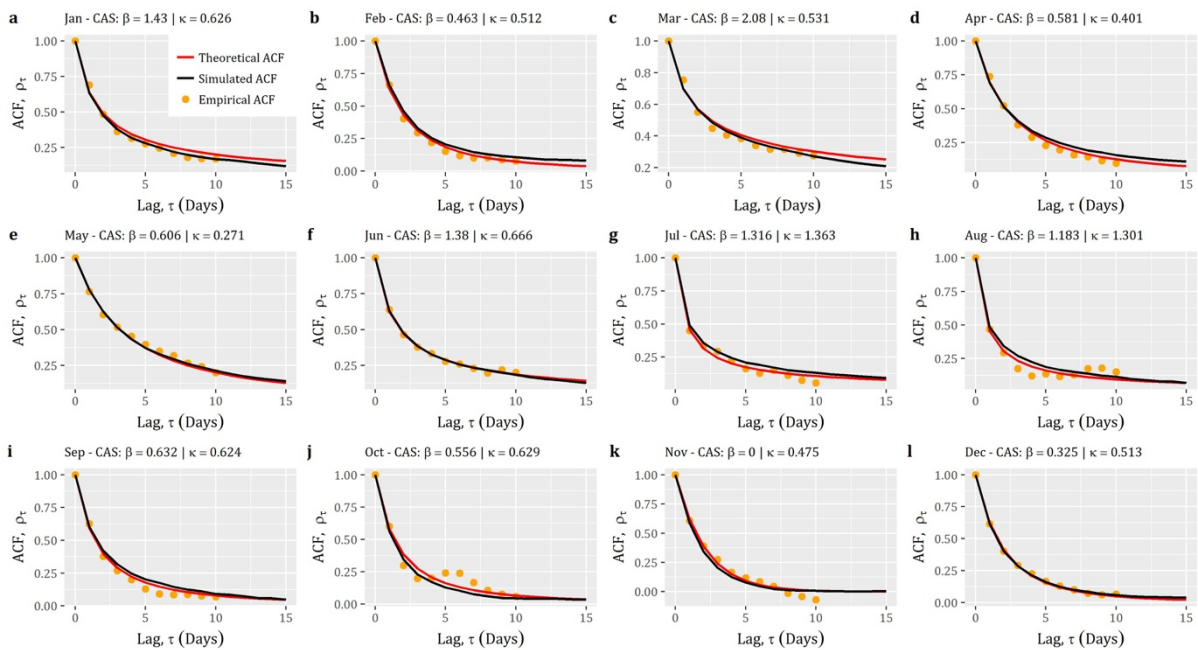
**Figure 7.11** | Daily non-zero rainfall - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.



**Figure 7.12** | Daily non-zero runoff - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.



**Figure 7.13** | Daily rainfall - monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF); the parameters of CAS are given on the title of each subplot.



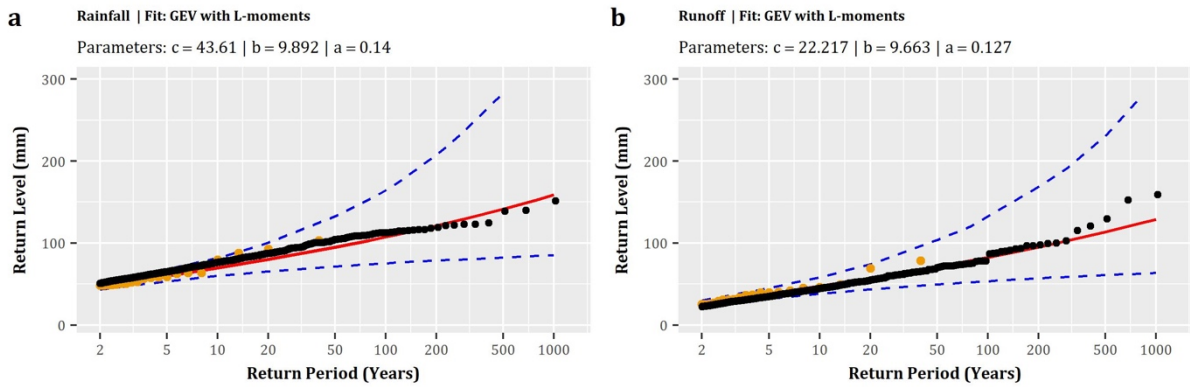
**Figure 7.14** | Daily runoff - monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF); the parameters of CAS are given on the title of each subplot.

A final assessment of model’s performance concerns its capabilities regarding the reproduction of the daily extremes. It is reminded that the distribution of the extremes is not explicitly modelled by the method. **Figure 7.15** depicts the empirical and simulated daily annual maxima, as well as the fitted (using the L-moments method), to the historical data, Generalized Extreme Value ( $\mathcal{GEV}$ ) distribution. The CDF of  $\mathcal{GEV}$  is given by,

$$F_{\mathcal{GEV}}(x; c, b, a) = \begin{cases} \exp\left(-\left(1 + a\frac{x-c}{b}\right)^{-\frac{1}{a}}\right), & a \neq 0 \\ \exp\left(-\exp\left(-\frac{x-c}{b}\right)\right), & a = 0 \end{cases}, \quad (7.6)$$

where  $a, c \in \mathbb{R}$  and  $b > 0$  are shape, location and scale parameters respectively.  $\mathcal{GEV}$  encompasses three distributions, the Fréchet ( $a > 0$  with  $\underline{x} \in [c - b/a, +\infty)$ ), the Gumbel ( $a = 0$  with  $\underline{x} \in (-\infty, +\infty)$ ) and the reversed Weibull ( $a < 0$  with  $\underline{x} \in (-\infty, c - b/a]$ ); the latter case is not considered herein, since it regards upper bounded random variables.

Inspection of **Figure 7.15**, shows that, in both cases, the model managed to resemble the distributional form of the identified  $\mathcal{GEV}$  distribution, which in both cases, is characterized by an arguably heavy-tailed behavior, expressed through the Fréchet distribution (since  $a > 0$ ). In our opinion, this behavior can be attributed to the concise reproduction of the distributions at the daily time scale, which in several instances was modelled using either, the power-type  $\mathcal{BrXII}$  or the  $\mathcal{LN}$  distribution.



**Figure 7.15** | Empirical (●) and simulated (●) daily annual rainfall-runoff maxima, as a function of the return period. The solid red line (—) depicts the fitted to historical data Generalized Extreme Value ( $\mathcal{GEV}$ ) distribution (parameters: location ( $c$ ), scale ( $b$ ) and shape ( $a$ )). The dashed blue line (---) represents the 95% confidence intervals (estimated using the parametric bootstrap method).

## 7.5 CASE STUDY B: DISAGGREGATION OF DAILY RAINFALL TO HOURLY SCALE

To demonstrate the flexibility provided by NDA, as well the potential to extend the three-level scheme of the previous section to even lower temporal levels, we now provide a two-level configuration for disaggregating a univariate daily sequence to the hourly scale.

Particularly, we employ an hourly rainfall dataset from the German Weather Service (Deutscher Wetterdienst; DWD). The historical hourly time series (**Figure 7.16b**), extend over the period 01/09/1995 – 31/12/2017 and concern data at Oberstdorf (station ID: 3730).

In this example, we do not aim to generate synthetic data that represents the actual process across multiple time scales of interest (such as in case study A). In contrast, our goal is to provide a synthetic hourly realization, under the following requirements:

- the synthetic data at the hourly scale reproduces the probabilistic and stochastic properties of the historical sample;

- the additive property is preserved between the aggregated hourly ( $k = 1$ ) synthetic data and the corresponding historical ones (i.e.,  $k^* = 24$ ; **Figure 7.16a**).

By definition, in disaggregation problems, the synthetic sequence has the same length with the given data.

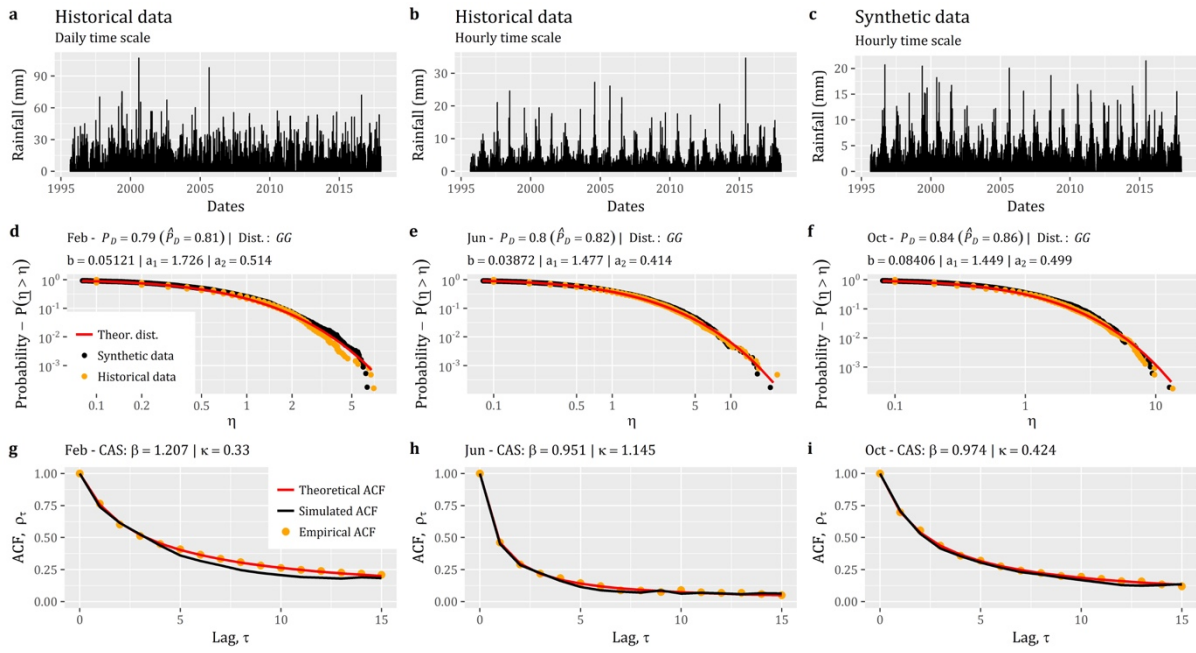
To cope with the effect of seasonality, we employ the typical assumption for fine-time scale rainfall processes (e.g., daily, hourly or finer), that of cyclical stationarity with annual period and monthly sub-period (see also section ). Assuming that the sequence  $\{w_{s,d}\}$  denotes the observed daily records for month  $s$ , we wish to simulate an hourly process, say,  $\{\underline{\eta}_{s,h}\}_{h \in \mathbb{Z}^>}$ , which is also considered stationary within the month  $s$ . This implies that the distribution function (i.e.,  $F_{\underline{\eta}_s}$ ) of the process, as well as its auto-correlation structure, i.e.,  $\rho_{\underline{\eta}_s;\tau} = \text{Corr}[\underline{\eta}_{s,h}, \underline{\eta}_{s,h+\tau}]$  remain invariant within the month  $s$ . Furthermore, to account for temporal consistency we impose the requirement of generating realizations of the process  $\{\underline{\eta}_{s,h}\}$  constrained by,  $w_{s,d} = \bar{w}_{s,d}$ , where  $\bar{w}_{s,d} := \eta_{s;l}^{(24)} = \sum_{h=(l-1)24+1}^{l24} \eta_{s,h}$  (analogous to Eq. (7.2)).

In order to simulate the hourly rainfall, we employ as generation mechanism the univariate version of CMARTA, which is known as ARTA [*Cario and Nelson, 1996*]. We remind that this model uses as an auxiliary Gp a Gaussian AR process (see section 5.4.3). The generation scheme is employed on a monthly basis, since the hourly process properties are reasonably considered seasonally varying.

Regarding the parameterization of ARTA, the marginal distribution of hourly rainfall of each month is modelled using the zero-inflated model of Eq. (4.45). In this case, for the continuous part we fitted (using L-moments) the  $\mathcal{GG}$  distribution (the parameters of the model are shown in **Figure 7.16d-e**). For the autocorrelation structure of the hourly rainfall, we fitted monthly-varying CAS models (i.e., Eq. (7.1)) to the corresponding empirical autocorrelation coefficients (red line in **Figure 7.16g-i**; including the identified parameters). Eventually, each individual hourly process is modeled using five parameters (three for the marginal distribution and two for the autocorrelation structure).

We note that, since our main purpose is demonstration, both the discrete and continuous part are estimated from historical data, yet it is noted that, in alternative situations, one could employ, regional information and/or rainfall's scaling properties.

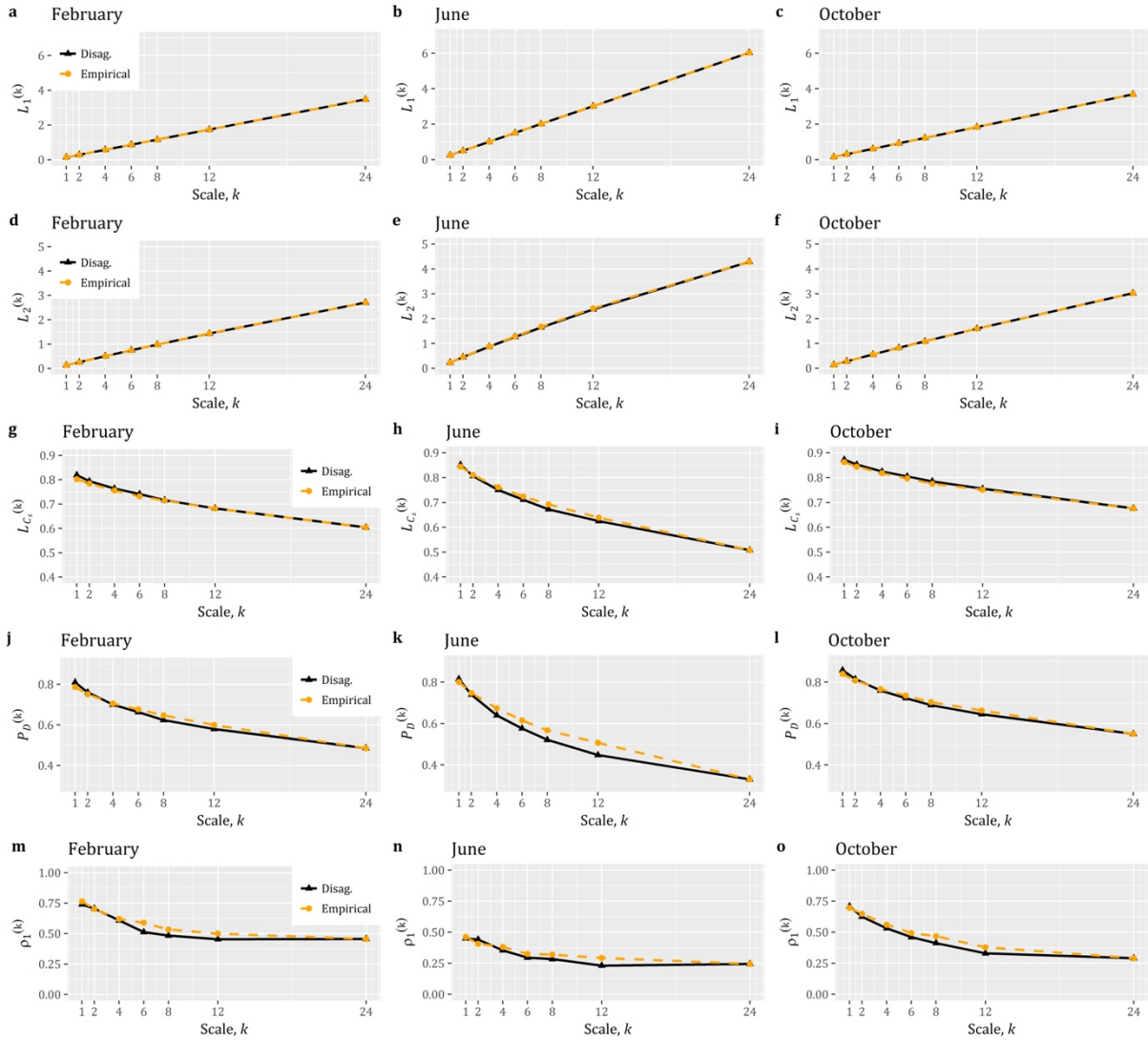
The results from this study are essentially identical for each month (the complete synthesized series is shown in **Figure 7.16c**), hence herein we shall present the results obtained from disaggregating daily rainfall to hourly for three months (i.e., February, June and October). As shown by **Figure 7.16d-e**, and **Figure 7.16g-i**, the model resembled the target distributions and autocorrelation structures respectively, with high precision. Similarly good performance is achieved for the rest of months (see Appendix **D.3**).



**Figure 7.16** | Historical a) daily and b) hourly rainfall series. c) Synthetic (disaggregated) hourly rainfall realization. d-f) Comparison of distribution function of non-zero amounts for hourly historical and disaggregated series for February, June and October respectively (the fitted theoretical model is shown with red line). g-i) Comparison of autocorrelation function (ACF) for hourly historical and disaggregated series for February, June and October respectively (the fitted theoretical model is shown with red line).

Furthermore, to investigate the behavior of the model at the intermediate time scales (i.e.,  $1 < k < 24$ ), in **Figure 7.17**, we depict for both the historical and synthetic series, the L-mean ( $L_1^{(k)}$ ), L-scale ( $L_2^{(k)}$ ), L-skewness ( $L_{CS}^{(k)}$ ), probability dry ( $p_D^{(k)}$ ) and lag-1 autocorrelation coefficient ( $\rho_1^{(k)}$ ), as a function of aggregation level  $k$  (estimated by aggregating the hourly sequences at the corresponding scale  $k$ ). Inspection of these plots reveals the potential of the approach to preserve the empirical scaling properties of rainfall, without requiring the use cascading techniques and direct simulation of rainfall at the intermediate temporal levels.

In our view, apart from the arguably good results of this study, the most important finding is the validation of the modular and scale-free character of NDA, which make it suitable for a wide range of hydrological stochastic simulation problems. Depending on the problem's needs, NDA can be easily applied, by making minimum adjustments or interventions on the algorithm of section 7.2.2.



**Figure 7.17** | Comparison of empirical and disaggregated, a-c) L-mean ( $L_1^{(k)}$ ), d-f) L-scale ( $L_2^{(k)}$ ), g-i) L-skewness ( $L_{CS}^{(k)}$ ), j-l) probability dry ( $p_0^{(k)}$ ) and m-o) lag-1 autocorrelation coefficient ( $\rho_1^{(k)}$ ), as a function of aggregation scale  $k$ , for February, June and October.

## 7.6 SUMMARY

In order to address the puzzle of multi-temporal simulation of hydrometeorological processes, we developed a puzzle-type approach, employing chain implementation of a novel generation procedure, called Nataf-based Disaggregation to Anything (NDA). This is built upon recent advances in stochastics by means of Nataf-based models (Chapter 4-6), coupled with the concepts of repetitive sampling and adjusting [e.g., *Harms and Campbell, 1967; Koutsoyiannis and Manetas, 1996; Koutsoyiannis, 2001*].

This coupling allows taking advantage of the primary ability of Nataf-based models to represent stationary processes that exhibit any distribution and any correlation structure. The recent extension of Nataf-based models to simulate cyclostationary as well as multivariate processes, offered the essential *generality* to handle challenging single-scale hydrometeorological simulation problems.

However, as widely discussed, the reproduction of a target probabilistic and stochastic behavior at a single temporal scale does not guarantee similarly consistent performance at higher

temporal scales. In this Chapter, the issue of *consistency across any pair of scales* is handled via the NDA approach (section 7.2), while the general puzzle-type framework (section 7.3) enables the transition to multi-scale simulations. We remind that NDA uses Nataf-based models at two independent scales as underlying data generators, and coupling mechanisms to adjust the lower-level data to the higher one.

The above approach ensures significant *flexibility*, since it allows establishing any configuration of scale-consistent simulators, through pairwise link of NDAs. This flexibility and the advantages of NDA itself have been mainly revealed by configuring a multivariate simulation scheme (section and 7.3.1) that reproduces the probabilistic and stochastic properties of the processes of interest at three characteristic temporal scales (i.e., annual, monthly and daily). In this configuration, we integrated different Nataf-based models for each scale, i.e., SMARTA [Tsoukalas et al., 2018d] for the annual, SPARTA [Tsoukalas et al., 2017a, 2018e] for the monthly, and CMARTA (Chapter 5) for the daily one.

The multi-temporal simulation capabilities of the integrated scheme were evaluated on basis of two simulation studies, one that regarded the generation of rainfall and runoff synthetic time series at a single location (section 7.4), and another that involves the synthesis of daily rainfall at four locations (Appendix D). As showed, the model reproduced with accuracy the characteristics of the underlying hydrometeorological processes, which exhibit substantial differences among processes and across scales and seasons. Key requirements in these studies were:

- the reproduction of a wide range of *target* distribution functions, varying across processes, scales and seasons;
- the simultaneous simulation of intermittent and/or continuous processes (e.g., daily rainfall and runoff), exhibiting significant correlations;
- the preservation of *target* short-term and long-term auto-dependence structures, at the annual scale, as well as the daily scale, on seasonal basis;
- the preservation of *target* season-to-season correlations at the monthly scale;
- the preservation of *target* lag-0 cross-correlations at all scales.

One can observe that in the above *bucket list* we make repeated use of term *target*, in order to highlight the multidimensional role of the user. Actually, before employing simulations, there are several critical modelling decisions to make, regarding the assignment of suitable distribution functions and correlation structures to the processes of interest (this also involves the selection of time scales to represent, thus the configuration of the puzzle). This flexibility may offer significant advantages. For instance, in this specific study, the careful selection of the daily distribution models resulted to resembling the heavy-tailed behavior of the observed daily extremes. We remind that the reproduction of extremes was not set as explicit requirement by the model, thus making this surprisingly outcome a promising topic for further research.

The model performance at even finer temporal scales (i.e., hourly) was demonstrated through a disaggregation example (section 7.5), where we employed NDA for the synthesis of hourly rainfall realizations that are consistent with the observed daily data. Similarly to the first study, the model faithfully reproduced the target behavior of the hourly process, simultaneously ensuring consistency with the daily scale. Moreover, it reproduced with accuracy important statistical properties of rainfall (expressed in terms of L-moments) at intermediate scales.

Above all, this study highlighted the scale-free character of NDA, as well as its ability to handle hydrological disaggregation problems.

Arguably, the potential applications of our puzzle-type approach extend beyond the realm of hydrometeorological time series generation (or disaggregation). Essentially, it is a general-purpose stochastic simulation scheme. Depending on the synthesis of the puzzle pieces (i.e., chain of NDAs), as well as the underlying decisions of each NDA (in terms of target marginal distributions and correlation structures), it is possible to apply the method for the simulation of a widely extended range of processes, geophysical and socioeconomic.

Beyond simulation, other applications of NDA may concern downscaling or disaggregation problems, which requires a) replacing the corresponding higher-level simulation model with the realizations provided by global or regional climate models, and b) identifying the marginal and stochastic properties of the lower-level model, using, e.g., in-situ gauging stations, regional information, and/or scaling laws.

Eventually, the proposed approach can be employed within broader Monte Carlo experiments, to provide long synthetic input data to deterministic simulation models. Given that the type and number of processes to simulate, as well as their temporal resolution, is dictated by the deterministic model, a major computational challenge arises. In particular, the repetitive sampling within NDA imposes a bottleneck, when applied to high-dimensional multivariate problems and/or long-term simulations at fine time scales. Potential remediation to this technical problem may be the use of parallel computing or the model implementation in low-level programming languages.

Regarding the modelling framework per se, potential future research may focus on two interesting aspects that have been revealed in the two case studies. The first involves the reproduction of extremes within synthetic data, while the second is the validation of the model behavior at intermediate time scales, on the basis of additional simulation studies, using multiple configurations at several time scales.

As a closure, this Chapter, by building-upon, as well as by merging the new developments of Chapter 4-6, into an integrated scheme, concludes the contributions of this Thesis to the stochastic modelling and simulation of hydrometeorological processes. The developed models and simulation schemes (Chapter 4-7) can overcome many of the limitations encountered in other state-of-the-art methods (see the review of section 2.3 and Chapter 3), and provide the means for a more accurate and realistic representation of hydrometeorological processes (hence input uncertainty).

Interesting research topics for future research regard, a) the investigation of NDA performance, in even lower (than hourly) temporal levels (e.g., disaggregation of hourly rainfall to 1-minute series), b) the exploration of the effect of employing different adjusting procedures, and c) the handling of the computational challenge (imposed by repetitive sampling) that arises in high-dimensional multivariate problems (e.g., using parallel computing or implementation in low-level programming languages).

As mentioned earlier, and often due to the typical size of historical data, which is not (neither will ever be) sufficient to extract safe conclusions about the long-term performance of a system, these time series can (and should) be used as input in a variety of risk-related water-system studies to represent the input (hydrometeorological) uncertainty, and it is anticipated to improve the quality of their outcomes, due to more accurate representation of the input processes. However, as discussed in section 1.3, the use of long stochastic inputs (regardless



the data generation model) in combination with simulation models and/or optimization techniques unwillingly pose a barrier in the practical application in simulation-optimization frameworks, since the required computational effort is increased by orders of magnitude. The following two Chapters, aim to address this important issue by developing suitable methodologies and algorithms, that can address challenging optimization problems at a fraction of time that is required by other state-of-the-art methods (e.g., evolutionary algorithms).

## MULTI-OBJECTIVE OPTIMIZATION ON A BUDGET: EXPLORING SURROGATE MODELLING FOR ROBUST MULTI-RESERVOIR RULES GENERATION UNDER HYDROLOGICAL UNCERTAINTY \*

---

### PREAMBLE

Next research steps (Chapter 8 and 9) regard the practical implementation of the new developments in modelling and simulation of hydrometeorological processes (Chapter 4-7), in uncertainty-aware water-system optimization problems (i.e., simulation-optimization frameworks driven by stochastic inputs). This challenge is related to the excessive computational budget imposed by both the use of long synthetic time series to represent the input uncertainty, and the use of objective functions that entail time *expensive* simulation models; which become the *norm* since the requirements for more detailed (hence *expensive*) models are increasing. Particularly, this Chapter considers the problem of handling time *expensive* multi-objective problems on a *budget*, under the prism of developing long term operation rules for multi-reservoir systems. This is a complicated task due to the number of decision variables, the non-linearity of system dynamics, and the computational effort required, which imposes barriers to the exploration of the solution space. These challenges are addressed by (a) employing a parsimonious multi-objective parameterization-simulation-optimization (PSO) framework, which incorporates hydrological uncertainty through stochastic simulation and allows the use of probabilistic objective functions and (b) by investigating the potential of multi-objective surrogate-based optimization (MOSBO) to significantly reduce the resulting computational effort. Three MOSBO algorithms are compared against two multi-objective evolutionary algorithms. Results suggest that MOSBOs are indeed able to provide robust, uncertainty-aware operation rules much faster, without significant loss of neither the generality of evolutionary algorithms nor of the knowledge embedded in domain-specific models.

This Chapter is structured as follows: section 8.1 introduces the problem, while section 8.2 and 8.3 present the overall methodology and the study area respectively. Next, section 8.4 and 8.5 regard the benchmarking and the experimental setup of the algorithms' performance respectively. Section 8.6 present the key results and findings, and finally section 8.7 concludes the Chapter.

---

\* Based on:

Tsoukalas, I., and C. Makropoulos (2015b), Multiobjective optimisation on a budget: Exploring surrogate modelling for robust multi-reservoir rules generation under hydrological uncertainty, *Environ. Model. Softw.*, 69, 396–413, doi:10.1016/j.envsoft.2014.09.023.

## 8.1 INTRODUCTION

Water reservoirs and the associated hydrosystems often serve multiple purposes including, *inter alia*, flood control, irrigation, water supply, restoration, navigation, recreation, hydropower generation, etc. The operation of a reservoir system involves a complex decision making process that strives to balance many variables and (often conflicting) objectives, aiming mostly at the quantification (and if possible minimization) of risk and uncertainty [Oliveira and Loucks, 1997]. The conflicting nature of the different objectives makes this decision process a classic multiobjective optimization problem as demonstrated in Haimes, [1977], seeking to derive optimal management strategies against various performance measures such as reliability of supply, cost minimization and environmental protection. Optimizing this decision process is complicated due to the existence of non-linear and interdependent parameters and processes, [Vink and Schot, 2002].

A common way to address this is the coupling of simulation models and multiobjective evolutionary optimization (MOEA) algorithms [Nicklow et al., 2010], typically yielding a set of efficient (Pareto-optimal) solutions. The solutions can then be used as a negotiation tool for decision makers, through the explicit representation of the objectives' tradeoff [Makropoulos and Butler, 2005] without having to embed *a priori* preferences in the decision process. Extensive reviews of MOEA schemes and their applications in water resources are available in the literature [e.g., Savic and Walters, 1997; Efstratiadis and Koutsoyiannis, 2010; Nicklow et al., 2010; Reed et al., 2013]. The ubiquity of MOEA approaches is justified by the generic nature and global search capabilities of these algorithms [Coello Coello et al., 2007; Zhou et al., 2011], but this choice results in a significant number of iterations needed to reach an adequate approximation of the Pareto front [Brockhoff and Zitzler, 2009]. The number of iterations, the large decision space, the complexity of the fitness landscape and the long simulation time required by the physical models to evaluate each instance of the objective function are major drawbacks when dealing with real-world applications, where computational time is necessarily limited [Maier et al., 2014]. This is especially true in cases, such as hydrosystem optimization, where performance (and hence optimal operation) is significantly affected by long-term hydrological uncertainties, primarily hydrological variability, which can adequately be addressed by using stochastic simulation (see Chapters 4-7, as well as [Koutsoyiannis, 2005b]), i.e., the generation and use of synthetic time series (typically 500-1000 years) whose statistical properties are consistent with historical data, to drive the simulation-optimization process and hence generate risk-based, robust operation rules. As discussed earlier in section 1.1 and 1.3, driving the simulation-optimization process with very long time series affects the computation burden of each and every evaluation of the objective function and drastically increases computation burden.

The current Chapter, builds on a parameterization-simulation-optimization (PSO) framework initially proposed by Koutsoyiannis and Economou [2003] to derive optimal reservoir operation rules. Although the PSO approach has already been implemented in single objective problem formulations [e.g., Nalbantis and Koutsoyiannis, 1997; Koutsoyiannis et al., 2002; Momtahan and Dariane, 2007; Celeste and Billib, 2009], little work has been done up to date on using the PSO with multiple objectives. Here we extend the capabilities of the PSO method towards multi-objective optimization, while incorporating hydrological uncertainty (also advocated in Maier et al. [2014]) into the optimizer using probabilistic objective functions, through stochastic simulation to improve the robustness of the resulting rules. To address the issue of computational practicability that ensues from the adopted stochastic simulation approach, we investigate the use of surrogate-based optimization (SBO) techniques (see section 1.3, as well

as Chapter 9), and specifically we explore the applicability of three multi-objective surrogate-based optimization (MOSBO) algorithms, and compare them against the well-known and extensively used algorithm NSGAII [Deb et al., 2002] as well as the more recently proposed SMS-EMOA [Beume et al., 2007]. An extensive benchmarking exercise is then undertaken to derive conclusions about the effectiveness and the efficiency of the MOSBO algorithms [Razavi et al., 2012a] as well as the robustness of the overall methodology. The proposed methodology is tested in the optimization of the multi-reservoir hydrosystem of Nestos in Northern Greece.

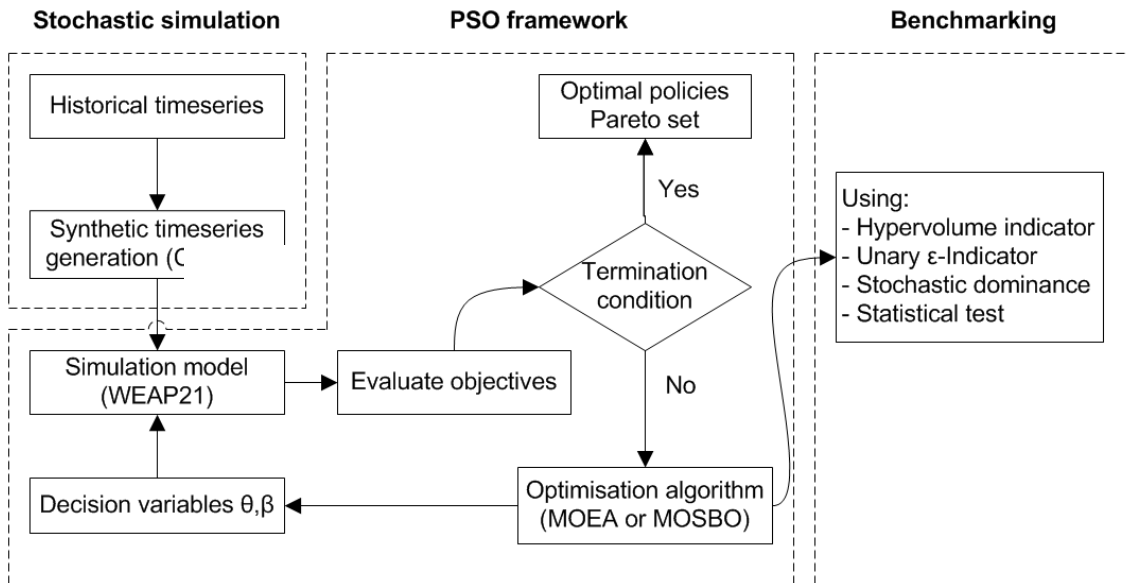
## 8.2 METHODOLOGY

### 8.2.1 Overall conceptual approach

This Chapter introduces a multi-objective extension of the PSO framework for the efficient (in terms of both time and Pareto front generation) optimization of multiple reservoirs' operation maximizing two conflicting objectives related to hydropower generation. The main advantage of the PSO [Koutsoyiannis and Economou, 2003] compared to other similar methods, based on implicit stochastic optimization (ISO) and explicit stochastic optimization (ESO), is on the one hand its parameter-parsimonious character and its ability to incorporate the hydrological uncertainty and on the other hand, the simple operation rules that it provides, which are well suited for real world operators [Celeste and Billib, 2009], thus bridging a gap between theoretical developments and real world applications [Yeh, 1985; Simonovic, 1992; Labadie, 2004; Celeste and Billib, 2009].

The methodological steps of our approach are presented in **Figure 8.1** and summarized below:

- 1) Representation of the main hydrological components (precipitation, evapotranspiration, inflow) through stochastic simulation, using stochastic simulation models that generate long synthetic time series able to capture and reproduce the statistical properties of the historical sample. The synthetic time series are used in step 3 to simulate the hydrosystem.
- 2) Parameterization of the management policy (operation rules) of the reservoirs system via a small number of parameters  $\theta$ ,  $\beta$  (decision variables).
- 3) Simulation of the hydrosystem (in this case with WEAP<sub>21</sub>) using the stochastically generated synthetic time series to drive the simulation and implementing the parameters that define management policy.
- 4) Definition of appropriate objective function(s) that express the desired performance metric(s). In this case, two objective functions related to hydro-energy characteristics have been employed and used in the optimization process (step 5).
- 5) Optimization to derive the best management policies (in view of the two objective functions) using 3 MOSBO and 2 MOEA algorithms (in Matlab and R environment).
- 6) Algorithms' performance benchmarking using appropriate performance indicators and methods.



**Figure 8.1** | Schematic representation of the conceptual approach.

The parameterization of the management policy (operation rules) of the reservoirs system (step 2) requires determining the appropriate decision variables that describe the hydrosystem's function. Typically, these variables express the release from the reservoirs and must be in a simple enough form to be understood by the system operators. Following an Occam's razor approach, we assign for each hydropower reservoir  $i$  a single variable  $\theta_i$  referring to a constant monthly hydropower production target (GWh). During the simulation this target  $\theta_i$  is translated into a minimum required discharge flow that has to pass through the turbines to produce the energy target  $\theta_i$ . The upper bound of variables  $\theta_i$  is set to the maximum theoretical hydropower capacity of the turbines. The lower bound of variables  $\theta_i$  is zero, to account for a null hydropower production target. A second variable, termed the reduction variable  $\beta$ , is applied uniformly to all reservoirs to simulate periods with low hydropower demands such as the period October-April. In order to simulate this demand variability, the constant monthly hydropower production targets  $\theta_i$  are multiplied by  $\beta$  for low hydropower demand periods. The lower bound of  $\beta$  is zero, which when multiplied the constant hydropower target  $\theta_i$  results in zero hydropower target for certain months. The upper bound is one, meaning that hydropower target is constant and equal to  $\theta_i$ . The purpose of  $\beta$  is to reduce the hydropower production targets during specific months (here October-April) and thus increase the hydropower potential available during the rest of the year (here May-Sept) when the demand is higher. Hence, for a hydrosystem consisting of  $n$  hydropower reservoirs, the total number of decision variables equals to  $n + 1$  ( $n$  values for the  $\theta$  parameter plus the uniformly applied variable  $\beta$ ). At every time-step, the energy target is translated to volume of water that passes through the turbines to achieve the specified hydropower target. Since hydropower generation is related to the current volume of the reservoir and the turbine efficiency, the release volume varies every month. This parsimonious parameterization can be easily implemented at most reservoir simulation models (here using WEAP21).

Two probabilistic performance metrics were introduced to shape the objective functions (step 4). The first is the maximization of the monthly guaranteed energy produced from the multi-reservoir system for a given reliability (e.g., reliability  $a = 99\%$ ), for the whole simulation period (i.e., Eq. (8.1)). This measure is also known as firm energy [Hamlet et al., 2002; Larson and Larson, 2007; Efstratiadis et al., 2012]. The second metric is the maximization of the

monthly firm power for a certain period (i.e., Eq. (8.2)) (in our case the period of May-September, where demands for hydropower are higher in Greece). The objective functions are non-linear, non-continuous and non-differentiable due to the probabilistic approach. These two performance metrics define two different operational policies. The first one ensures the stability of energy production for a given reliability level. The second performance metric describes a policy where the guaranteed energy production is maximized only in a period of interest, such as the summer period. The overall probabilistic approach (probabilistic constraints) used to calculate the objective functions accounts for the hydrological uncertainty of the inflows to the reservoirs (which is the main uncertainty in such cases) and thus, we argued, provides more robust, uncertainty-aware operation rules.

**Objective 1:**  $\max\{E^c\}$

$$\text{s. t. : } a = P(E_t > E^c) = \frac{n^s}{n^{tot}} = 99\% \quad (8.1)$$

where,  $E_t = \sum_{i=1}^h E_t^i$ ,  $E_t$  is the hydropower energy time series for given  $\theta_i$  and  $\beta$ , for  $t = 1, \dots, n^{tot}$ , where  $n^{tot}$  is the total number of time-steps for all months and all years in simulation. While,  $i = 1, \dots, h$ ,  $h$  denotes the number of hydropower reservoirs.  $E^c$  denotes the firm energy of the system for the whole year, and  $a$  the desired reliability (=99%), calculated simply by dividing  $n^s$  and  $n^{tot}$ , where  $n^s$  is the number of time-steps that exceed  $E_t > E^c$  and  $n^{tot}$  is the total number of time-steps in  $E_t$ .

**Objective 2:**  $\max\{E^p\}$

$$\text{s. t. : } a = P(E_t^p > E^p) = \frac{n^p}{n^{ptot}} = 99\% \quad (8.2)$$

where,  $E_t^p = \sum_{i=1}^h E_t^{p;i}$ ,  $E_t^p$  is the hydropower energy time series for given  $\theta_i$  and  $\beta$ , for  $t = 1, \dots, n^{ptot}$ , where  $n^{ptot}$  is the total number of time-steps only for high demand months (here May-Sept) during the entire simulation, while  $i = 1, \dots, h$ , where  $h$  denotes the number of hydropower reservoirs.  $E^p$  denotes the firm energy of the system for the high demand period (here May-Sept), and  $a$  is the desired reliability (=99%), calculated simply by dividing  $n^p$  and  $n^{ptot}$ , defined as the number of the time-steps that exceed  $E_t^p > E^p$  and the total number of time-steps in  $E_t^p$  respectively. Note, that parameters  $\theta_i$  and  $\beta$  enter the equation indirectly, through  $E_t$  and  $E_t^p$ .

### 8.2.2 Models and tools

For the simulation of the hydrosystem (in step 3) a widely used water management model, the Water Evaluation and Planning System (WEAP21<sup>3</sup>) was used. To enable a two-way communication between the simulation and optimization models (step 5) a coupling was established through the COM-API<sup>4</sup> function of WEAP21 as discussed in *Tsoukalas and Makropoulos [2013, 2015a]*.

<sup>3</sup> <http://www.weap21.org>

<sup>4</sup> Component Object Model Application Programming Interface

For the stochastic simulation of key hydrological variables (in step 1) synthetic time series of 500 years were generated using the Castalia software [Efstratiadis et al., 2014b]<sup>5</sup>. Castalia is a multivariate stochastic simulation tool developed for the study of monthly hydrological variables such as rainfall, evapotranspiration and inflow. It generates synthetic time series on the basis of historical data, that reproduce the statistical properties (mean value, standard deviation, skewness, lag-1 autocorrelation and lag-0 cross-correlation coefficients) of the observed data sets. Furthermore, Castalia reproduces the long-term persistence of hydrological processes, also known as Hurst-Kolmogorov dynamics [Koutsoyiannis, 2011b], in both annual and inter-annual scale, as well as the periodicity in infra-annual scale and intermittent behavior on daily scales. In this work, 500 years of monthly synthetic rainfall, evapotranspiration and inflow time series were generated, based on historical time series available for the periods 1968-1982 and 1991-1995. Unfortunately, longer historical time series that would result to more representative statistics were not available for this hydrosystem. We note that the use of Castalia in this Chapter is supported by the fact that the focus is in the investigation of surrogate modelling techniques (i.e., MOSBO algorithms in particular) to handle computationally expensive simulation-optimization problems. In an operational context, it could be preferable to employ the stochastic modelling and simulation approach of Chapter 4-7, since it overcomes many of limitations of the current synthetic data generation schemes.

Nonetheless, the synthetic datasets generated were used as inputs to WEAP21 to drive each simulation. Although, the use of synthetic time series leads to more robust solutions that incorporate hydrological uncertainty into the operational policy design it also results in much longer simulation times and hence significantly increases computational effort. For instance, for synthetic monthly time series of 500 years the model needs approximately 90 sec for a single simulation run on a 3.0 GHz Intel Core i5 processor with 4 GB of RAM, running on Windows 8 Operating System. By contrast, a typical multiobjective evolutionary algorithm requires an order of 10 000 iterations to adequately approximate the Pareto front. Consequently, the whole process would last 250 hours, which makes it unrealistic.

To address this issue, we replaced the typical multiobjective evolutionary algorithms with MOSBO algorithms and their performance was assessed. Specifically, we used the following surrogate-based (MOSBO) algorithms (see section 8.2.4): ParEGO [Knowles, 2005], SUMO [Gorissen et al., 2010] and SMS-EGO [Ponweiser et al., 2008]. Their performance has been evaluated for various configurations related to the maximum number of function evaluations allowed. That is because the computational time needed by the MOSBO to construct and search (optimize) the metamodel is minimal. e.g., the total computational time needed to construct a metamodel of 400 samples and then search (optimize) it is less than 2 sec. This effort compared to 90 sec required by the simulation model (WEAP21) is negligible and thus can be ignored. The performance of the 3 MOSBOs has been benchmarked against the well-known NSGAII algorithm [Deb et al., 2002] and the recently proposed SMS-EMOA algorithm [Beume et al., 2007], using performance indicators and methods proposed by Razavi et al. [2012a], Asadzadeh and Tolson [2013] and Matott et al. [2012] (see section 8.4).

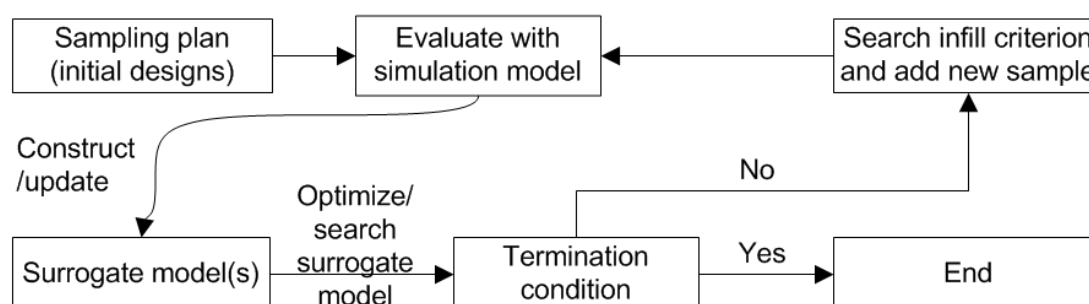
### 8.2.3 Fundamentals of SBO algorithms

The basic concept of SBO methods is to replace most of the expensive simulations with much less expensive surrogate models within the optimization cycle. Examples can be found in

---

<sup>5</sup> See also the R language implementation, i.e., *CastaliaR* package [Tsoukalas et al., 2018c].

literature [e.g., [Queipo et al., 2005](#); [Yan and Minsker, 2006, 2011](#); [Forrester et al., 2008](#); [Knowles and Nakayama, 2008](#); [Kleijnen, 2009](#); [Gorissen et al., 2010](#); [Keating et al., 2010](#); [Santana-Quintero et al., 2010](#); [Jin, 2011](#)]. In SBO, most evaluations are performed with the surrogate model, while the expensive model is used periodically within the optimization process to improve the accuracy of the results. The nature of the true function is not known *a priori*. Therefore, the selection of the most appropriate surrogate model is a challenging and critical task for the optimization process [[Santana-Quintero et al., 2010](#)]. The basic stages of SBO are described in [Figure 8.2](#) below and briefly discussed in the following paragraphs.



**Figure 8.2** | Flowchart of the surrogate-based optimization process.

### 8.2.3.1 Initial sampling plan

SBO attempts to develop a (surrogate/inexpensive) model *approximator* able to capture the response of the expensive model for a limited number of *optimally* selected data points. The first step of the SBO procedure is therefore to select these points, using a sampling plan (also termed initial design) and evaluate them with the expensive model. The sampling plan is usually implemented with Design of Experiments (DoE) methods, which aim to maximize the amount of information gained from a limited number of sample points. Typically, the samples need to spread across the design space in order to capture global trends. DoE methods include Latin Hypercube Sampling (LHS), Orthogonal Array Design (OAD) and Uniform Design (UD) [[Giunta et al., 2003](#)].

### 8.2.3.2 Surrogate models

With the initial points selected, the next step is to construct the appropriate surrogate model to approximate the expensive simulator. Typical surrogate models include polynomials [[Sudret, 2008](#); [Crestaux et al., 2009](#)], Radial Basis Functions (RBFs) [[Mugunthan et al., 2005](#); [Regis and Shoemaker, 2007a](#); [Shoemaker et al., 2007](#)], Artificial Neural Networks (ANNs), Support Vector Machines [SVMs - [Dibike et al., 2001](#); [Zhang et al., 2009](#)] and Kriging [[Sacks et al., 1989](#); [Santner et al., 2003](#)]. ANN, in particular, have been used extensively in water resources research [e.g., [Broad et al., 2005](#); [May et al., 2008](#); [Behzadian et al., 2009](#); [Fu et al., 2010, 2012](#)]. [Khu et al. \[2007\]](#), examined various applications of evolutionary computation based surrogate models to augment or replace the conventional use of numerical simulation and optimization within the context of hydro-informatics. Key issues of surrogate models for multiobjective optimization are presented and discussed in [Razavi et al. \[2012b\]](#). They also reported that surrogate models are not suitable for optimization problems with many decision variables, mainly due to the large search space and the number of samples required to adequately sample the objective space [[Razavi et al., 2012b](#)]. Techniques that are able to handle high dimension problems are presented in [Shan and Wang \[2010\]](#).



In our work, this issue was addressed using the PSO parsimonious approach to keep the number of decision variables fairly small. In the work presented here Kriging is used as the surrogate model of choice. Kriging is based on the idea that a value of a random field at an unobserved location can be statistically interpolated using observations at nearby locations. The method was originally proposed by *Krige [1951]* and applied by *Sacks et al. [1989]* to approximate computer experiments. Kriging is a widespread surrogate model capable of approximating deterministic noise-free data, in our case deterministic computer experiments (simulations), and has adequately performed in challenging tasks [*Jones et al., 1998*]. The mathematical background of Kriging is presented in [*Santner et al., 2003; Forrester et al., 2008*]. Kriging consists of two terms:

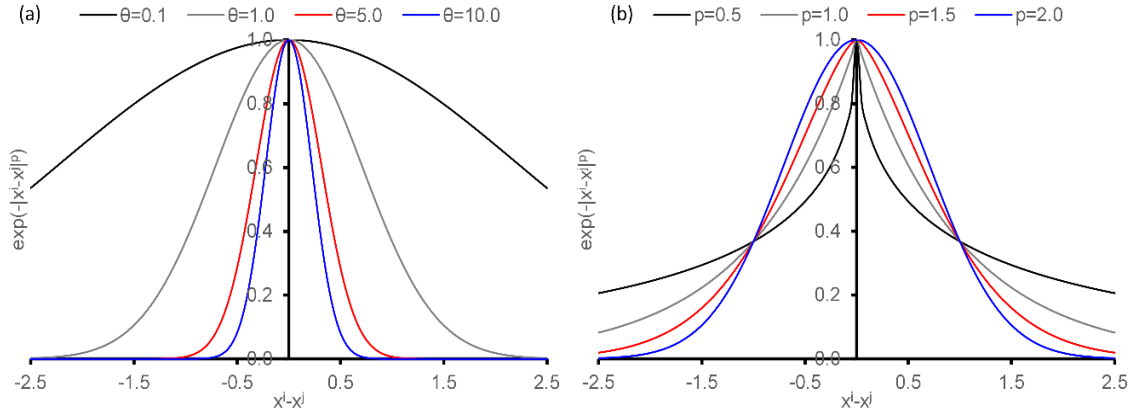
$$Y(\mathbf{x}) = f(\mathbf{x}) + Z(\mathbf{x}) \quad (8.3)$$

where  $f(\mathbf{x})$  is a regression function and  $Z(\mathbf{x})$  is a Gaussian process with a mean of zero and nonzero covariance. There exists a variety of Kriging techniques, including simple Kriging, ordinary Kriging, universal Kriging, etc. The difference of the various methods is the form of the regression function. *Jones et al. [1998]* suggested the use of ordinary Kriging whenever there is no rationalization to indicate a suitable *trend* function, and this suggestion is adopted in our work. This implies that  $f(\mathbf{x})$  is not a function, but a constant term.

Given  $n$  samples of  $\mathbf{x}^{(i)} = [x_1^{(i)}, \dots, x_d^{(i)}] \in \mathbb{R}^d$ , each with dimension  $d$  and a response  $\mathbf{y} = [y^{(1)}, \dots, y^{(n)}]^T$  the following pairs exist,  $(\mathbf{x}^{(i)}, y^{(i)})$  for  $i = 1, \dots, n$ . The covariance of  $Z(\mathbf{x})$  is given by,  $\text{Cov}[Z(\mathbf{x}^{(i)}), Z(\mathbf{x}^{(j)})] = \sigma^2 \Psi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$  where,  $\Psi(\cdot, \cdot)$  is a correlation function. The choice of correlation function is vital to create an accurate Kriging model regardless of the Kriging method. A commonly used correlation class, known as generalized exponential correlation function, is defined by:

$$\Psi(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp\left(-\sum_{h=1}^d \theta_h |x_h^{(i)} - x_h^{(j)}|^{p_h}\right), \quad \theta_h \geq 0, p_h \in [1, 2] \quad (8.4)$$

where  $\boldsymbol{\theta} = [\theta_1, \dots, \theta_d]$  and  $\mathbf{p} = [p_1, \dots, p_d]$  are the *hyperparameters* of correlation function. It is noted that this type of correlation class depends only on the Euclidian distance, with the correlation being inversely proportional to the distance. These parameters guide the rate and the shape of this relationship: Particularly, parameters  $\mathbf{p}$  determine the initial drop in correlation as distance increases. When  $\mathbf{p}$  is equal to 2 we have the popular Gaussian correlation function, which considers that the data follow a continuous and smooth surface. Lower values of  $\mathbf{p}$  are more suitable for rough response surfaces as they permit a more significant variance in function values for closer points (**Figure 8.3b**). **Figure 8.3a** depicts the influence of parameters  $\boldsymbol{\theta}$  (or *width*), showing how far a sample point's impact extends. These parameters are convenient as they describe the amount of variation in each dimension  $h$ . High values of  $\theta_h$  represent a non-linear behavior in dimension  $h$ , with similar points having significantly diverse responses. On the other hand low values of  $\theta_h$  indicate a more linear behavior in dimension  $h$ .



**Figure 8.3** | Behavior of generalized exponential correlation function (univariate case; for simplicity the index  $i$  is omitted) with, (a) varying  $\theta$  and fixed  $p = 2$  and (b) with varying  $p$  and fixed  $\theta = 1$ .

This study focuses on the use of Gauss correlation function (i.e.,  $p = 2$  in Eq. (8.4)). The correlation function  $\Psi(\cdot; \cdot)$  is parameterized by a set of hyperparameters  $\theta = [\theta_1, \dots, \theta_d]$ , whose identification is achieved using a Maximum Likelihood Estimation (MLE) method, in which the negative concentrated log-likelihood needs to be minimized [Forrester et al., 2008; Couckuyt et al., 2012].

### 8.2.3.3 Infill criteria

Once the surrogate model(s) has been constructed, the next design samples need to be evaluated. Infill criteria (also known as sample selection techniques or update strategies) aim to extract as much information as possible from the (cheap) surrogate and thus determine the next sample site (locating potential points) to be evaluated by the expensive simulator. These samples can be used to improve the accuracy and validate the performance of the surrogate. The iterative process of using infill criteria is known as adaptive sampling or active learning. There is a variety of strategies to design this process, including *inter alia* random, density, error based, and balanced exploration-exploitation strategies [e.g., Sasena et al., 2002; Forrester and Keane, 2009; Wagner et al., 2010; Zaefferer et al., 2013]. In this work we focus on balanced exploration-exploitation strategies, and some of them are described below.

#### 8.2.3.3.1 Infill strategies for single objective optimization

Two popular criteria which ensure a balance between exploitation and exploration are Probability of Improvement (PoI) and Expected Improvement (EI). Both of them became popular by the work of Jones et al. [1998] where they were implemented in a single objective SBO algorithm (EGO). PoI denotes the probability of a sample  $\mathbf{x}$  to lead to an improvement over the current minimum observed value  $y_{min}$  (calculated with the expensive function). By considering  $\hat{y}(\mathbf{x})$ , the prediction of Kriging, as a realization of a Gaussian random variable with variance  $\hat{s}^2(\mathbf{x})$ , the variance of kriging prediction, the probability of improvement  $I = y_{min} - \hat{y}(\mathbf{x})$  upon  $y_{min}$  is calculated as:

$$P[I(\mathbf{x})] = \frac{1}{\hat{s}(\mathbf{x})\sqrt{2\pi}} \int_{-\infty}^0 \exp\left(-\frac{(I - \hat{y}(\mathbf{x}))^2}{2\hat{s}^2(\mathbf{x})}\right) dI \quad (8.5)$$

Therefore  $P[I(\mathbf{x})]$  is the cumulative density function of  $\mathbf{x}$  and can be calculated using the error function as:

$$P[I(\mathbf{x})] = \Phi\left(\frac{y_{min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{y_{min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})\sqrt{2}}\right)\right] \quad (8.6)$$

The EI extends the definition of PoI and calculates the amount of improvement expected for a given  $\mathbf{x}$  and not only the PoI. The calculation of EI is shown below, where  $\Phi(\cdot)$ ,  $\varphi(\cdot)$  and  $\hat{s}(\mathbf{x})$  denote the Gaussian cumulative distribution function and probability density function, as well as Kriging error, respectively:

$$E[I(\mathbf{x})] = \begin{cases} (y_{min} - \hat{y}(\mathbf{x}))\Phi\left(\frac{y_{min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) + \hat{s}(\mathbf{x})\varphi\left(\frac{y_{min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right), & \text{if } \hat{s}(\mathbf{x}) > 0 \\ 0, & \text{if } \hat{s}(\mathbf{x}) = 0 \end{cases} \quad (8.7)$$

The ParEGO algorithm, used in this work, uses this single objective EI of an augmented Tchebycheff aggregation. Further details about this approach are given in a next section.

### 8.2.3.3.2 Infill strategies for multi-objective optimization

When dealing with multi-objective optimization problems a common practice is to aggregate the objective functions and use a single objective version of PoI or EI [Knowles, 2005]. Multi-objective versions of PoI and EI based on Euclidian distance have also been proposed [Keane, 2006; Forrester and Keane, 2009]. Emmerich et al. [2011] recently proposed a hypervolume-based EI criterion.

In our work we applied the strategy proposed by Couckuyt et al. [2013] implemented in the SUMO toolbox [Gorissen et al., 2010] which is based on an efficient method of calculation of hypervolume-based PoI and EI criteria. Here we adopt the hypervolume-based PoI criterion as described by [Couckuyt et al., 2013].

In order to define the concept of the hypervolume-based PoI criterion the multiobjective version of PoI needs to be defined. For notation reasons the output of each of  $m$  Kriging models (and thus objectives) can be viewed as independent Gaussian variables,  $Y_j(\mathbf{x})$  where  $j = 1, \dots, m$ . Hence:

$$Y_j(\mathbf{x}) = \mathcal{N}\left(\mu_j(\mathbf{x}), \hat{s}_j^2(\mathbf{x})\right), \text{ for } j = 1, \dots, m \quad (8.8)$$

Considering that given  $n$  points  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$  exist in  $d$  dimensions, a Pareto set  $\mathbf{P}$  can be derived, consisting of  $\nu \leq n$  non-dominated solutions.

$$\mathbf{P} = [f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_\nu^*)] \quad (8.9)$$

where  $f(\mathbf{x}_i^*)$  is a vector that contains the objective function values for the corresponding input  $\mathbf{x}_i^*$ ,  $i = 1, \dots, \nu$ . Generally, the Probability of Improvement over the Pareto set that a point  $\mathbf{x}$  can yield is calculated as [Couckuyt et al., 2013]:

$$P[I(\mathbf{x})] = \int_{y \in A} \prod_{j=1}^m \varphi_j[Y_j] dY_j \quad (8.10)$$

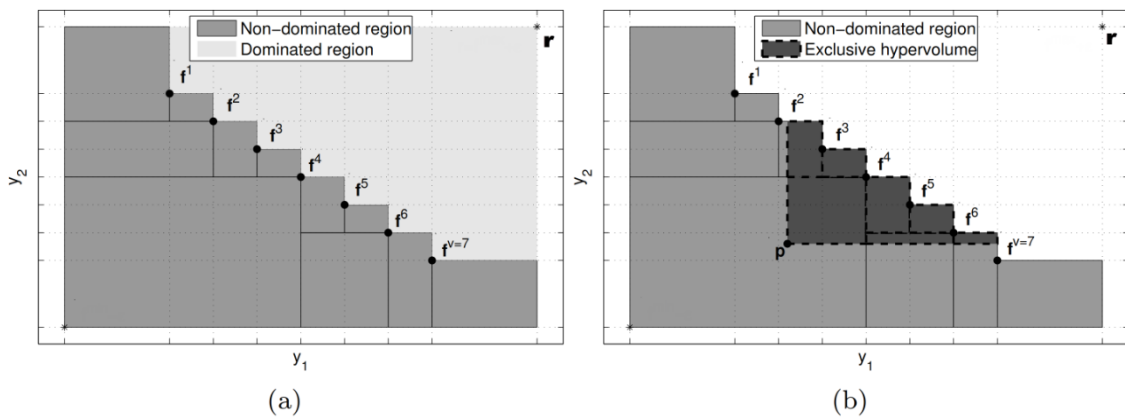
The basic concept of multiobjective PoI is to evaluate the probability of point  $\mathbf{x}$  to yield an objective function value located in a region A. Thus, A can be the non-dominated area of the objective space. In order to evaluate the multiobjective PoI, one has to decompose the area A in  $q$  rectangular cells; this yields a finite summation of contributing terms (Figure 8.4b). Thus  $P[I(\mathbf{x})]$  is transformed to:

$$P[I(\mathbf{x})] = \sum_{k=1}^q \pm \prod_{j=1}^m (\varphi_j[u_j^k] - \varphi_j[l_j^k]) \quad (8.11)$$

where  $[l^k, u^k]$  are the lower and upper bounds of each  $q$  cells. There are numerous approaches to calculate the hypervolume (i.e. the space within the aforementioned bounds) [e.g., Fleischer, 2003; Beume, 2009; Beume et al., 2009; Bringmann and Friedrich, 2009; Bader et al., 2010]. The SUMO toolbox implements the Walking Fish Group (WFG) algorithm [While et al., 2012] which was originally introduced by Fleischer [2003] which is deemed faster for practical optimization problems.

So far we have defined the concept of multiobjective PoI, yet in order to derive a hypervolume-based PoI criterion the hypervolume-based improvement function needs to be defined. Here we adopt the Hypervolume metric (HV) [Zitzler and Thiele, 1999; Zitzler et al., 2003]. The hypervolume  $HV(\mathbf{P})$  of a given Pareto set  $\mathbf{P}$  essentially measures how much volume the non-dominated set  $\mathbf{P}$  dominates relative to a reference point  $\mathbf{r}$  (termed *anti-ideal*). The reference point  $\mathbf{r}$  should be dominated by every point of the Pareto set. The HV value is proportional to the quality of the Pareto set  $\mathbf{P}$ . Another essential and interesting notion is hypervolume contribution ( $HV_{\text{con}}$ ), which measures the contribution of a point  $\mathbf{p}$  to the overall Pareto set  $\mathbf{P}$ . Therefore,  $HV_{\text{con}}$  can be used to define an improvement function. It should be noted that  $HV_{\text{con}}$  does not require normalization of the objective space [Knowles, 2002]. Figure 8.4 depicts the concepts of HV and  $HV_{\text{con}}$ , as well as the integration area A of the hypervolume-based PoI which is decomposed into smaller cells by a binary partitioning procedure [Couckuyt et al., 2013].

$$I(\mathbf{p}, \mathbf{P}) = \begin{cases} H(\mathbf{P} \cup \{\mathbf{p}\}) - H(\mathbf{P}), & \text{if } \mathbf{p} \text{ is not dominated by } \mathbf{P} \\ 0, & \text{otherwise} \end{cases} \quad (8.12)$$



**Figure 8.4** | Graphical representation of hypervolume and hypervolume contribution. a) The light grey depicts the hypervolume of the Pareto set (non-dominated region) b) The dark grey depicts the exclusive (contribution) hypervolume of a point  $\mathbf{p}$  (adopted from: [Couckuyt et al., 2013]).

Having defined an improvement function we are able to specify a hypervolume-based PoI criterion which is derived by multiplying  $I(\boldsymbol{\mu}, \mathbf{P})$  and  $P[I(\mathbf{x})]$ . It should be noted that the integration area  $A$  of  $P[I(\mathbf{x})]$  corresponds to the non-dominated region, and thus the  $P_{HV}[I(\mathbf{x})]$  can be calculated in closed-form. Area  $A$  can be derived from the same set of cells used to evaluate  $P[I(\mathbf{x})]$  [Couckuyt et al., 2013] (see Figure 8.4b). These observations are formulated in Eq. (8.13) and (8.14), where  $\boldsymbol{\mu} = [\mu_1(\mathbf{x}), \dots, \mu_\mu(\mathbf{x})]$  is a vector that contains the surrogate model's (Kriging) mean predictions:

$$P_{HV}[I(\mathbf{x})] = \left( \sum_{k=1}^q \pm \text{Vol}(\boldsymbol{\mu}, l^k, u^k) \right) P[I(\mathbf{x})] \quad (8.13)$$

where,

$$\text{Vol}(\boldsymbol{\mu}, \mathbf{l}, \mathbf{u}) = \begin{cases} \prod_{j=1}^m (u_j - \max(l_j, \mu_j(\mathbf{x}))), & \text{if } u_j > \mu_j \text{ for } j = 1, \dots, m \\ 0, & \text{otherwise} \end{cases} \quad (8.14)$$

#### 8.2.3.4 Model assessment and validation

In order to assess the quality of the surrogate model, and hence its prediction capabilities, an error estimation method needs to be specified. The most common method is cross-validation (CV), according to which training data are divided into  $q$  equally sized subsets. An iterative procedure of  $q$  iterations then begins. In each iteration one subset is removed and the model is fitted to the remaining data. For every iteration the excluded subset has the role of the validation set and hence it is used to calculate a selected error measure. The general model error is computed using the  $q$  error measures obtained (e.g., is the average error of  $q$  models). The parameters that yield the minimum general model error are then selected. In CV, most if not all of the available data are used. If the mean square error (MSE) is selected as the error measure the following equation is used:

$$\text{err}_{cv}^{\text{mse}} = \frac{1}{q} \sum_{i=1}^q (y_i - \hat{y}_i)^2 \quad (8.15)$$

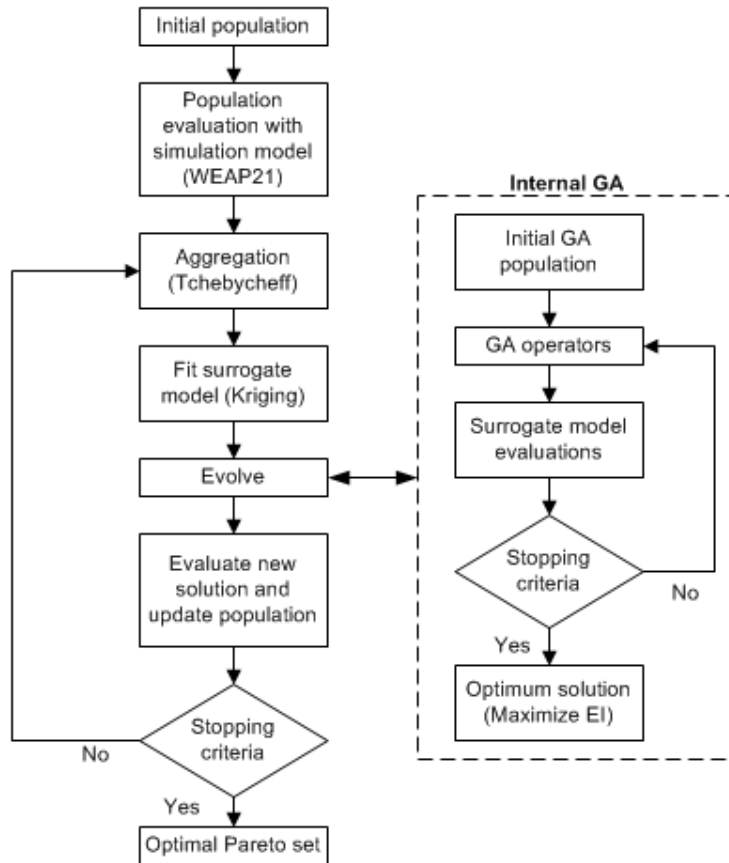
where,  $y_i$  is the real function values and  $\hat{y}_i$  is the predicted value of sample  $\mathbf{x}^i$  of the surrogate model constructed without the use of  $(\mathbf{x}^i, y_i)$ . Cross-validation is a reliable method particularly when the use of a separate validation set is computationally expensive.

#### 8.2.4 The deployed MOSBO algorithms

##### The ParEGO algorithm

The ParEGO algorithm was introduced by Knowles [2005] and is an global optimization algorithm for expensive multi-objective optimization problems. ParEGO is essentially a multi-objective conversion of EGO [Jones et al., 1998], making use of scalarizing weight vectors at each step. The algorithm is based on the Design and Analysis of Computer Experiments (DACE) model [Sacks et al., 1989]. ParEGO uses DACE (i.e., Kriging) to fit previously evaluated points and uses the fitted model to locate interesting (i.e. potentially better) new

points to visit subsequently. One of the advantages of DACE, which is used in both EGO and ParEGO, is that a confidence interval of prediction is available and used to guide the search. The iterative process of ParEGO is shown in the flow chart of **Figure 8.5**. An initial set of solutions is generated using the Latin Hypercube design [Press et al., 1992] and evaluated with the expensive simulation function (in our case WEAP21). Then the objectives are aggregated using Tchebycheff function (see below) and the initial DACE model is generated by fitting these solutions. The algorithm then tries to predict a trial solution which is most likely to improve the best fit found so far. An internal genetic algorithm (GA) is used to find the solution that maximizes Expected Improvement (EI) and to update the solution set of points evaluated by the expensive function. The DACE model is then updated and the next iteration begins.



**Figure 8.5** | ParEGO algorithm flowchart.

In order to solve multi-objective problems, the algorithm updates the weighting between the objectives using the non-linear Tchebycheff function (i.e., Eq. (8.16)) as suggested by Knowles [2005] which combines  $m$  objectives to a single objective, thus gradually building up the whole Pareto front. The Tchebycheff function is described below:

$$f_{\lambda}(\mathbf{x}) = \max_{j=1}^n \left( \lambda_j f_j(\mathbf{x}) \right) + \rho \sum_{j=1}^m \lambda_j f_j(\mathbf{x}) \quad (8.16)$$

where  $f_j(\mathbf{x})$  and  $\lambda_j$  ( $j = 1, 2, \dots, m$ ) are the  $j^{\text{th}}$  normalized objective values with respect to the known (or estimated) limits of the cost space, so that each cost function and its weight lie in the range  $[0, 1]$ , and  $\rho$  is typically set equal to 0.05 [Knowles, 2005]. In order to gradually build the whole Pareto front a weight vector  $\lambda$  is drawn uniformly at random from the set of evenly

distributed vectors defined by,  $\Lambda = \left\{ \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m) \mid \sum_{j=1}^m \lambda_j = 1 \forall j, \lambda_j \in \left\{ 0, \frac{1}{s}, \dots, 1 \right\} \right\}$ , with  $|\Lambda| = \binom{s+k-1}{k-1}$ , so the choice of  $s$  determines the total number of the weight vectors [Knowles, 2005].

### The SURrogate MOdeling Toolbox

The SURrogate MOdeling (SUMO) Toolbox<sup>6</sup> [Gorissen et al., 2010] is capable of building global surrogate models of a given data source (data set, external code, script). The toolbox supports a variety of DoE techniques (Latin Hypercube, etc.), surrogate model types (e.g. Kriging, SVMs, RBF, ANN, etc.) and also includes several optimization algorithms (particle swarm optimization, simulated annealing, genetic algorithm, etc.) to optimize surrogate model parameters (hyperparameters), such as the parameters of the Gauss correlation function. Moreover, it includes model selection (cross validation, leave-out set, etc.) and sample (infill criteria) selection methods (including random, error based, density based, expected improvement and hypervolume PoI, etc.). Each component of SUMO is configured through an XML file. The basic workflow of SUMO used in this work is as follows: 1) initial sampling with DoE (Latin Hypercube design) and evaluation with expensive function; 2) building of models for each objective (with ordinary Kriging in our case); 3) determination of the next design sample using appropriate infill criteria (in our case the hypervolume-based PoI criterion has been chosen as described in section 8.2.3.3.2); 4) evaluating of the design sample and updating the non-dominated set of points, 5) iteration between steps 2-4 until the stopping criteria are satisfied (in our case when a maximum number of function evaluations is reached).

### The SMS-EGO algorithm

SMS-EGO is a surrogate based multiobjective algorithm originally proposed by [Ponweiser et al., 2008]. The key idea of the algorithm is to utilize a hypervolume-based criterion termed hypervolume contribution  $HV_{\text{con}}$ . In order to evaluate potential solutions, the algorithm uses the lower conditional bound (LCB) criterion as suggested by [Emmerich et al., 2006]:

$$\hat{\mathbf{y}}_{\text{pot}}(\mathbf{x}) = \hat{\mathbf{y}}(\mathbf{x}) - a\hat{\mathbf{s}}(\mathbf{x}) \quad (8.17)$$

where,  $\hat{\mathbf{y}}(\mathbf{x})$ ,  $\hat{\mathbf{s}}(\mathbf{x})$  and  $a$  are the Kriging prediction, error and a weight factor respectively. The value of  $a$  can be derived by a user defined confidence probability given by  $P_a = (1 - 2\Phi(\alpha))^m$  as suggested by Emmerich et al. [2006] and Ponweiser et al. [2008]. Furthermore, SMS-EGO applies additive  $\varepsilon$ -dominance [Zitzler et al., 2003] in order to ensure good distribution of the Pareto set. Therefore, if a potential point  $\hat{\mathbf{y}}_{\text{pot}}$  is dominated or  $\varepsilon$ -dominated, a penalty value  $p$  is assigned. If it is non-dominated, the hypervolume contribution will be calculated and the point with the higher  $HV_{\text{con}}$  value will be chosen for evaluation with the expensive function (simulator).

$$p = \sum_{\hat{\mathbf{y}}^{(i)} \in \mathbf{P}} \begin{cases} -1 + \prod_{j=1}^m \left( 1 + \left( \hat{\mathbf{y}}_{\text{pot},j} - \hat{\mathbf{y}}_j^{(i)} \right) \right), & \text{if } \hat{\mathbf{y}}^{(i)} \leq \hat{\mathbf{y}}_{\text{pot}} \\ 0, & \text{otherwise} \end{cases} \quad (8.18)$$

<sup>6</sup> <http://www.sumowiki.intec.ugent.be>

The use of LCB allows the algorithm to explore unvisited areas (high variance) without requiring integration over the objective space as opposed to other strategies [Couckuyt et al., 2012, 2013].

### 8.3 THE STUDY AREA: THE HYDRO-SYSTEM OF NESTOS, GREECE

The Nestos basin, with a total area of 6,219 km<sup>2</sup>, is a trans-boundary basin which extends between Bulgaria (60% share) and Greece [Paraskevopoulos – Pangaea, 1994]. Nestos is the second largest river in Thrace River Basin District and one of the major rivers in Greece (Figure 8.7). It flows from Mount Rila in Bulgaria (a region with the highest altitude of the Balkans, about 2,925 m) and has a total length of 234 km of which 130 are in Greek territory. Nestos flows into Greece

Figure 8.6

from the plateau of Nevrokopi of Drama. The river forms a natural boundary between Bulgaria and Greece for a few kilometers. The river flows into the Sea of Thrace, forming a large delta area about 50 km<sup>2</sup>. It is worth mentioning that the Nestos estuary is an area protected by the Ramsar Convention and is part of the NATURA 2000 network.

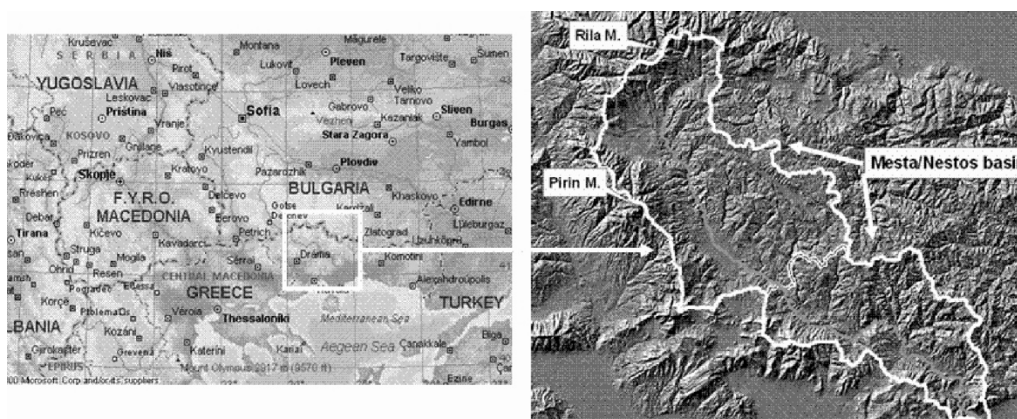


Figure 8.7 | Geographical representation of the river basin Mesta / Nestos [Skoulikaris et al., 2008].

The first dam constructed in the region was Toxotes (1960-1966). It is an irrigation dam with a length of 280 meters located in the neck of the estuary and diverts the quantities of water to the east (city of Xanthi) and western bank (city of Kavala) of the main stream of the river. The first feasibility study of the Toxotes dam was undertaken in 1954 [YDE, 1954]. Later on, during 1971-1972 a feasibility study was undertaken for the construction of three more upstream dams. Construction began ten years later (mid-1980's) based on an interim agreement with Bulgaria on minimum incoming water quantities in Greece. In 1995 an agreement was signed with Bulgaria to allow at least 29% of the total river flow to reach Greece. The initial plan of the feasibility study was to build 3 hydroelectric power stations (serially) with the first two of them reversible (pump-storage). These stations are the Thysavros (381 MW), the Platanovryssi (116 MW) and the Temenos station (19 MW). These projects are multi-purpose, providing water for irrigation and potable water to small towns and industrial areas, and for energy production. The hydroelectric plant of the Thysavros reservoir is at the head of the system and provides scaling, while controlling the annual runoff of the river. However, the overall project has not been completed yet due to lack of funds: two reservoirs have been constructed so far (Thysavros and Platanovryssi). The third dam, Temenos, remains unconstructed [Skoulikaris et al., 2008] but privatization policies have recently revived interest in its completion. In our study we include Temenos and explore the performance of the complete hydrosystem.



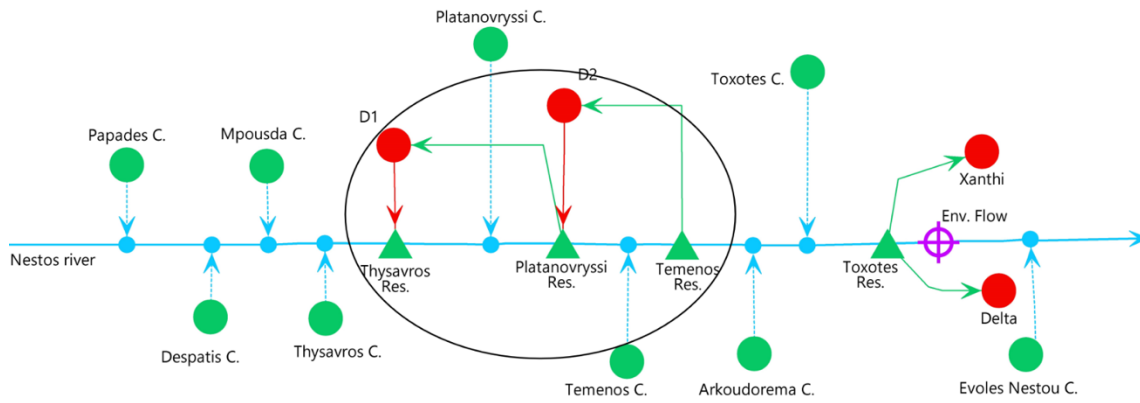
The hydrosystem was modeled (**Figure 8.8** using WEAP21. The calibration of the hydrosystem was successfully undertaken during a previous study [Tsoukalas and Makropoulos, 2013, 2015a] yielding a Nash-Sutcliffe coefficient equal to 0.87. WEAP21 doesn't have a built-in module to simulate pump-storage processes; As this is an important functionality for the Nestos hydrosystem, pump storage was simulated using *dummy* demand nodes (D1 and D2) with the appropriate connectivity (**Figure 8.8**, inside the circle). In order to define the maximum monthly turbine/pump discharge of the reservoirs it was assumed that the hydroelectric plants operate 18 hours a day as turbines (energy production) and 6 hours as pumps (consumption). Parameters in D1 and D2 that control the volume of water used in the pump-storage process were assumed to equal their maximum discharge capacity. The reason behind that is the fact that the pump-storage contributes significantly to the produced hydropower energy and the (monthly) simulation time-step is not able to depict clearly the variations in the storage volume of the reservoirs due to pump-storage. Irrigation water demand is significant downstream the Toxotes reservoir (**Table 8-1**) and a constant environmental flow requirement is set to 6 m<sup>3</sup>/s by legislation. The demand priorities of environmental flow, irrigation and hydropower production were set to 1, 2 and 3, respectively.

WEAP21 uses a linear programming solver to allocate the available water resources across the hydrosystem. The objective of the LP solver is to maximize satisfaction of demands, subject to demand priorities, mass balances and other constraints. This routine ensures a physically-consistent description of the hydrosystem and the satisfaction of all constraints.

**Table 8-1** | Present and future irrigation demand below Toxotes reservoir.

| Month/Demand (m <sup>3</sup> /s) | Apr  | May  | Jun  | Jul  | Aug  | Sept |
|----------------------------------|------|------|------|------|------|------|
| Delta                            | 11.5 | 15.7 | 18.5 | 20.9 | 20.0 | 13.0 |
| Xanthi (future)                  | 5.7  | 7.8  | 9.2  | 10.4 | 10.0 | 6.5  |

The parametric rule discussed earlier tries to identify *optimal*  $\theta_i$  and  $\beta$  (where  $i$  is the number of hydroelectric reservoirs). To implement this parameterization within WEAP21 the build-in model parameter “*Energy Demand*” that can be specified at each reservoir was used. Therefore, in this case study we had four (4) decision variables. More specifically for low energy demand months (October-April), “*Energy Demand*” was set to  $\theta_i$  multiplied by the reduction coefficient  $\beta$  and for remaining year (May-Sept) it was set to  $\theta_i$  without any reduction coefficient applied. For example, to implement  $\theta_i = 40$  and  $\beta = 0.5$  for Reservoir 1, the reservoir’s “*Energy Demand*” parameter was set to 20 GWh for October-April and to 40 GWh for May-Sept. This yields a workable and simple operation rule, provided of course that  $\theta_i$  and  $\beta$  are selected after an optimization process accounting for the hydrological variability of inflows over the longer run. It should be noted that the current operation of the reservoirs does not follow specific rules and is based mostly on ad hoc expert judgment. Hence, unfortunately, no meaningful comparison of current practice versus optimization results can take place.



**Figure 8.8** | Hydrosystem modelled in WEAP21 and detail (inside the circle) of simulation of pump-storage. Symbol (●) represents the catchments, (●) represents the demand nodes, (⊗) represents in-stream flow requirements, (→) represents the river, (●) represents river nodes or junctions and (▲) represents the reservoirs.

#### 8.4 BENCHMARKING THE ALGORITHMS' PERFORMANCE

Contrary to single-objective optimization, comparing multi-objective optimization algorithms is more complicated due to the fact that the outcome is an approximation of the Pareto front consisting of many different solutions. A possible way to address this problem is to condense performance into scalar metrics – also known as performance indicators. A comprehensive review of performance indicators was recently presented by *Zitzler et al. [2008]*. In earlier work, *Zitzler et al. [2003]* argue that some of the most typical features that performance indicators have to assess are: a) the precision of the solutions in the set, i.e. how well they approximate the ideal Pareto front; b) the number of solutions contained in the final Pareto-optimal set and c) the spread and distribution of the solutions. Following this rationale, we chose two indicators to compare the optimization algorithms in this work: the hypervolume indicator and the unary  $\epsilon$ -indicator. Both of them are briefly described in the following paragraphs.

Furthermore, in order to accurately benchmark the performance of MOSBO and MOEA, some basic concepts of the methodology proposed by *Razavi et al. [2012a]*, *Asadzadeh and Tolson [2013]* and *Matott et al. [2012]* were adopted: Due to the stochastic nature of MOSBO and MOEA, multiple independent runs of each algorithm are utilized. The performance indicators are then calculated for each run. For each performance indicator the empirical cumulative distribution function (CDF) is calculated to depict the probability of obtaining an equivalent or better solution. The concept of stochastic dominance [*Levy, 1992*] is then utilized. Specifically, the first degree of stochastic dominance (SD) is used to compare the CDFs of the algorithms [*Matott et al., 2012; Razavi et al., 2012a; Asadzadeh and Tolson, 2013*]. To explain this, consider the comparison of two algorithms A and B, with CDFs  $\Phi_A$  and  $\Phi_B$  based on a performance metric  $m$ , such that smaller values of  $m$  are preferred. SD of A over B applies only if  $\Phi_A(m) \geq \Phi_B(m)$  for all  $m$ . The SD does not apply when the CDFs are crossed. In order to statistically assess the differences between the CDFs, the non-parametric Mann–Whitney U test (MWU) is used. The null hypothesis of the MWU test is that data in  $\Phi_A$  and  $\Phi_B$  are samples from continuous distributions with equal medians. The confidence level of the MWU test was set to 90%. Finally, to visualize results from multiple runs, the empirical attainment function (EAF) proposed by *da Fonseca et al. [2001]* is employed which describes the probabilistic distribution of the outcomes obtained by a MOOA. The functionality and the properties of the EAF are also briefly described next.

### 8.4.1 Hypervolume indicator

The hypervolume indicator ( $HV$ ) was first proposed by *Zitzler and Thiele [1999]*, who called it S-metric. The  $HV$  essentially measures how much volume a non-dominated set  $\mathbf{P}$  dominates relative to a reference point  $\mathbf{r}$  (anti-ideal). The reference point  $\mathbf{r}$  should be dominated by every point of the Pareto set. This indicator has become popular due to its ability to depict the accuracy and the spread of the approximation set. The choice of the reference point  $\mathbf{r}$  is crucial due to the fact that the contribution of extreme points essentially depends on it [*Knowles and Corne, 2003*].

In this work we consider the normalized hypervolume ratio indicator (NHVR) which is the  $HV$  in the normalized objective space  $[0,1]$ , divided by the  $HV$  of the *true* (optimal or reference) Pareto front. In our work the *true* Pareto front was not available and thus a reference (ideal) Pareto front was created compiling data from all optimization runs. The reference point selected is the maximum of the  $j^{\text{th}}$  objective shifted by  $a = 10\%$ . In other words, for  $m$  objectives, we defined a vector  $\mathbf{f} = [f_1, f_1, \dots, f_m]$  with  $f_j = \max_j + a(\max_j - \min_j)$ ,  $j \in \{1, \dots, m\}$ , where  $\max_j$  and  $\min_j$  represent the maximum and minimum of the  $j^{\text{th}}$  objective respectively.

### 8.4.2 Unary $\epsilon$ -indicator (epsilon indicator)

This indicator ( $I_\epsilon$ ) was first proposed by *Zitzler et al. [2003]*, aiming to identify the minimum distance between a given approximation of the Pareto set and the true or a benchmark Pareto front. In this case the objectives are also normalized in  $[0,1]$  space before calculating the indicator. Lower values of the  $I_\epsilon$  indicate that the approximation set is closer to the reference front (our case it is the same as in NHVR). A detailed description of  $I_\epsilon$  is also included in [*Fonseca et al., 2005*].

### 8.4.3 Empirical attainment function

The Empirical Attainment Function (EAF) was originally proposed by *da Fonseca et al. [2001]*. Later on, *Zitzler et al. [2008]* investigated in depth the properties of the EAF, while *Fonseca et al. [2011]* formalized the problem of its computation and proposed efficient algorithms for two and three dimensional computation. The key concept of the EAF is to calculate the probability that an algorithm will dominate an arbitrary point in the objective space at one run. Because of the stochastic nature of evolutionary algorithms there is no guarantee that the algorithm will achieve the same Pareto front at each run. A key advantage of the EAF is that the whole Pareto can be observed, therefore strong and weak areas of the front can be easily identified. Using color gradients, the EAF depicts the relative number of times that each region of the objective space is dominated. In our work we used an R package tool<sup>7</sup> [*López-Ibáñez et al., 2010*] which also provides the ability to compare two algorithms in a single plot using the differential empirical attainment function (Diff-EAF), which was particularly useful for our purposes. Diff-EAF expresses the probability that a point in the solution space is dominated by only one of the compared algorithms. Therefore, Diff-EAF depicts the difference of two EAFs in a single graph. Thus, one can visually distinguish which algorithm performs better in certain regions of the objective space.

---

<sup>7</sup> available from <http://iridia.ulb.ac.be/~manuel/eaftools>

## 8.5 EXPERIMENTAL SETUP

In order to evaluate and compare the performance of ParEGO, SMS-EGO and SUMO, each algorithm was run multiple times. The configuration of the ParEGO, SMS-EGO and SUMO parameters was based on *Knowles [2005]*, *Ponweiser et al. [2008]* and *Couckuyt et al. [2013]* respectively, and it was attempted to keep them as similar as possible for comparison purposes. Experimental setup similarities include the initial population sample (set to 54), the DoE method (Latin hypercube), and finally the surrogate models (ordinary Kriging). The hyperparameters of the models were optimized in all cases using GAs. The main difference between the algorithms regards the infill criteria strategy. ParEGO uses the Tchebycheff function, aggregates the objectives into a single objective, and then uses the EI criterion. SMS-EGO uses the hypervolume contribution in combination with LCB and a penalty function for dominated solutions. SUMO uses the hypervolume-based PoI criterion to locate promising samples. All MOSBO algorithms were tested with Gauss correlation functions for the Kriging models, and for 200 and 400 function evaluations (FE).

The NSGAI and SMS-EMOA were used as benchmark algorithms. For that purpose, 10 independent optimization runs were performed for both algorithms for 200, 400 and 1 000 FE. An additional 5 independent optimization runs were performed for 2 000 FE and another 2 for 5 000 FE in order to evaluate the performance of MOEAs for larger numbers of FE (in these last cases 10 runs were impractical due to the excessive computational time required). The parameter settings used for NSGAI are similar to those proposed by Deb's KANGAL web page. The population size was set to 50 and the maximum generation number was set accordingly to the maximum FE allowed. The same setup was used for SMS-EMOA. A summary of the optimization runs and configurations applied in this work is presented in **Table 8-2**.

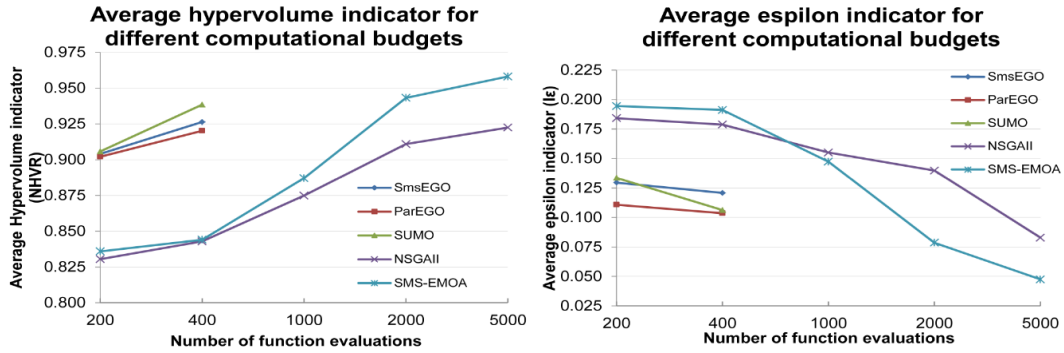
**Table 8-2** | Summary of optimization runs and configurations.

| Algorithm | Correlation function | Initial population size | Function Evaluations (Optimization runs)      |
|-----------|----------------------|-------------------------|---|
| ParEGO    | Gauss                | 54                      | 200(10), 400(10)                              |
| SMS-EGO   | Gauss                | 54                      | 200(10), 400(10)                              |
| SUMO      | Gauss                | 54                      | 200(10), 400(10)                              |
| NSGAI     | -                    | 50                      | 200(10), 400(10), 1000(10), 2000(5), 5 000(2) |
| SMS-EMOA  | -                    | 50                      | 200(10), 400(10), 1000(10), 2000(5), 5 000(2) |

## 8.6 RESULTS AND DISCUSSION

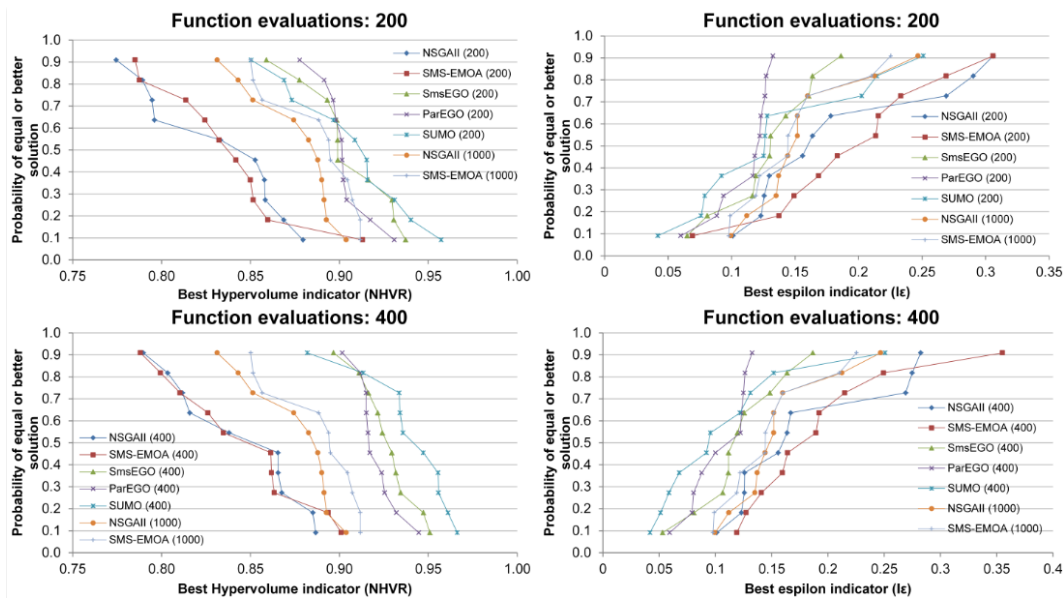
### 8.6.1 Comparison and benchmarking results

The average performance of the applied MOSBO and MOEA algorithms in our case based on the selected performance indicators is presented in **Figure 8.9** for each investigated configuration (computational budget). This figure depicts the superiority of MOSBO algorithms over MOEAs in few FE (200, 400). In both cases, all MOSBOs manage to attain larger values of NHVR compared to MOEAs, and simultaneously achieve lower values of I<sub>e</sub>. It seems that beyond the 400 FE the performance of the MOEAs improves substantially and finally reaches the MOSBOs' performance (at around 1 000 to 2 000 FE). Further FE (2 000-5 000) have the potential to lead to MOEAs to outperform MOSBOs. This is further discussed in a following paragraph where MOSBO algorithms with 200 and 400 FE are compared with MOEAs with larger computational budget.



**Figure 8.9** | Average performance of algorithms for various computational budgets.

Furthermore, **Figure 8.10** presents the empirical CDFs of all MOSBOs and MOEAs for 200 and 400 FE, as well as the CDFs of MOEAs for 1 000 FE. It is observed that for 200 and 400 FE all MOSBOs stochastically dominate MOEAs in both NHVR and  $I_\epsilon$ .



**Figure 8.10** | Empirical CDF of all MOSBO and MOEA for 200 and 400 function evaluations; also the CDFs of MOEA for 1000 function evaluation are depicted.

**Comparison between equal budgets**

**Table 8-3** summarizes the results of all algorithms for 200 and 400 FE and presents some key statistics. In **Table 8-4** the algorithms are ranked according to their performance, with regards to the NHVR and  $I_\epsilon$ . As anticipated the top three ranks are occupied by the MOSBO algorithms. The CDFs of the MOSBOs intersect, hence stochastic dominance (SD) does not apply. In order to rank the algorithms, the following procedure was applied: the preferred algorithm is the one with the highest median in the case of NHVR and the one with the lowest median in the case of  $I_\epsilon$ . The alternative algorithm is the one with the highest P-value in the Mann–Whitney U test (MWU) when compared with the preferred algorithm.

Furthermore, Diff-EAF plots depict the best (upper line) and the worst (lower line) EAF of a pair of algorithms. The median EAF (dotted line in the middle) of each algorithm is plotted. For example, the worst empirical attainment function (EAF) represents the smallest objective space that was captured from all runs and for each pair of algorithms. **Table 8-4** also presents the P-values from the MWU test for all pairs. The high P-values indicate the acceptance of the null hypothesis (i.e. that data are samples from continuous distributions with equal medians).

In most of the cases the P-values are above 0.1 (confidence level 90%). The MWU test illustrates that the medians of the algorithms do not differ significantly. In the case of 200 FE there is a *tie* between SUMO and ParEGO, with  $I_\epsilon$  indicating that the preferred algorithm is ParEGO and NHVR indicating SUMO. However, the high P-values (in both cases) suggest that there is no significant difference between the medians. Therefore, the performance of all MOSBO in 200 FE could be assumed equivalent with very small differences. In the case of 400 FE both NHVR and  $I_\epsilon$  agree that the preferred algorithm is SUMO. Also in this case the P-values are high, except when comparing SUMO and ParEGO, where the P-value is less than 0.1. This warrants further exploration as discussed below.

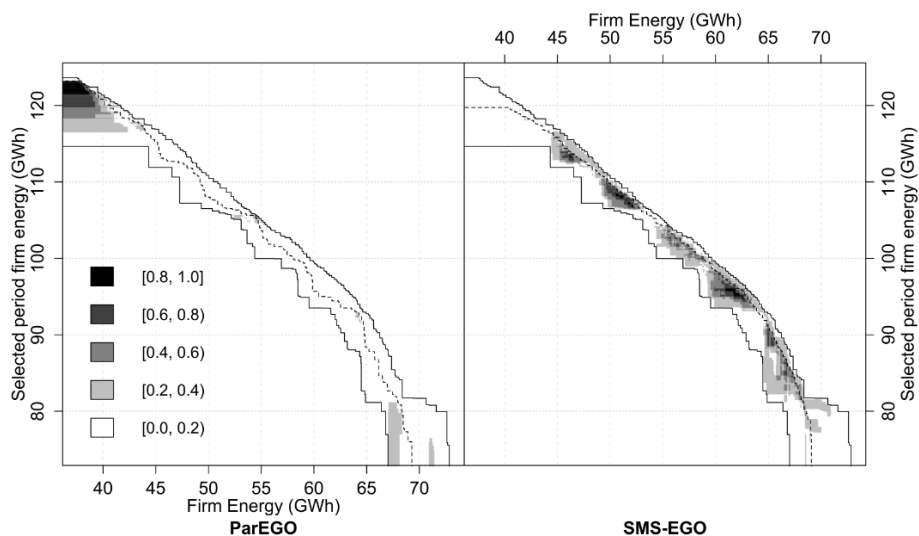
**Table 8-3** | Results summary for 200 and 400 function evaluations.

| Hypervolume indicator |           |        |         |       |          |       |
|-----------------------|-----------|--------|---------|-------|----------|-------|
| Budget                | Statistic | ParEGO | SMS-EGO | SUMO  | SMS-EMOA | NSGAI |
| 200                   | Average   | 0.902  | 0.904   | 0.906 | 0.836    | 0.830 |
|                       | Median    | 0.901  | 0.899   | 0.912 | 0.837    | 0.843 |
|                       | St. dev.  | 0.014  | 0.025   | 0.034 | 0.037    | 0.038 |
|                       | Max       | 0.931  | 0.937   | 0.957 | 0.913    | 0.879 |
|                       | Min       | 0.878  | 0.859   | 0.850 | 0.785    | 0.774 |
| 400                   | Average   | 0.920  | 0.926   | 0.938 | 0.844    | 0.843 |
|                       | Median    | 0.917  | 0.927   | 0.941 | 0.848    | 0.852 |
|                       | St. dev.  | 0.012  | 0.016   | 0.025 | 0.039    | 0.036 |
|                       | Max       | 0.945  | 0.951   | 0.966 | 0.901    | 0.887 |
|                       | Min       | 0.902  | 0.897   | 0.882 | 0.788    | 0.790 |
| Epsilon indicator     |           |        |         |       |          |       |
| Budget                | Statistic | ParEGO | SMS-EGO | SUMO  | SMS-EMOA | NSGAI |
| 200                   | Average   | 0.111  | 0.130   | 0.134 | 0.194    | 0.184 |
|                       | Median    | 0.120  | 0.130   | 0.126 | 0.198    | 0.160 |
|                       | Std       | 0.023  | 0.037   | 0.068 | 0.068    | 0.076 |
|                       | Max       | 0.133  | 0.186   | 0.251 | 0.306    | 0.306 |
|                       | Min       | 0.060  | 0.065   | 0.042 | 0.069    | 0.101 |
| 400                   | Average   | 0.104  | 0.121   | 0.106 | 0.191    | 0.179 |
|                       | Median    | 0.111  | 0.116   | 0.094 | 0.177    | 0.160 |
|                       | St. dev.  | 0.026  | 0.039   | 0.062 | 0.070    | 0.070 |
|                       | Max       | 0.133  | 0.187   | 0.251 | 0.355    | 0.282 |
|                       | Min       | 0.059  | 0.053   | 0.042 | 0.119    | 0.101 |

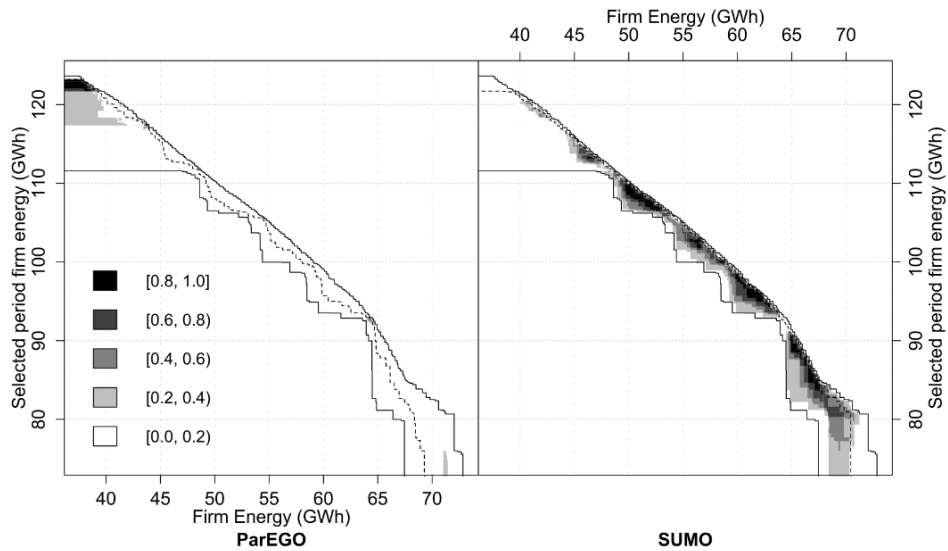
**Table 8-4** | Comparison of MOSBO and MOEA under equal budget.

| Alg/Budget               | $I_\epsilon$ |         | NHVR    |         |
|--------------------------|--------------|---------|---------|---------|
|                          | 200          | 400     | 200     | 400     |
| Preferred <sup>1</sup>   | ParEGO       | SUMO    | SUMO    | SUMO    |
| Alternative <sup>2</sup> | SUMO         | ParEGO  | SMS-EGO | SMS-EGO |
| Third <sup>3</sup>       | SMS-EGO      | SMS-EGO | ParEGO  | ParEGO  |
| P-value <sub>12</sub>    | 0.571        | 0.623   | 0.910   | 0.104   |
| P-value <sub>13</sub>    | 0.186        | 0.345   | 0.623   | 0.031   |
| P-value <sub>23</sub>    | 0.791        | 0.427   | 0.999   | 0.308   |

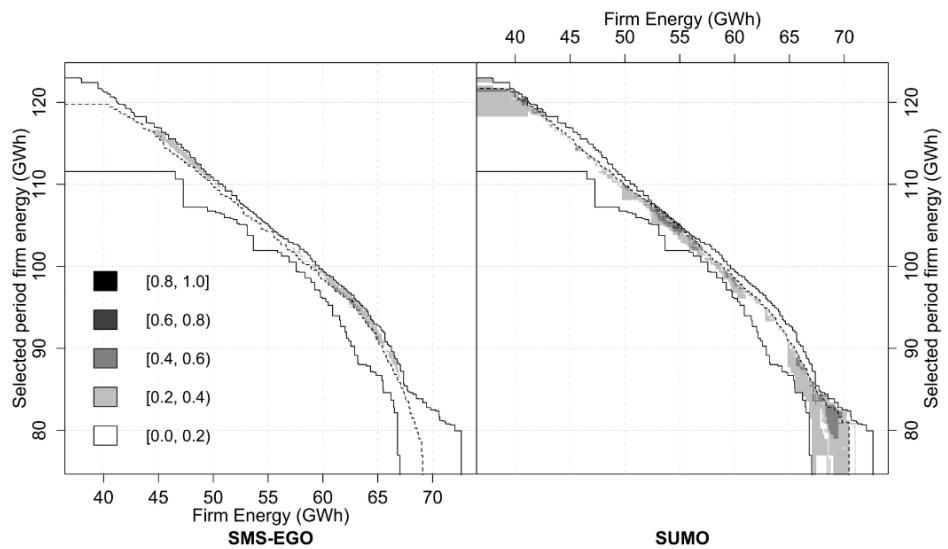
As mentioned before the differential empirical attainment function (Diff-EAF) was used to provide a comprehensive visualization of the results and identify areas where an algorithm outperforms another. For the sake of limited space Diff-EAF plots are presented only for configurations with 400 FE. Diff-EAF uses color gradients to express the probability that a point in the solution space is dominated by only one of the compared algorithms. We used gray scale to distinguish the areas in the objective space where each algorithm performs better. Dark grey areas indicate high probability that this algorithm dominates that area and respectively light grey to white areas indicate the opposite. Furthermore, Diff-EAF plots depict the best (upper line) and the worst (lower line) EAF of a pair of algorithms. The median EAF (dotted line in the middle) of each algorithm is plotted. For example, the worst empirical attainment function (EAF) represents the smallest objective space that was captured from all runs and for each pair of algorithms. **Figure 8.11** depicts the Diff-EAF of the ParEGO algorithm as compared to SMS-EGO for 400 FE. The only difference between these two algorithms is the infill criteria used. SMS-EGO outperforms ParEGO in all of the objective space (dark areas in **Figure 8.11**, right plot) except in a small region in the upper left side and in the lower right side (dark areas in **Figure 8.11**, left plot). This probably occurs due to the weights in the Tchebycheff aggregation function used by ParEGO which provide a greater exploration power at the edges of the design space. The median EAF of both algorithms provide good approximation sets. The worst EAF can be also seen as the *guaranteed* approximation set. However, the spread between median and best surface in the case of ParEGO could be considered large as compared to other configurations examined in this work. **Figure 8.12** represents the Diff-EAF of ParEGO and SUMO algorithms. The plots depict a superiority of the SUMO algorithm in almost all of the objective space (dark areas in **Figure 8.12**, right plot). Once again, the ParEGO is dominating larger areas towards the edges (as in **Figure 8.11**). In this case the median EAF of SUMO exhibits better performance than ParEGO, dominating a larger space, especially in the middle of the front (where the more balanced, realistic solutions tend to lie). It is interesting to note that all algorithms, exhibit a small weakness to capture the left and right extremes as compared to their best EAF. **Figure 8.13** compares the SMS-EGO and SUMO algorithms. Also, in this configuration both median EAF seems to perform well in the central area of objective space. A minor superiority (grey area values) of SUMO can be observed throughout the objective space. This is also depicted in the performance statistics presented in **Table 8-3**. Nevertheless, both algorithms provide good results considering the small spread between best and median EAF and the small number of FE.



**Figure 8.11** | EAF difference plot for ParEGO and SMS-EGO for 400 function evaluations.



**Figure 8.12** | EAF difference plot for ParEGO and SUMO for 400 function evaluations.



**Figure 8.13** | EAF difference plot for SMS-EGO and SUMO for 400 function evaluations.

It seems clear that all MOSBO algorithms examined in this work manage to yield better results than MOEAs in the case of 200 and 400 FE as underpinned by the statistics (**Table 8-3**) and the CDF plots (all MOSBOs stochastically dominate the MOEAs) of the performance indicators NHVR and  $I_{\epsilon}$ .

### Comparison between unequal budgets

Subsequently, we compare the performance of MOSBOs for 200 and 400 FE against the MOEAs for 1 000 FE. A comparison among the MOEAs is initially performed to indicate the preferred algorithm. Then, both MOEAs (1 000 FE) are compared with all MOSBO algorithms (200 and 400 FE) in terms of NHVR and  $I_{\epsilon}$ . Furthermore, auxiliary Diff-EAF plots are used to infer the performance of the algorithms. **Table 8-5** compares the performance of MOEAs with 1 000 FE budget. In the case of NHVR, SMS-EMOA stochastically dominates NSGAI2 but has a P-value around 0.1 (which is very close to the preferred confidence level), and its median and the average values (**Table 8-6**) are higher than those of the NSGAI2. Hence, in terms of the NHVR indicator the preferred algorithm is the SMS-EMOA. In the case of  $I_{\epsilon}$ , the CDFs of the algorithms cross and hence stochastic domination does not apply. The median and the



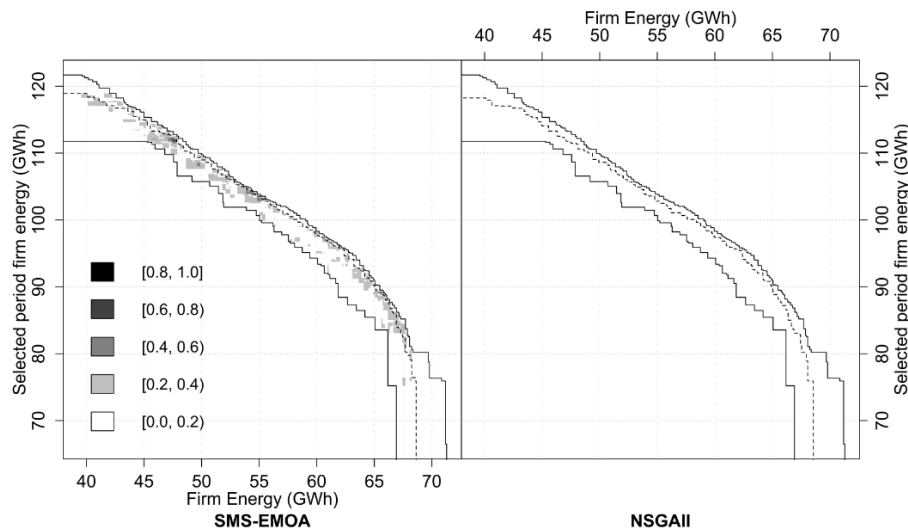
average (Table 8-6) of the SMS-EMOA are lower (better) than NSGAI. Therefore, the SMS-EMOA should also be considered the preferred algorithm for  $I_\epsilon$ . It has to be noted that the P-value in this case is significantly large and indicates high probability of null hypothesis acceptance. This comparison is also visually validated using the Diff-EAF plot (Figure 8.14) where no areas in the objective space exist clearly showing an algorithm extremely outperforming the other.

**Table 8-5** | Comparison of MOEAs for 1 000 function evaluations.

|             | $I_\epsilon$ | NHVR     |
|-------------|--------------|----------|
| Preferred   | SMS-EMOA     | SMS-EMOA |
| Stoch. Dom. | No           | Yes      |
| P-value     | 0.623        | 0.104    |

**Table 8-6** | Results summary for 1 000 function evaluations.

| Budget | Statistic | Hypervolume indicator |       | Epsilon indicator |       |
|--------|-----------|-----------------------|-------|-------------------|-------|
|        |           | SMS-EMOA              | NSGAI | SMS-EMOA          | NSGAI |
| 1000   | Average   | 0.887                 | 0.875 | 0.147             | 0.155 |
|        | Median    | 0.894                 | 0.885 | 0.144             | 0.148 |
|        | St. dev.  | 0.025                 | 0.024 | 0.043             | 0.044 |
|        | Max       | 0.912                 | 0.904 | 0.225             | 0.247 |
|        | Min       | 0.850                 | 0.831 | 0.098             | 0.100 |



**Figure 8.14** | EAF difference plot for SMS-EMOA and NSGAI for 1 000 function evaluations.

Subsequently, in order to compare the performance of MOSBOs (with 200 and 400) and MOEAs with 1 000 FE the following procedure was used: firstly, we check for stochastic dominance (SD) across pairs. If SD applies then the preferred algorithm is the one that dominates the other, and the P-value of the MWU test is calculated to infer significance. If the CDFs of the algorithms cross, then the preferred algorithm is determined by the median, and the P-value of the MWU test is also calculated. Table 8-7 summarizes the results of those comparisons.

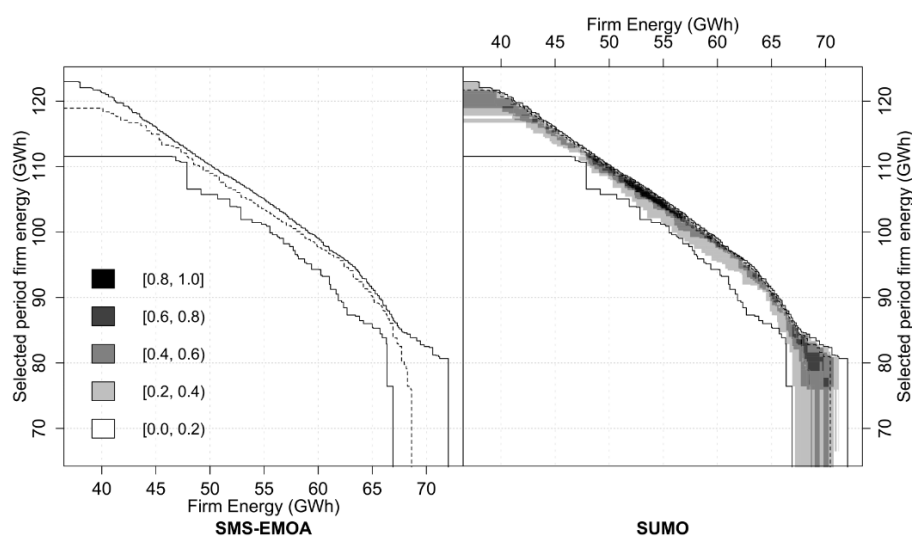
When comparing the ParEGO and SMS-EGO with 400 evaluation functions against the NSGAI with 1 000 FE, the former stochastically dominate the NSGAI with compatible P-values with regards to  $I_\epsilon$ . Interestingly, SUMO does not stochastically dominate NSGAI, yet it demonstrates a lower median and a compatible P-value. This can be explained by the fact that

the CDFs of the algorithms cross at probability 90%, hence this could be an outlier. In the case of the NHVR metric, all MOSBOs stochastically dominate NSGAI and achieve P-values lower than 0.001, outlining their superiority. When comparing the MOSBOs with 400 evaluation functions against the SMS-EMOA with 1 000 FE, the former stochastically dominate, yet with large (incompatible) P-values with regards to the  $I_\epsilon$  metric. The higher P-values indicate also the advantage of SMS-EMOA over NSGAI. Finally, when the NHVR metric is examined all MOSBOs stochastically dominate SMS-EMOA and achieve low P-values ( $<0.001$ ).

**Table 8-7** | Comparison of MOSBO and best MOEA under different budgets.

| Indicator   |             | $I_\epsilon$ |          |          |          | NHVR     |          |          |          |
|-------------|-------------|--------------|----------|----------|----------|----------|----------|----------|----------|
|             |             | 200/1000     |          | 400/1000 |          | 200/1000 |          | 400/1000 |          |
| MOSBO/ MOEA | Algorithm   | NSGAI        | SMS-EMOA | NSGAI    | SMS-EMOA | NSGAI    | SMS-EMOA | NSGAI    | SMS-EMOA |
| SUMO        | Preferred   | SUMO         | SUMO     | SUMO     | SUMO     | SUMO     | SUMO     | SUMO     | SUMO     |
|             | Stoch. Dom. | No           | No       | No       | No       | Yes      | Yes      | Yes      | Yes      |
|             | P-value     | 0.212        | 0.427    | 0.021    | 0.045    | 0.045    | 0.121    | $<0.001$ | $<0.001$ |
| SMS-EGO     | Preferred   | SMS-EGO      | SMS-EGO  | SMS-EGO  | SMS-EGO  | SMS-EGO  | SMS-EGO  | SMS-EGO  | SMS-EGO  |
|             | Stoch. Dom. | Yes          | Yes      | Yes      | Yes      | Yes      | Yes      | Yes      | Yes      |
|             | P-value     | 0.307        | 0.520    | 0.090    | 0.273    | 0.014    | 0.185    | $<0.001$ | $<0.001$ |
| ParEGO      | Preferred   | ParEGO       | ParEGO   | ParEGO   | ParEGO   | ParEGO   | ParEGO   | ParEGO   | ParEGO   |
|             | Stoch. Dom. | Yes          | Yes      | Yes      | Yes      | Yes      | No       | Yes      | Yes      |
|             | P-value     | 0.007        | 0.053    | 0.004    | 0.038    | 0.006    | 0.345    | $<0.001$ | $<0.001$ |

For the sake of limited space differential empirical attainment function (Diff-EAF) plots are going to be presented only for SUMO with 400 FE compared to SMS-EMOA with 1 000 FE (**Figure 8.15**) since SUMO was the preferred algorithm for 400 FE (**Table 8-4**) and SMS-EMOA the preferred MOEA for 1 000 FE (see **Table 8-5**). **Figure 8.15** depicts that SUMO is able to dominate SMS-EMOA in all the objective space and with high probability values (dark grey areas). This inference is also validated by the results shown in **Table 8-3**, **Table 8-6** and **Table 8-7**.



**Figure 8.15** | EAF difference plot for SMS-EMOA and SUMO for different computational budget: 1 000 and 400 function evaluations respectively.

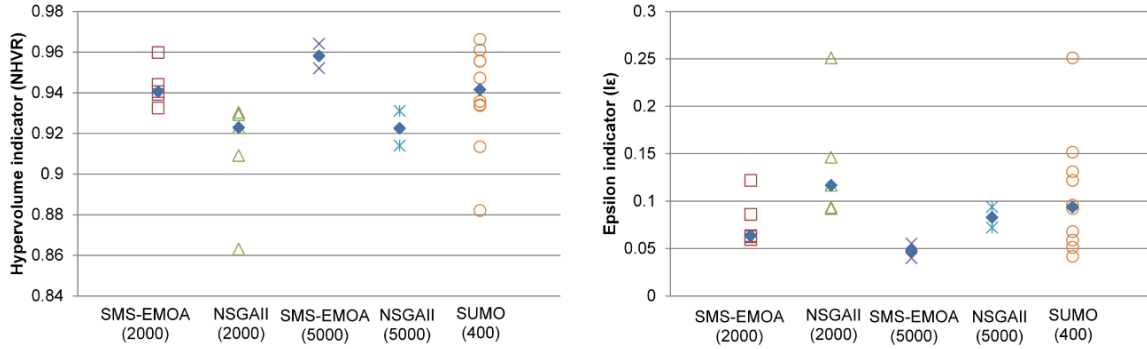
To sum up, the performance of MOSBOs for 400 FE yielded better results in terms of performance indicators when compared with MOEAs for much higher FE (1 000). For example, SUMO achieves better results in both NHVR and  $I_{\epsilon}$  than SMS-EMOA using 60% lower computational budget (as visually confirmed in the Diff-EAF plot). It has to be noted that even when comparing the performance of MOSBOs and MOEAs with 200 and 1 000 FE budget respectively (a significant difference in terms of computational burden) the MOSBOs again yield better results (**Table 8-3**, **Table 8-6** and **Table 8-7**). Yet, it should be clear that the selection of the benchmarking algorithm is critical [*Razavi et al., 2012b*] to any discussion on issues of superiority or not of an algorithm due to impact on relative performance.

Finally, as discussed before, MOEAs were also run for computational budgets of 2 000 and 5 000 FE. For those budgets 5 and 2 replicates has been generated respectively. More replicates would be preferable, but this was not possible due to the high computational time needed for the simulation: one simulation (function evaluation) in WEAP21 required 90 sec. Therefore, in these cases the stochastic dominance and statistical test cannot be used. The performance assessment is only based on statistics presented in **Table 8-3** and **Table 8-8**.

When comparing MOSBOs (with 200 FE) against the MOEAs (with 2 000 and 5 000 FE), the latter outperformed the former in terms of the median value. In the case of the NSGAI the differences are small: MOSBOs NHVR median values range from 0.899-0.911 as compared to the 0.920 and 0.922 NSGAI values (for the 2 000 and 5 000 respectively); MOSBOs  $I_{\epsilon}$  median values range from 0.12-0.13 as compared to the 0.117 and 0.083 NSGAI values (for the 2 000 and 5 000 respectively). When comparing MOSBOs with the SMS-EMOA these differences are larger (SMS-EMOA achieves median NHVR 0.941/0.958, and median  $I_{\epsilon}$  0.063/0.047, for 2 000 and 5 000 FE respectively).

Finally, when comparing MOSBOs (with 400 FE) against the MOEAs (with 2 000 and 5 000 FE). ParEGO, SMS-EGO and SUMO yield median NHVR 0.916, 0.926, 0.941 and median  $I_{\epsilon}$  0.111, 0.115, 0.093 respectively. In this case the performance of median NHVR of NSGAI for 2 000 and 5 000 FE is outperformed by SMS-EGO and SUMO, though ParEGO is close behind. When comparing the  $I_{\epsilon}$  the MOSBOs manage to outperform NSGAI only for 2 000 FE with very small differences. As compared to SMS-EMOA however, only SUMO manages to compete and only in terms of NHVR. In the case of median NHVR, SMS-EMOA achieves better results than ParEGO and SMS-EGO and equal to SUMO. This is, however, not the case when comparing the median of  $I_{\epsilon}$ . In this case the differences are larger and all MOSBO are outperformed by SMS-EMOA. **Figure 8.16** summarizes the results and depicts the performance of SUMO with a budget of 400 FE relative to MOEAs with larger budgets.

From these results it is inferred that MOSBOs with 400 FE depict similar performance to MOEAs with 2 000 FE. Thus, the efficiency of MOSBOs presents an upper limit of approximately 80% ( $=2\ 000-400/2\ 000$ ).



**Figure 8.16** | Comparison of SUMO with MOEA for 2 000 and 5 000 function evaluations. Blue diamond represents the median of each column. In left panel (NHVR) higher values are preferred. In the right panel (Iε) lower values are preferred.

**Table 8-8** | Results summary for 2 000 and 5 000 function evaluations.

| Budget | Statistic | Hypervolume indicator |       | Epsilon indicator |       |
|--------|-----------|-----------------------|-------|-------------------|-------|
|        |           | SMS-EMOA              | NSGAI | SMS-EMOA          | NSGAI |
| 2000   | Average   | 0.943                 | 0.911 | 0.079             | 0.140 |
|        | Median    | 0.941                 | 0.920 | 0.063             | 0.117 |
|        | Std       | 0.010                 | 0.028 | 0.026             | 0.066 |
|        | Max       | 0.960                 | 0.930 | 0.122             | 0.251 |
|        | Min       | 0.932                 | 0.863 | 0.059             | 0.092 |
| 5000   | Average   | 0.958                 | 0.922 | 0.047             | 0.083 |
|        | Median    | 0.958                 | 0.922 | 0.047             | 0.083 |
|        | Std       | 0.009                 | 0.012 | 0.011             | 0.015 |
|        | Max       | 0.964                 | 0.931 | 0.055             | 0.094 |
|        | Min       | 0.952                 | 0.914 | 0.040             | 0.072 |

**8.6.2 Results for the case study**

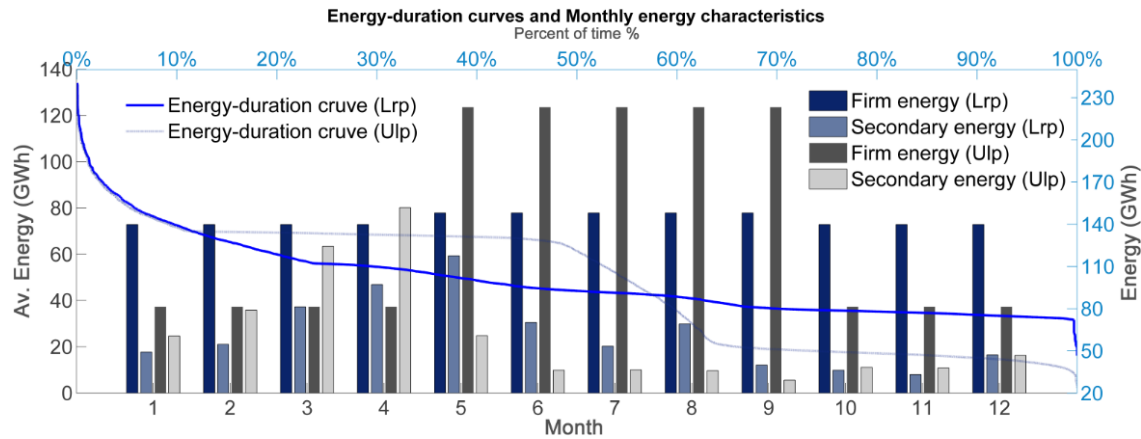
The MOSBO algorithms, after demonstrating their adequate performance through the benchmarking exercise discussed above, were applied for the optimization of the Nestos hydrosystem. **Figure 8.17** illustrates the results in terms of the optimization of the energy-related operation rules. The figure presents results from the upper left (Ulp) and the lower right points (Lrp) of the best Pareto front. **Table 8-9** presents the variables and objective function values from these operation rules. The operation rules are in simple form and can be easily translated to real decisions and thus applied by the reservoir system operators. For instance,  $\theta_1$  in Lrp is translated to 52.42 GWh set as a target of energy production for the period May-Sept and  $52.42 \times 0.94=49.29$  GWh for Oct-Apr.

**Table 8-9** | Operation rules Ulp and Lrp.

| Operational rule | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\beta$ | Obj. 1 | Obj. 2 |
|------------------|------------|------------|------------|---------|--------|--------|
| Ulp              | 28.43      | 10.80      | 8.27       | 0.00    | 37.24  | 123.59 |
| Lrp              | 52.42      | 11.57      | 2.74       | 0.94    | 72.85  | 77.84  |

**Figure 8.17** depicts the rationale behind the objective functions used in this work. The average monthly energy is distinguished between firm and secondary energy. Firm energy is the guaranteed energy and secondary is the excess of energy generated each month. The simulation of Lrp illustrates constant firm energy each month, as well as some excess of secondary energy. This is also depicted in the energy-duration curve of Lrp where the energy exceeds 72 GWh for

99% of time. In the case of Ulp the firm energy for the selected time period (99% reliability) is considerably greater, close to 123 GWh with 37 GWh guaranteed energy for the remaining months. This is also shown in the corresponding energy-duration curve where there is a leap for about 50% of the time. Such behavior was anticipated because Lrp and Ulp favor objective functions 1 and 2, respectively; therefore, they advocate two contradictory operation policies.



**Figure 8.17** | Energy-duration curves and monthly energy characteristics of the case study for the upper left point (Ulp) and Lower right point (Lrp) of the best Pareto front.

## 8.7 SUMMARY

This Chapter presented a multi-objective version of the parameterization-simulation-optimization (PSO) framework which is able to incorporate hydrological uncertainty (through stochastic simulation) and thus develop uncertainty-aware reservoir operation rules. The multi-objective version of PSO is able to handle multiple and conflicting criteria. Hence, it represents different and conflicting operational policies without the need for decision makers to *a priori* express their preferences. Visualizing the objectives' tradeoffs is particularly useful and can be used as a negotiation tool for further decision making between different stakeholders in reservoir management (e.g., energy and water companies, farmers and municipalities).

Furthermore, we presented a comparative study of the potential for using multi-objective surrogate-based optimization methods (MOSBOs) in multi-reservoir management, by benchmarking three MOSBOs against two well-known MOEAs. Different experiments with regards to computational budget were investigated yielding very promising results for the MOSBOs. Results suggest that MOSBOs are able to adequately reproduce Pareto fronts, without significant damage to the problem detail or their ability to identify robust, (uncertainty-aware in the context described earlier) operational policies. Furthermore, we have demonstrated the potential of MOSBO algorithms to perform well even under significant computational budget restrictions (of as few as 200-400 function evaluations) comparing favorably to more *standard* MOEAs with higher computational budget (e.g. 1 000-2 000 function evaluations), achieving efficiencies of 60-80% in terms of computational time for the same (or even better) results. This indicates their usefulness in addressing realistic problems where, as in the case of multi-reservoir operation management, a key barrier to properly incorporating uncertainty is the excessive computational burden. If MOSBOs are indeed able to alleviate much of this burden, as suggested in this Chapter, they have a significant role to play in guaranteeing reliability in real-world applications within a highly uncertain climatic and socio-economic context. An input uncertainty, that, in an arguably non-Gaussian world,

can potentially be better captured using the novel models and schemes of Chapter 4-7. In this vein, and motivated by these encouraging findings, in Chapter 9 we step forward, and conclude the research developments of this Thesis, by introducing a novel surrogate-enhanced global optimization algorithm, that combines both, surrogate-based modelling approaches, as well as different, global and local, optimization approaches (evolutionary search, simulated annealing, downhill simplex), into a single algorithm.

## SURROGATE-ENHANCED EVOLUTIONARY ANNEALING SIMPLEX ALGORITHM FOR EFFECTIVE OPTIMIZATION OF WATER REOURCES PROBLEMS ON A BUDGET \*

---

### PREAMBLE

This Chapter is motivated by the promising findings of Chapter 8 and the fact that typical water resources optimization problems involve an objective function that presumes the use of a simulation model and the subsequent evaluation of its outputs. Long simulation times, which may arise due to uncertainty incorporation or time *expensive* simulation models, may pose significant barriers to the procedure. Often, to obtain a solution within a reasonable time, the user has to substantially restrict the allowable number of function evaluations, thus terminating the search much earlier than required. As shown earlier, a promising strategy to address these shortcomings is the use of surrogate modelling techniques. Here we introduce the Surrogate-Enhanced Evolutionary Annealing-Simplex (SEEAS) algorithm that couples the strengths of surrogate modelling with the effectiveness of the evolutionary annealing-simplex method. SEEAS combines three different optimization approaches (evolutionary search, simulated annealing, downhill simplex). Its performance is benchmarked against other surrogate-based algorithms in several test functions and two water resources applications (model calibration, reservoir management). Results reveal the significant potential of using SEEAS in challenging optimization problems on a budget.

This Chapter is structured as follows: section 9.1, through a literature, provides an introduction to the problem of time *expensive* optimization methods. Section 9.2 describes the proposed algorithm. Section 9.3 concerns the established benchmarking protocol, devised to assess the performance of the algorithms. Section 9.4 to 9.6, describe and present the benchmarking results based on three distinct problems, i.e., mathematical test functions (section 9.4), a hydrological calibration problem (section 9.5), and a multi-reservoir management problem (section 9.6). Finally, section 9.7 summarizes the key findings of this Chapter.

---

\* Based on:

Tsoukalas, I., P. Kossieris, A. Efstratiadis, and C. Makropoulos (2016), Surrogate-enhanced evolutionary annealing simplex algorithm for effective and efficient optimization of water resources problems on a budget, *Environ. Model. Softw.*, 77, 122–142, doi:10.1016/j.envsoft.2015.12.008.

## 9.1 INTRODUCTION

This Chapter introduces the Surrogate-Enhanced Evolutionary Annealing-Simplex algorithm (SEEAS), which is a novel global surrogate-based optimization (SBO) method, focused on time-expensive functions. Our motivation arises from challenging simulation-optimization problems that are commonly found in water resources problems (see section 1.1 and 1.3), and they impose, in the everyday practice, very limited computational budgets, e.g., of few hundred function evaluations. SEEAS has been designed for typical hydrological optimization problems, i.e., decision-making and calibration, suffering from different peculiarities and complexities, which are in turn reflected in the different geometry of the associated response surfaces.

For convenience, we consider that all criteria are aggregated in a single objective function representing a global performance measure of the system (an alternative approach would require the formulation of a multiobjective function and the identification of acceptable tradeoffs among conflicting criteria, which is not the case here). We also assume that all *internal* constraints (i.e., constraints associated with the system dynamics) are handled through the simulation model [e.g., *Koutsogiannis and Economou, 2003*], while any additional *external* constraints, which are usually associated with decision-making problems, are embedded in the objective function, typically as penalty terms. Under this premise, the combined simulation-optimization problem is formalized as the determination of the global optimum (for convenience, minimum) of a nonlinear objective function  $f(\mathbf{x})$ , where  $f(\cdot)$  represents the simulation model and  $\mathbf{x}$  is the vector of control variables. The search space is a hypervolume, since the unique constraints of the problem are the lower and upper bounds of parameters. As  $f(\mathbf{x})$  is a black-box function, its analytical expression as well as its derivatives are not available, which prohibits the use of gradient-based optimization. Given also that, due to uncertainties and complexities of the system,  $f(\mathbf{x})$  is non-convex, and thus multimodal (i.e., it contains multiple local optima), derivative-free methods [e.g., *Rios and Sahinidis, 2013*] combined with stochastic search approaches are essential to solve this so-called global optimization problem.

Surrogate-based optimization (SBO) methods, often termed response surface approaches go back to 70's [*Blanning, 1975*], have been popularized since the pioneering work by *Jones et al. [1998]*, who developed the Efficient Global Optimization (EGO) algorithm. EGO uses Kriging as surrogate model and an acquisition function (named Expected Improvement), in order to locate potential good samples that should be evaluated through expensive simulation functions [*Sacks et al., 1989; Jones et al., 1998*]. Later, *Sasena et al. [2002]* implemented and investigated various acquisition functions for EGO. Literature also reports multi-objective versions of EGO [e.g., *Knowles, 2005; Ponweiser et al., 2008; Couckuyt et al., 2013*].

Beyond Kriging, other commonly used surrogate models are Radial Basis Functions [*RBFs - Powell, 1992; Buhmann, 2003*], polynomials [*Myers and Montgomery, 1995*], artificial neural networks, and support vector machines [*Cortes and Vapnik, 1995; Dibike et al., 2001*]. The use RBFs within the context of evolutionary algorithms was popularized after the publication *Regis and Shoemaker [2004]*. Other typical examples of RBFs are the Multistart Local Metric Stochastic RBF (MLMSRBF) and the ConstrLMSRBF, which handles inequality constraints [*Regis and Shoemaker, 2007a; Regis, 2011*]. Additionally, *Regis [2014]* and *Tang et al. [2012]* proposed hybridizations of the particle swarm optimization algorithm [*Kennedy and Eberhart, 1995*] that use RBFs to assist the search. *Shoemaker et al. [2007]* developed an evolutionary algorithm that uses an RBF approximation and benchmarked its performance against several test problems, with dimensions ranging from 8-D to 14-D. Finally, *Regis and Shoemaker [2013]* developed the DYnamic COordinate Search (DYCORS) that uses Response Surface models to



handle high-dimensional expensive optimization problems. DYCORS was benchmarked against other RBF-based algorithms in a variety of test problems, ranging from 14-D to 200-D.

Comprehensive reviews of surrogate-based optimization methods can be found in the broader optimization literature [e.g., *Jin, 2005, 2011; Forrester and Keane, 2009*], while *Razavi et al. [2012b]* summarize the use of surrogate modelling techniques in water resource systems, also classifying the existing meta-modelling frameworks. The literature reports several successful applications in time-demanding hydrological problems [e.g., *Broad et al., 2005; Mugunthan et al., 2005; Mugunthan and Shoemaker, 2006; Zou et al., 2007; Regis and Shoemaker, 2007b; Kourakos and Mantoglou, 2009; Drosou et al., 2015; Tsoukalas et al., 2015a, 2015b; Kossieris et al., 2015; Tsoukalas and Makropoulos, 2015b, 2015a; Tegos et al., 2017*], highlighting their potential to alleviate the computational burden that accompanies simulation-optimization problems.

SEEAS is built upon the Evolutionary Annealing-Simplex (EAS) method [*Efstratiadis and Koutsoyiannis, 2002*], which is a hybrid scheme combining global and local search strategies and assisted by a RBF surrogate model. SEEAS uses an external archive to maintain all visited solutions in order to formulate, update and exploit the surrogate model during search. There are also some improvements in the key core of EAS, regarding the simplex transitions and the mutation operator. SEEAS is compared and benchmarked against the original version of EAS and three state-of-the-art optimization algorithms (see section 9.3.3). Namely, the Dynamic Dimension Search (DDS) [*Tolson and Shoemaker, 2007*], the MLMSRBF [*Regis and Shoemaker, 2007b*], and DYCORS [*Regis and Shoemaker, 2013*]. Evaluations are made on the basis of 12 mathematical problems (i.e., six test functions for two alternative dimensions, 15-D and 30-D), a hydrological calibration problem with 11 parameters, configured with both real and synthetic data, and a multi-reservoir management problem with 20 decision variables, using synthetic inflows of 500 years length. The use of synthetic data is one of the novelties of our testing framework. Moreover, most of the known surrogate-based schemes have been only evaluated in calibration problems and not in time-demanding water management applications, with few exceptions [e.g., *Razavi et al., 2012b; Tsoukalas and Makropoulos, 2015a, 2015b*]. The results of this extended analysis are very encouraging, since the proposed method is effective and efficient, in terms of locating a satisfactory solution as close as possible to the global optimum, within reasonable computational time, and outperforms the other examined approaches, in almost all tests.

## 9.2 OPTIMIZATION METHODOLOGY

### 9.2.1 Evolutionary Annealing-Simplex

EAS<sup>s</sup> is a heuristic, population-based global optimization technique, originally developed by *Efstratiadis and Koutsoyiannis [2002]*, that couples the strength of simulated annealing in rough search spaces along with the efficiency of the downhill simplex method [*Nelder and Mead, 1965*] in smoother spaces. Its key idea is the introduction of an external variable  $T$ , which plays a role similar to temperature in a real-world annealing process and determines the degree of randomness of the search procedure. This is expressed through a stochastic term that is relative to temperature and is added to the initial objective function  $f(\mathbf{x})$ , thus getting a modified function  $g(\mathbf{x}) = f(\mathbf{x}) \pm \mathbf{u}T$  (where  $\mathbf{u}$  is a vector of uniformly distributed random numbers).

---

<sup>s</sup> EAS and SEEAS are available online at: <http://www.itia.ntua.gr/en/softinfo/29/>

Search is based on an evolving population of feasible points, where critical decisions are driven by the modified function. The genetic operators are either quasi-stochastic geometric transformations, inspired by the downhill simplex method, or fully-probabilistic transitions (mutations). As search proceeds, the system temperature reduces according to an adaptive annealing cooling schedule, and all transitions become more deterministic.

EAS has been successfully employed in several hydrological applications [e.g., *Rozos et al., 2004; Nalbantis et al., 2011; Kossieris et al., 2013; Efstratiadis et al., 2014b*]. It has been also incorporated within advanced modelling tools, i.e., Hydronomeas [*Efstratiadis et al., 2004*], Hydrogeios [*Efstratiadis et al., 2008*] and HyetosR [*Kossieris et al., 2016*] to solve challenging simulation-optimization problems. The original algorithm has been also adapted to handle multiobjective problems [*Efstratiadis and Koutsoyiannis, 2008*] and stochastic (i.e., noisy) objective functions [*Kossieris et al., 2013*]. Here we introduce an improved version of EAS, called Surrogate-Enhanced Evolutionary Annealing-Simplex (SEEAS) algorithm, which is presented in detail herein.

## 9.2.2 Surrogate-Enhanced Evolutionary Annealing-Simplex

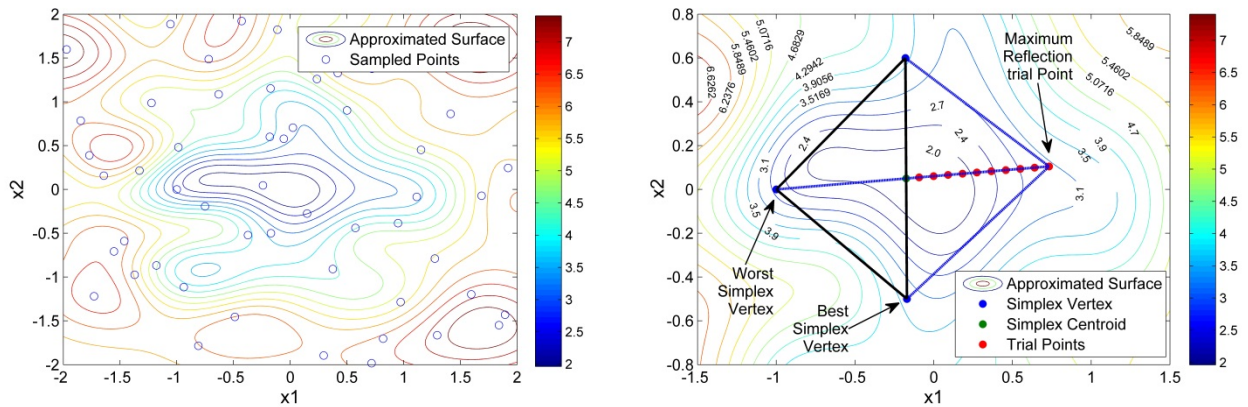
### 9.2.2.1 Overview of SEEAS algorithm

The algorithm is a surrogate-enhanced extension of EAS in a way that builds, maintains and exploits surrogate modelling (SM) techniques that generate approximated response surfaces, which allow effectively guiding search towards promising areas of the real response surface. The model used is the RBF, which is a well-known interpolation technique (**Figure 9.1**, left). During the iterative procedure, the algorithm maintains an external archive of all visited points, already evaluated through the (expensive) objective function. This archive is used to update the SM, in an attempt to progressively provide more accurate approximations of the current region of interest (i.e. the area around the current best point). In SEEAS, the surrogate model has a double role. The first is providing new points that are added to the current population, and the second is assisting the genetic operators of the downhill simplex scheme to identify suitable directions across the search space (e.g., favorable slopes and new areas of attraction).

In order to balance exploration (i.e., detailed sampling) and exploitation (i.e., blind use of SM), SEEAS uses a weighted metric, termed acquisition function (AF), which accounts for the predictions provided by the SM as well as the spread of all previously evaluated points (by means of a distance quantity). In opposite to common practices that use a standard expression of the AF with constant weights, in our approach the weights are dynamically adjusted, thus improving the efficiency of the algorithm. Details about the acquisition function (AF) are given in Section **9.2.2.3**.

SEEAS follows an iterative search procedure. At the end of each iteration cycle (or generation, according to the terminology of evolutionary theory), we obtain at least one new point that enters the population and replaces one of its existing members. A typical iteration cycle of SEEAS starts by fitting the surrogate model to the current population (initially, this population is randomly generated through Latin Hypercube Sampling, LHS). Next, we run an internal global optimization algorithm (particularly, the original version of EAS) across the surrogate response surface, using as objective the acquisition function (AF), in order to locate a candidate solution to enter the population (provided that this solution outperforms the current worst point). Thereafter, we follow a search procedure that is mostly based on the genetic operators of EAS, enhanced by surrogate-assisted steps in simplex-based transformations.

The general idea is to utilize the information gained by the SM, in order to enhance the current knowledge in the selection of simplex transitions. A characteristic example involving the reflection step is illustrated in **Figure 9.1**, right (for simplicity, we demonstrate the predictions of the surrogate model and not the AF). In the original version of EAS, after specifying the direction of reflection (defined by the difference between the worst vertex of the simplex and the centroid of all rest vertices), the algorithm employs a blind trial-and-error procedure, i.e., it generates subsequent random points along this direction and evolves according to their values. In this scheme, the original objective function is called whenever a new trial point is generated. Since the expansion continues as long as the function value improves, this procedure may be quite expensive, in terms of function evaluations. In opposite, in SEEAS we employ a candidate screening procedure using the SM, which allows making multiple trials with negligible computational cost and guiding search using all prior information. Similar screening is employed within all simplex transformations (except shrinkage), thus providing significant aid to the associated decisions.



**Figure 9.1** | Approximated surface (RBF) in a 2-D example (Ackley function) using all available sample points (left panel). The right panel demonstrates a randomly selected simplex and the modified surrogate-enhanced reflection movement using candidate points on the line formed from the simplex centroid and the maximum reflection point. The simplex is reflected at the candidate point with the minimum function value.

### 9.2.2.2 Surrogate model (RBF)

SEEAS implements the Radial Basis Function (RBF) interpolation method [Powell, 1992; Buhmann, 2003], and more specifically the RBF with cubic basis functions and linear polynomial tail. This is a commonly used surrogate model of proven effectiveness, as reported in numerous studies [e.g., Mugunthan et al., 2005; Regis and Shoemaker, 2007b, 2007a; Shoemaker et al., 2007; Müller and Shoemaker, 2014].

The computational procedure of RBF is the following. Given  $N_s$  samples  $\mathbf{x} \in R^n$  with response  $y$ , we get the pairs  $(\mathbf{x}_i, y_i)$ . The prediction  $s(\mathbf{x})$  of RBF model at sample point  $\mathbf{x}$  is given by:

$$s(\mathbf{x}) = \sum_{i=1}^{N_s} \lambda_i \varphi(\|\mathbf{x} - \mathbf{x}_i\|) + p(\mathbf{x}) \quad (9.1)$$

where  $\lambda_i \in R$ ,  $\varphi$  is a basis function of the form  $\varphi(r) = r^3$ ,  $\|\cdot\|$  is the Euclidean distance (norm) and  $p(\mathbf{x})$  is a polynomial tail of the form  $p(\mathbf{x}) = \mathbf{b}^T \mathbf{x} + a$ , where  $\mathbf{b} = (b_1, \dots, b_n)^T$  and  $a \in R$ . The model parameters  $\lambda$ ,  $\mathbf{b}$ , and  $a$  are determined by solving the linear system:

$$\begin{bmatrix} \Phi & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \quad (9.2)$$

where  $\Phi$  is an  $N_s \times N_s$  matrix with elements  $\varphi_{ij} = \varphi(\|\mathbf{x}_i - \mathbf{x}_j\|)$ ,  $\mathbf{P}$  is a  $N_s \times (n + 1)$  matrix, the  $i^{\text{th}}$  row of which is  $(1, \mathbf{x}_i^T)$ ,  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{N_s})^T$ ,  $\mathbf{c} = (b_1, \dots, b_n, a)^T$ , and  $\mathbf{y} = (y_1, \dots, y_{N_s})^T$ . We mention that the matrix of Eq. (9.2) is invertible if and only if  $\text{Rank}(\mathbf{P}) = n + 1$  [Powell, 1992].

### 9.2.2.3 Acquisition function

Acquisition functions (AF) are well-established techniques, aiming to balance exploration-exploitation in surrogate-based optimization algorithms [e.g., Sasena et al., 2002; Forrester and Keane, 2009]. SEEAS implements a novel scheme, in which the weights are automatically adjusted during the iterative process, according to the current number of function evaluations and the maximum allowed number of evaluations.

Consider a set of  $N_s$  points,  $\mathbf{x}_s^j$ , with known response value,  $f(\mathbf{x}_s^j)$ , and another set of  $N_c$  points  $\mathbf{x}_c^i$ , with approximated response values  $s(\mathbf{x}_c^i)$ . These are conventionally called candidate points, in the sense that they are used within infilling or internal search procedures, e.g., selection of the most appropriate reflection point in the graphical example of Figure 9.1. The acquisition function is estimated as follows:

**Step A.** Standardize the approximated response values of all candidate solutions by setting  $s^*(\mathbf{x}_c^i) = [s(\mathbf{x}_c^i) - s^{\min}] / [s^{\max} - s^{\min}]$ , where  $s^{\min}$  and  $s^{\max}$  are the corresponding minimum and maximum values.

**Step B.** Calculate the minimum Euclidean distance of each candidate point  $\mathbf{x}_c^i$  from all previously evaluated points,  $\mathbf{x}_s^j$ , i.e.,  $d_i = d^*(\mathbf{x}_c^i) = \min_{1 \leq j \leq N_s} \|\mathbf{x}_c^i - \mathbf{x}_s^j\|$ , and standardize them by setting  $d_i^* = (d_i - d^{\min}) / (d^{\max} - d^{\min})$ , where  $d^{\min}$  and  $d^{\max}$  are the corresponding minimum and maximum distances.

**Step C.** Calculate the weighted value of AF for every candidate point using the formula:

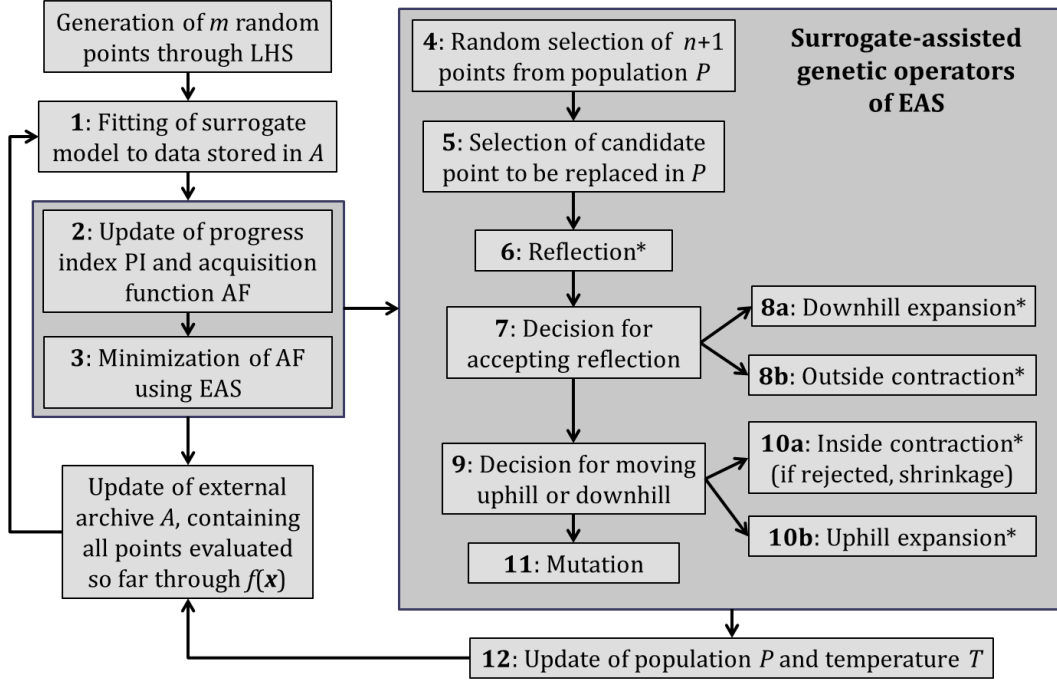
$$\text{AF}_i = w s^*(\mathbf{x}_c^i) + (1 - w) d^*(\mathbf{x}_c^i) \quad (9.3)$$

where  $w$  is a dimensionless weighting coefficient, ensuring balance between exploitation and exploration. To finalize the infilling routine, the candidate with the minimum AF value will be selected and assessed through the objective function. As mentioned before, the minimization of the AF across the surrogate search space is carried out through the original EAS algorithm.

### 9.2.2.4 Detailed description of SEEAS

Let  $f(\mathbf{x})$  be a nonlinear objective function in the feasible space  $\mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U$ , where  $\mathbf{x}$  is an  $n$ -dimensional vector of continuous control variables (in practice,  $f(\mathbf{x})$  represents the performance measure of a simulation model). For convenience, we search for the global minimum of  $f(\mathbf{x})$ , allowing a budget of MFE function evaluations. The algorithm uses two archives. The first is the population  $P^{[t]}$ , which is evolved during the search procedure (where  $t$  denotes the iteration cycle or generation), and the second is the so-called external archive  $A^{[t]}$ , which contains all visited points from the beginning of the optimization ( $t = 0$ ), including the members of the current population. Whenever a new point  $\mathbf{x}$  is evaluated through the objective function  $f(\mathbf{x})$ , it enters the archive  $A^{[t]}$  (the archive may be updated several times within a generation). At the beginning of each new generation  $t$ , the surrogate model is re-evaluated by

considering the current elements of  $A^{[t]}$ . The size of the population is  $m \geq n + 1$  (i.e., the minimum number of points required to fit a RBF with linear polynomial as well as to formulate a simplex in the  $n$ -dimensional space), and remains constant, while the size of the external archive progressively increases, thus ensuring more accurate approximations of the response surface and, consequently, more reliable predictions. The initial population  $P^{[0]}$  is generated via the Latin Hypercube Sampling (LHS) technique, which ensures satisfactory spread across the feasible space [Giunta et al., 2003]. Apparently, the initial archive  $A^{[0]}$  is identical to  $P^{[0]}$ .



**Figure 9.2** | Outline of SEEAS algorithm following the steps explained in section 9.2.2.4 (\* denotes the use of the surrogate model within the associated simplex transformations).

Similarly to EAS, the surrogate-enhanced algorithm also uses an auxiliary parameter,  $T^{[t]}$ , called temperature. The concept originates from simulated annealing, where the key role of temperature is ensuring balance between randomness and determinism. In SEEAS, temperature is dynamically adjusted (i.e., reduced) using empirical rules, considering the extreme values,  $f_{\min}^{[t]}$  and  $f_{\max}^{[t]}$ , of the current population  $P^{[t]}$ , and a dimensionless progress index, defined as,

$$PI = \log(FE) / \log(MFE) \quad (9.4)$$

where FE is the current number of function evaluations and MFE is the maximum allowable number of FE, which is a user-specified termination criterion.

A typical iteration cycle of SEEAS, an outline of which is illustrated in Figure 9.2, comprises the following steps (generation index  $t$  is omitted for simplicity):

**Step 1.** The interpolation surface  $s(\mathbf{x})$  is updated using the current information stored in the external archive  $A$  (i.e., all points evaluated so far through the original objective function).

**Step 2.** The weighting coefficient of the AF is updated using the empirical formula:

$$w = \max(0.75, \min(PI, 0.95)) \quad (9.5)$$

The above formula ensures that at the early stages of optimization, more weight is given to exploration (up to 0.25), but gradually its contribution diminishes thus not exceeding 0.05.

**Step 3.** A new point  $\mathbf{x}_p$  is generated by minimizing AF, using the original version of EAS for internal optimization. The new point is evaluated through  $f(\mathbf{x})$  and replaces the worst point of the current population, if the latter is worse (higher) than  $f(\mathbf{x}_p)$ .

**Step 4.** A set of  $n + 1$  points is randomly selected from the current population, in order to formulate the vertices of a simplex in the  $n$ -dimensional search space, symbolized  $S = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1}]$ . The elements of  $S$  are sorted such as  $f(\mathbf{x}_1)$  corresponds to the best (lowest) and  $f(\mathbf{x}_{n+1})$  to the worst value of the objective function.

**Step 5.** From the subset  $[\mathbf{x}_2, \dots, \mathbf{x}_{n+1}]$  we select a candidate point  $\mathbf{x}_w$  to be replaced in the population, based on the modified, quasi-stochastic objective function:

$$g(\mathbf{x}) = f(\mathbf{x}) + uT \quad (9.6)$$

where  $u$  is a uniform random number in the interval  $[0, 1]$ . By adding the stochastic component  $uT$  to the objective function  $f(\mathbf{x})$ , the algorithm behaves as in between random and downhill search. At the early stages of optimization, when temperature is still high, any point except for the best one can be replaced. On the other hand, in the limiting case  $T \rightarrow 0$ , the actually worst point, i.e.,  $\mathbf{x}_{n+1}$ , is replaced, as considered in the original downhill simplex method.

**Step 6.** A set of  $N_r$  trial points  $\mathbf{x}_{cr}^k$  are generated by reflecting the simplex according the formula:

$$\mathbf{x}_{cr}^k = \mathbf{g} + (0.5 + \delta k) (\mathbf{g} - \mathbf{x}_w) \quad (9.7)$$

where  $\mathbf{g}$  is the centroid of the subset  $[\mathbf{x}_2, \dots, \mathbf{x}_{n+1}]$  and  $\delta k$  is a scale coefficient equally spread in the interval  $[0, 1]$ , thus  $\delta k = (k - 1)/(N_r - 1)$ , for  $k = 1, \dots, N_r$ . Among all candidates, we select the one that minimizes AF, which we will next call the reflection point,  $\mathbf{x}_r$ . The reflection point is evaluated on the basis of the objective function and enters the external archive.

**Step 7.** If  $f(\mathbf{x}_r) < f(\mathbf{x}_w)$ , we replace  $\mathbf{x}_w$  by  $\mathbf{x}_r$  in the population and move to steps 8a or 8b, according to the outcome of its comparison with the current best vertex, i.e.,  $f(\mathbf{x}_r) < f(\mathbf{x}_1)$ . Otherwise, we move to step 9, to decide whether  $\mathbf{x}_r$  should be accepted or withdrawn, thus seeking another candidate.

**Step 8a.** If  $f(\mathbf{x}_r) < f(\mathbf{x}_1)$ , the vector  $\mathbf{x}_r - \mathbf{x}_1$  defines a direction of minimization. We remark that the detection of downhill slopes in high-dimensional spaces of complex geometry is not an often case. This makes essential to take advantage in order to accelerate the search procedure, by employing a sequence of  $N_e$  trial expansion steps through the recursive formula:

$$\mathbf{x}_{ce}^k = \mathbf{g} + \delta_k (\mathbf{x}_r - \mathbf{g}) \quad (9.8)$$

where  $\delta_k$  is a scale coefficient given by  $\delta_k = \delta_{k-1} + (k - 1)/(N_e - 1)$ , for  $k = 1, \dots, N_e$ . The expansion continues as long as the AF value is improved (or until reaching the bounds of the feasible space). The optimal (in terms of AF) trial point,  $\mathbf{x}_e$ , is kept in the external archive and replaces  $\mathbf{x}_r$  in the current population, provided that  $f(\mathbf{x}_e) < f(\mathbf{x}_r)$ . In that case, the algorithm moves to step 12 to finalize the cycle.

**Step 8b.** If  $f(\mathbf{x}_r) > f(\mathbf{x}_1)$ , we attempt detecting a promising solution in the neighborhood of  $\mathbf{x}_1$ , by employing  $N_c$  trial contractions of the simplex in the interval between the centroid and the reflection point, according to the formula:

$$\mathbf{x}_{cc}^k = \mathbf{g} + (0.25 + 0.5\delta_k) (\mathbf{x}_r - \mathbf{g}) \quad (9.9)$$

where  $\delta_k = (k - 1)/(N_c - 1)$ , for  $k = 1, \dots, N_c$ . The optimal (in terms of AF) trial point,  $\mathbf{x}_c$ , is kept in the external archive and replaces  $\mathbf{x}_r$  in the current population, provided that  $f(\mathbf{x}_c) < f(\mathbf{x}_r)$ . In that case, the algorithm moves to step 12 to finalize the generation cycle.

**Step 9.** If  $f(\mathbf{x}_r) > f(\mathbf{x}_w)$ , we use the modified objective function (Eq. (9.7)) to decide whether employing inside contraction of the simplex, thus seeking for a potential local optimum, or expanding towards a non-optimal (i.e., uphill) direction, in an attempt to escape from the current area of attraction. In this respect, if  $g(\mathbf{x}_r) > g(\mathbf{x}_w)$  we move to step 10a, otherwise we move to step 10b.

**Step 10a.** We reject  $\mathbf{x}_r$  and implement  $N_c$  trial inside contractions of the simplex in the interval between the centroid and the worst point, according to the formula:

$$\mathbf{x}_{cc}^k = \mathbf{g} - (0.25 + 0.5\delta_k) (\mathbf{g} - \mathbf{x}_r) \quad (9.10)$$

where  $\delta_k = (k - 1)/(N_c - 1)$ , for  $k = 1, \dots, N_c$ . The optimal (in terms of AF) trial point,  $\mathbf{x}_c$ , is kept in the external archive and replaces  $\mathbf{x}_w$  in the current population, provided that  $f(\mathbf{x}_c) < f(\mathbf{x}_w)$ . Otherwise, the simplex shrinks towards the best vertex  $\mathbf{x}_1$ , such as:

$$x_{s,i} = 0.5(\mathbf{x}_1 + \mathbf{x}_i) \text{ for } i = 2, \dots, n + 1 \quad (9.11)$$

We remark that the above transformation is the sole evolving mechanism of the algorithm allowing the simultaneous generation of multiple points; particularly,  $n$  new points are generated that replace all previous vertices in the current population. This can be considered as milestone of the search procedure, in the sense that a local minimum, lying in the neighborhood of  $\mathbf{x}_1$ , has been surrounded. This is the time to reduce the temperature of the optimization system by a reduction factor  $\psi$ . In contrast to EAS, where  $\psi$  is a constant parameter of the annealing cooling schedule, usually taking values into the interval 0.90–0.99, in its surrogate-enhanced version  $\psi$  is automatically adjusted to also account for the progress index PI, using the following expression:

$$\psi = \max(1 - \text{PI}, 0.50) \quad (9.12)$$

The threshold of 0.50 prohibits a fast reduction of temperature and therefore maintains enough randomness within decisions, which in turn prohibits early convergence to local optima. After reducing  $T$ , the iteration cycle is finalized (step 12).

**Step 10b.** The reflection point  $\mathbf{x}_r$  is accepted although being worse than  $\mathbf{x}_w$ . Next,  $N_u$  uphill (i.e., maximization) movements are performed using the same formula with multiple expansion (Eq. (9.9)), in an attempt to pass the hill and discover adjacent regions of attraction. This geometrical transformation was introduced by *Pan and Wu [1998]*, to facilitate the simplex escaping from local minima. Similarly to previous steps, we use the AF to determine the optimum uphill point,  $\mathbf{x}_u$ . If  $f(\mathbf{x}_u) < f(\mathbf{x}_r)$ , this point is kept in the external archive and replaces  $\mathbf{x}_r$  in the current population, while the algorithm moves to step 12 to finalize the generation cycle. Otherwise, none of the simplex transformations results to a better solution than the worst vertex  $\mathbf{x}_w$ , thus the last option is to attempt a pure stochastic generator, referred to as mutation (step 11).

**Step 11.** We seek a random point out of the typical range of the current population, defined on the basis of the mean,  $\mu_P$ , and standard deviation,  $\sigma_P$ , of all members of  $P$ . In this respect, we generate a normally-distributed point  $\mathbf{x}_m$  out of the interval  $[\mu_P - \sigma_P, \mu_P + \sigma_P]$ , which is accepted if  $f(\mathbf{x}_m) < f(\mathbf{x}_r)$ . Otherwise, we account for a user-specified mutation probability  $p_m$  in order to accept or not the randomly generated point,  $\mathbf{x}_m$ , and replacing  $\mathbf{x}_r$  in the current population. Anyway, since  $\mathbf{x}_m$  is evaluated through the objective function, it enters the external archive.

**Step 12.** Considering the new member (or members, in the particular case of simplex shrinkage) of the population, we re-evaluate the current minimum,  $\mathbf{x}_{\min}$ , and maximum,  $\mathbf{x}_{\max}$ , and their function values,  $f_{\min}$  and  $f_{\max}$ . We also re-evaluate the current number of function evaluations, FE, and check whether this hasn't exceeded the termination criterion, MFE. Finally, we re-evaluate the temperature so that  $T \leq \xi(f_{\max} - f_{\min})$ , where  $\xi \geq 1$  is a user-specified parameter of the annealing schedule, usually set between 2 to 5. This restriction prevents  $T$  taking extremely high values, which would deteriorate the efficiency of SEEAS, as far as search would become too random.

To run the algorithm, it is essential providing values for all input arguments, which are the number of desirables steps within different simplex transitions ( $N_r, N_e, N_c, N_u$ ), the mutation probability  $p_m$ , and the adjusting factor  $\xi$  of the annealing cooling schedule. Recommended values, also used in all next benchmarking tests, are  $N_r = N_e = N_c = N_u = 20$ ,  $p_m = 0.10$  and  $\xi = 2$ . These values were determined on the basis of extended investigations within the development of SEEAS, and they have been also validated through the sensitivity analysis of section 9.4.4.

## 9.3 BENCHMARKING METHODOLOGY

### 9.3.1 Benchmarking protocol

To assess the performance of SEEAS we compared it with the original version of EAS as well as three state-of-the-art optimization algorithms, which are synoptically presented in section 9.3.3. Two of the benchmark algorithms, i.e., DYCORS (DYnamic COordinate Search) and MLMSRBF (Multistart Local Metric Stochastic RBF), are surrogate-assisted, while EAS and DDS (Dynamic Dimension Search) do not employ surrogate models through search.

A variety of test problems were examined, theoretical as well as real-world. Briefly, the hereafter called benchmarking *suite* includes six mathematical test functions, formulated with 15 and 30 control variables, a hydrological calibration problem with real and synthetic data, and a time-expensive multi-reservoir management problem ( $6 \times 2 + 1 \times 2 + 1 = 15$  problems, in total).

To ensure fair comparison and safely infer about the performance of the algorithms we attempted to ensure as much as similar configurations, as summarized in Table 9-1. In all problems we employed multiple independent runs, using the same population size and the same random generation technique, i.e., LHS. The population size was set equal to  $m = 2(n + 1)$ , as recommended by Regis and Shoemaker [2007b], [2013], where  $n$  is the problem dimension (i.e., the number of control variables). We remark that other researchers relate the initial population size (also referred to as design of experiment, DoE) to the available computational budget, quantified in terms of MFE, in order to design a more detailed metamodel; for instance, Razavi et al. [2012b] suggest that  $m = \max[2(n + 1), 0.1 \text{ MFE}]$ . However, in our tests we avoided associating  $m$  with MFE, in order to investigate the impacts of the problem dimension to the performance of the examined algorithms. Furthermore, we preferred saving resources for the evolutionary procedure, instead of spending a non-negligible part of our budget to the initial DoE.



Each problem but the last was solved considering two alternative computational budgets, MFE (500 and 1000). We run all tests with two different budgets (instead of the maximum of them) since all examined algorithms (except EAS) involve parameters depending on MFE (in particular, SEEAS uses the progress index PI, defined in eq. (4), within the annealing cooling schedule). Finally, for the three surrogate-based methods (SEEAS, DYCORS, MLMSRBF) we employed the same metamodel (RBF with cubic basis functions and linear polynomial tail), thus ensuring similar computational effort for building, updating and exploiting the RBF [Razavi et al., 2012a]. We remark that in real-world problems the effort of the optimization routines (including metamodel fitting) is much less than the effort of simulation, and therefore the runtime of the overall search procedure is practically relative to MFE.

All computations were implemented in MATLAB mathematical environment using a 3.0 GHz Intel Core i5 processor with 4 GB of RAM, running on Windows 8 OS. For the SEEAS method we employed the typical input arguments given in section 9.2.2.4, while for the other algorithms, i.e., EAS, DDS, MLMSRBF and DYCORS, we used the default values suggested in the associated articles [Efstratiadis and Koutsoyiannis, 2002; Regis and Shoemaker, 2007a, 2013; Tolson and Shoemaker, 2007].

**Table 9-1** | Configuration of benchmarking suite.

| Problem                             | Algorithms             | Number of control variables, $n$ | Max. function evaluations (MFE) | Independent Runs with random initial populations | Population size | Surrogate model (metamodel)                               |
|-------------------------------------|------------------------|----------------------------------|---------------------------------|--|-----------------|---|
| Test functions                      | All                    | 15                               | 500, 1000                       | 30   | 32              |   |
| Test functions                      | All                    | 30                               | 500, 1000                       | 30   | 62              |   |
| Model calibration with real data    | All                    | 11                               | 500, 1000                       | 30   | 24              | RBF with cubic basis functions and linear polynomial tail |
| Toy calibration with synthetic data | All                    | 11                               | 500, 1000                       | 30   | 24              |   |
| Multireservoir management problem   | SEEAS, DYCORS, MLMSRBF | 20                               | 500                             | 10   | 42              |   |

### 9.3.2 Performance evaluation approach

Following the ideas of Razavi et al. [2012a] and Matott et al. [2012], after implementing all runs for each specific optimization problem solved with a specific algorithm, we plotted the cumulative distribution function (CDF) of the optimal values of  $f(\mathbf{x})$  obtained within the specific budget. In order to quantify the probability of attaining an equal or better solution, we used the concept of stochastic dominance (SD), introduced by Levy [1992], to compare the CDFs of the algorithms. Let  $\Phi_A$  and  $\Phi_B$  be the CDFs of algorithms A and B, respectively. Assuming the minimization of a random quantity  $q$ , we assume that A dominates B if  $\Phi_A(q) > \Phi_B(q)$  for all  $q$ , and vice versa. On the contrary, if the two CDFs are intersected at some point  $q_w$ , then SD is not applicable. In this case, we evaluated the median point, i.e., the one with 50% probability of exceedance, and considered as better the algorithm with the best performance at this point. In fact, to ensure that the difference of the two algorithms at the point of interest is statistically significant, we employed the non-parametric Mann–Whitney U-test [MWU - Mann and Whitney, 1947]. The null hypothesis of the MWU test is that data in  $\Phi_A$  and  $\Phi_B$  are samples from continuous distributions with equal medians. The confidence level of the MWU test was set to 95%.

### 9.3.3 Brief description of benchmarking optimization algorithms

#### 9.3.3.1 Dynamically Dimensioned Search (DDS)

Dynamically Dimension Search<sup>9</sup> (DDS), is a stochastic, single-solution based algorithm, developed by *Tolson and Shoemaker [2007]* to locate near-optimal solutions with few function evaluations. DSS is designed to search globally at the early stages and more locally when approaching a user-specified number of maximum function evaluations (MFE). It evolves by perturbing the current best solution in randomly selected dimensions, using an evolutionary operator based on the normal distribution. The probability of selecting a dimension to perturb is proportional to the current number of function evaluations and MFE. The transition from global to local search is employed by dynamically reducing the number of perturbed dimensions. In the literature are reported several successful applications of DDS [e.g., *Tolson et al., 2009; Razavi et al., 2010, 2012a; Matott et al., 2012; Regis and Shoemaker, 2013*].

#### 9.3.3.2 Multistart Local Metric Stochastic RBF algorithm (MLMSRBF)

*Regis and Shoemaker [2007a]* developed the Multistart Local Metric Stochastic RBF<sup>10</sup> (MLMSRBF), which is surrogate-assisted optimization algorithm that can be considered as extension of DDS. The first step is the implementation of the initial DoE to fit the surrogate model (particularly, RBF), which evolves by perturbing the current best point (similar to DDS), using normal distribution with zero mean and a specified covariance matrix. Additionally, in order to locate promising candidates, the algorithm uses a metric that balances the RBF prediction and the minimum distance from previously evaluated points (this is similar to the acquisition function introduced in 2.2.3, but with constant weights). The global character of the algorithm is further enhanced by implementing multiple DoEs. This multistart strategy is enabled only if the algorithm appears to have been trapped to a local minimum. The authors, demonstrated the efficiency of MLMSRBF in several benchmark problems, including 17 multimodal test functions and a 12-dimensional groundwater bioremediation problem. We note that, groundwater problems are particularly demanding, due to the numerous constraints and the typical non-linear nature of the employed objective functions [e.g., *Karatzas and Pinder, 1993, 1996*]. The literature are also reported other successful applications of the MLMSRBF method [e.g., *Mugunthan et al., 2005; Mugunthan and Shoemaker, 2006; Regis and Shoemaker, 2013*].

#### 9.3.3.3 DYNAMIC COordinate Search-Multistart Local Metric Stochastic RBF (DYCORS-LMSRBF)

The DYCORS framework was recently proposed by *Regis and Shoemaker [2013]* for surrogate-based optimization of high-dimensional expensive functions. The authors presented two versions, DYCORS-LMSRBF and DYCORS-DDSRBF<sup>11</sup>. The former is extension of LMSRBF and the latter is a surrogate-assisted DDS (here we use DYCORS-LMSRBF that performed slightly better than DYCORS-DDSRBF). DYCORS employs a strategy similar to DDS by dynamically and probabilistically reducing the number of perturbed dimensions until reaching the MFE. In order to generate trial candidate points (on the selected/perturbed dimensions)

---

<sup>9</sup> <https://github.com/akameloo1/Dynamic-Dimension-Search>

<sup>10</sup> <https://courses.cit.cornell.edu/jmueller/> or <http://people.sju.edu/~rregis/pages/software.html>

<sup>11</sup> <https://courses.cit.cornell.edu/jmueller/>

the algorithm uses a normal distribution with zero mean and standard deviation  $\sigma_n$ , but this does not remain constant, since  $\sigma_n$  is dynamically adjusted to control the range of perturbation. Moreover, DYCORS-LMSRBF is cycling through a set of weights in order to balance exploration and exploitation of the surrogate model. The authors assessed the performance of the two algorithms against several optimization schemes in a variety of test problems, among which a 14-D hydrological calibration problem.

## 9.4 TEST FUNCTIONS

### 9.4.1 Setup of optimization problems

The first suite of benchmark problems involves the optimization of six well-known mathematical problems (test functions), combining two alternative formulations in terms of number of variables ( $n = 15$  and  $30$ ), and two algorithmic configurations in terms of MFE (500 and 1000). This setting allowed for assessing the performance of the algorithms against increasing levels of dimensionality and increasing computational budget. Considering two alternative dimensions and two computational budgets, we configured four different problems for each test function, i.e., 24 optimization problems in total. According to the benchmarking protocol explained in section 9.3.1, for all problems, we employed 30 independent runs, thus randomly changing the initial population of each search experiment. The population size of all algorithms we set equal to 32 and 62, for the 15-D and 30-D formulations, respectively.

**Table 9-2** summarizes the main characteristics of the examined test functions, which represent search spaces of different complexity. Two of them (Sphere and Zakharov) are unimodal, while the rest are multimodal (Ackley, Griewank, Rastrigin, Levy). In all cases the global minimum is known and equal to zero. The analytical expression of the test functions and the bounds of their variables are given in the Appendix of *Tsoukalas et al. [2016]*.

**Table 9-2** | Summary characteristics of test functions (see also the Appendix of *Tsoukalas et al. [2016]*).

| Problem | Test function | Response surface properties                             |
|---------|---------------|---|
| OF1     | Sphere        | Unimodal and convex                                     |
| OF2     | Ackley        | Multimodal with many local minima                       |
| OF3     | Griewank      | Multimodal with many regularly distributed local minima |
| OF4     | Zakharov      | Unimodal with a plate-shaped valley                     |
| OF5     | Rastrigin     | Multimodal with many local minima                       |
| OF6     | Levy          | Multimodal with many local minima and parabolic valleys |

### 9.4.2 Statistical evaluation of optimal solutions

An initial assessment of the performance of the five examined algorithms was made on the grounds of mean and standard deviation of the best function values obtained from each optimization set (i.e., 30 independent runs of the algorithm). The closest to zero is the mean and the lowest the standard deviation indicates that the algorithm reaches the theoretical optimum with high accuracy and reliability.

The statistical superiority of SEEAS is exhibited in all problem configurations, as shown in **Table 9-3** and **Table 9-4**, for problem dimensions  $n = 15$  and  $30$ , respectively. Specifically, for the 15-D formulation (**Table 9-3**), SEEAS achieves the best performance (i.e., the lowest mean) in three out of six (OF1, OF3, OF6) and four out of six problems (OF1, OF2, OF3, OF6), for MFE = 500 and 1000, respectively. By doubling the dimensionality of the test functions to  $n = 30$ , thus significantly increasing the complexity of the associated optimization problems, SEEAS outperforms the other algorithms in four out of six (OF1, OF2, OF3, OF6) and three

out of six problems (OF<sub>1</sub>, OF<sub>3</sub>, OF<sub>6</sub>), for MFE = 500 and 1000, respectively (Table 9-4). Considering all alternative configurations, SEEAS is optimal for 14 out of 24 problems, DYCORS and EAS are optimal for 4 out of 24, and DDS is optimal for 3 out of 24. MLMSRBF does not outperform in none of the 24 test problems.

As expected, the increase of computational budget from 500 to 1000 improves the performance of all algorithms. In general, the most significant improvement is achieved by EAS and DDS, which is reasonable since these algorithms are not surrogate-assisted, thus they are by definition designed to proceed slower than the other schemes. The convergence behavior of the algorithms is further investigated in next section.

It is also worth mentioning that all algorithms exhibit poor performance against functions OF<sub>4</sub> (Zakharov) and OF<sub>5</sub> (Rastrigin), since they fail locating satisfactory solutions for the given budgets. In particular, the plate-shaped valley of Zakharov function makes extremely difficult fitting metamodels, which degenerates to hyperplane with practically zero slopes. It is not surprising that EAS ensures the best solutions, although these are still far from the theoretical optimum. EAS has been designed to also handle flat response surfaces, which are often met in water management optimization problems. On the other hand, DDS is the algorithm that generally ensures the best solution of the Rastrigin problem. Again, this is not surprising, since the search space of this function is extremely rough, with multiple local minima, thus the most stochastic of all schemes is expected to be the most efficient.

**Table 9-3** | Mean and standard deviation of best solutions in 15-D test problems (optimal results are highlighted).

| MFE  | Test function | EAS           |        | DDS           |        | SEEAS        |        | DYCORS       |        | MLMSRBF |        |
|------|---------------|---------------|--------|---------------|--------|--------------|--------|--------------|--------|---------|--------|
|      |               | Mean          | StDev  | Mean          | StDev  | Mean         | StDev  | Mean         | StDev  | Mean    | StDev  |
| 500  | OF1           | 1.938         | 0.978  | 0.852         | 0.479  | <b>0.002</b> | 0.001  | <b>0.002</b> | 0.001  | 0.019   | 0.014  |
|      | OF2           | 7.159         | 1.723  | 6.025         | 1.314  | 0.812        | 0.233  | <b>0.809</b> | 0.372  | 2.231   | 0.658  |
|      | OF3           | 7.682         | 2.997  | 2.626         | 1.269  | <b>0.538</b> | 0.118  | 0.885        | 0.084  | 1.085   | 0.052  |
|      | OF4           | <b>39.434</b> | 14.894 | 137.447       | 52.366 | 59.144       | 28.023 | 158.669      | 47.788 | 150.411 | 49.875 |
|      | OF5           | 86.245        | 14.148 | <b>24.887</b> | 7.081  | 46.268       | 15.359 | 38.958       | 12.340 | 45.920  | 18.803 |
|      | OF6           | 1.905         | 0.877  | 0.681         | 0.314  | <b>0.203</b> | 0.105  | 1.208        | 1.406  | 1.344   | 2.129  |
| 1000 | OF1           | 0.378         | 0.177  | 0.150         | 0.079  | <b>0.001</b> | 0.001  | 0.001        | 0.000  | 0.011   | 0.007  |
|      | OF2           | 3.523         | 0.936  | 3.847         | 0.528  | <b>0.437</b> | 0.208  | 0.607        | 0.092  | 1.862   | 0.556  |
|      | OF3           | 2.444         | 1.061  | 1.505         | 0.299  | <b>0.368</b> | 0.140  | 0.809        | 0.082  | 1.040   | 0.037  |
|      | OF4           | <b>26.828</b> | 17.895 | 97.541        | 38.226 | 41.290       | 26.639 | 121.266      | 36.925 | 121.359 | 37.730 |
|      | OF5           | 59.735        | 17.012 | <b>11.233</b> | 3.136  | 29.733       | 12.838 | 33.585       | 13.490 | 35.784  | 11.031 |
|      | OF6           | 0.767         | 0.292  | 0.234         | 0.104  | <b>0.124</b> | 0.060  | 0.536        | 0.860  | 0.524   | 0.863  |

**Table 9-4** | Mean and standard deviation of best solutions in 30-D test problems (optimal results are highlighted).

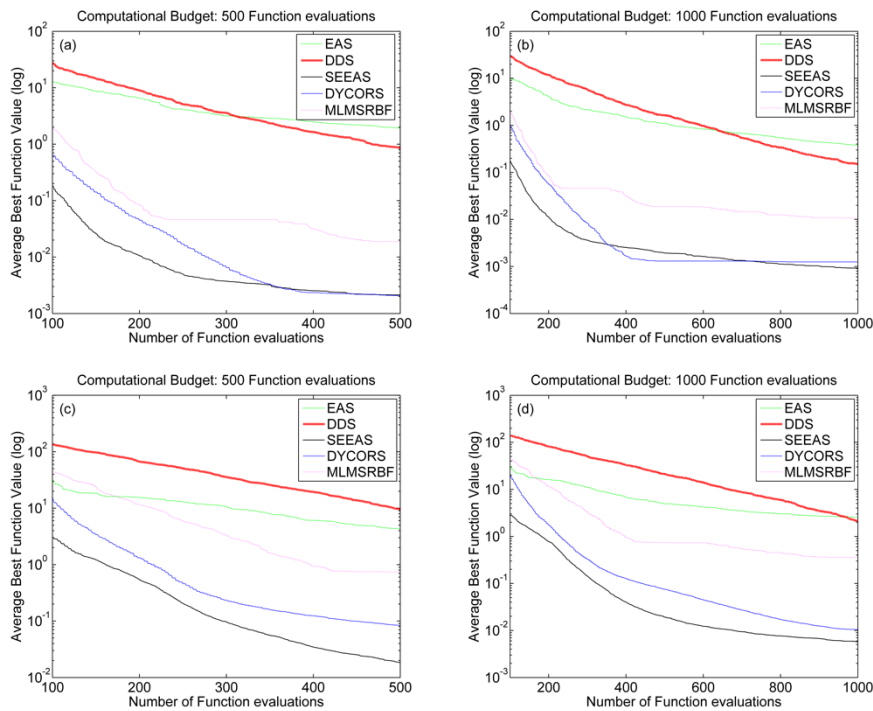
| MFE  | Test function | EAS            |        | DDS           |         | SEEAS        |        | DYCORS         |        | MLMSRBF |         |
|------|---------------|----------------|--------|---------------|---------|--------------|--------|----------------|--------|---------|---------|
|      |               | Mean           | StDev  | Mean          | StDev   | Mean         | StDev  | Mean           | StDev  | Mean    | StDev   |
| 500  | OF1           | 4.305          | 1.163  | 9.516         | 2.737   | <b>0.019</b> | 0.006  | 0.083          | 0.034  | 0.739   | 0.708   |
|      | OF2           | 9.923          | 1.160  | 12.872        | 1.329   | <b>1.878</b> | 0.301  | 4.297          | 3.721  | 6.193   | 4.362   |
|      | OF3           | 17.866         | 3.455  | 38.398        | 12.050  | <b>0.782</b> | 0.118  | 1.265          | 0.079  | 3.459   | 1.927   |
|      | OF4           | <b>117.821</b> | 28.757 | 562.145       | 113.230 | 173.240      | 44.185 | 472.815        | 90.897 | 575.424 | 174.073 |
|      | OF5           | 228.693        | 18.442 | 132.149       | 24.567  | 122.658      | 19.427 | <b>112.046</b> | 23.076 | 165.437 | 46.846  |
|      | OF6           | 6.338          | 2.652  | 15.823        | 5.481   | <b>0.659</b> | 0.184  | 3.407          | 2.540  | 7.326   | 10.944  |
| 1000 | OF1           | 2.529          | 0.933  | 2.112         | 0.791   | <b>0.006</b> | 0.004  | 0.011          | 0.004  | 0.358   | 0.177   |
|      | OF2           | 6.516          | 0.845  | 7.670         | 0.924   | 1.206        | 0.297  | <b>1.085</b>   | 0.168  | 3.643   | 1.103   |
|      | OF3           | 8.836          | 2.617  | 8.273         | 2.679   | <b>0.549</b> | 0.093  | 1.020          | 0.026  | 2.420   | 0.713   |
|      | OF4           | <b>94.598</b>  | 20.317 | 412.238       | 118.573 | 151.472      | 54.097 | 403.812        | 93.081 | 491.425 | 146.097 |
|      | OF5           | 198.335        | 16.587 | <b>71.598</b> | 15.028  | 98.371       | 19.505 | 85.267         | 22.956 | 134.864 | 39.193  |
|      | OF6           | 2.683          | 0.736  | 3.921         | 2.215   | <b>0.443</b> | 0.126  | 4.213          | 5.440  | 2.865   | 4.583   |

### 9.4.3 Evaluation of convergence behavior

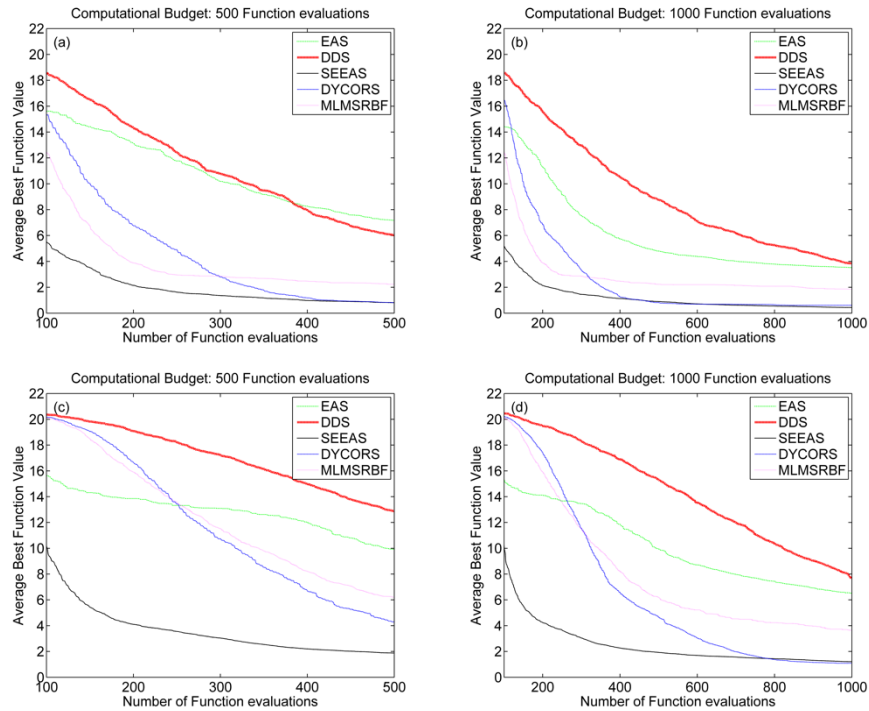
In order to further investigate the convergence behavior of the algorithms, we plotted the average (out of 30 trials) value of the best point found so far against the number of function evaluations (**Figure 9.3-Figure 9.8**). Each figure refers to a specific test function and comprises four charts, for the alternative configurations (two dimensions  $\times$  two MFE).

In most cases, SEEAS exhibits the faster convergence, evidently because the expansion mechanisms supported by the metamodel (which provides enhanced overview of the surface geometry), allow implementing steep downhill transitions. In general, the great advantage of the simplex-based transitions is the indirect use of the concept of gradient, which favors quick location of regions of attraction of local optima. This is of particular importance in computational expensive problems, where the algorithm should quickly detect promising descent directions. In fact, SEEAS is found superior to the other two surrogate-assisted algorithms (DYCORS and MLMSRBF) in all problems, except for Rastrigin. The most impressive case is the Levy problem, where SEEAS locates a very good solution after the first one hundred of function evaluations (**Figure 9.9a**), while the mean best value found by other algorithms so far is even two orders of magnitude higher. Similar are the results for the Griewank function (**Figure 9.9b**), which could be interpreted as a rough, multimodal version of sphere. A plausible explanation for this is the combined effect of the knowledge gained by the metamodel, which easily recognizes the spherical structure of Griewank, and the simplex-based operators, using approximations of the gradient of the function.

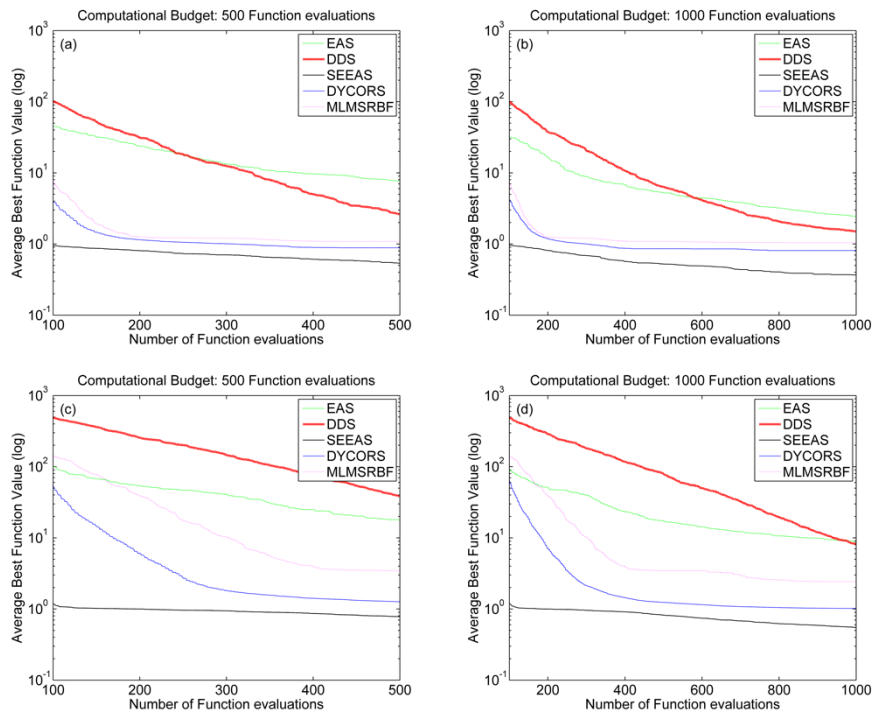
An interesting conclusion is that, regarding SEEAS, the increase of the computation budget has mild effects in the improvement of the mean best solution. This is another evidence of the suitability of SEEAS for extremely time-demanding optimization problems, in which the desirable number of function evaluations should be minimal.



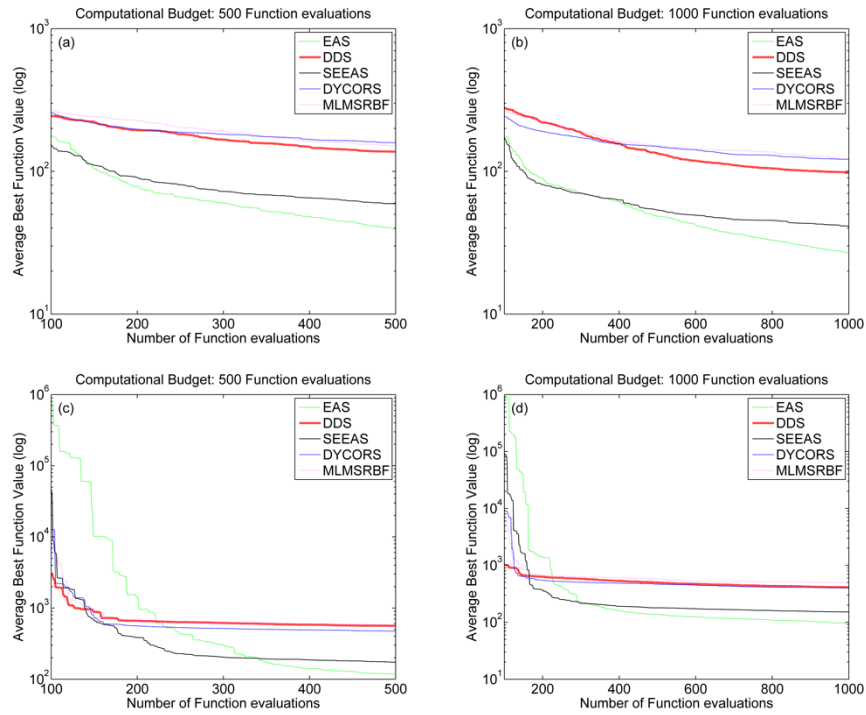
**Figure 9.3** | Convergence curves for test function OF<sub>1</sub> (Sphere) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d).



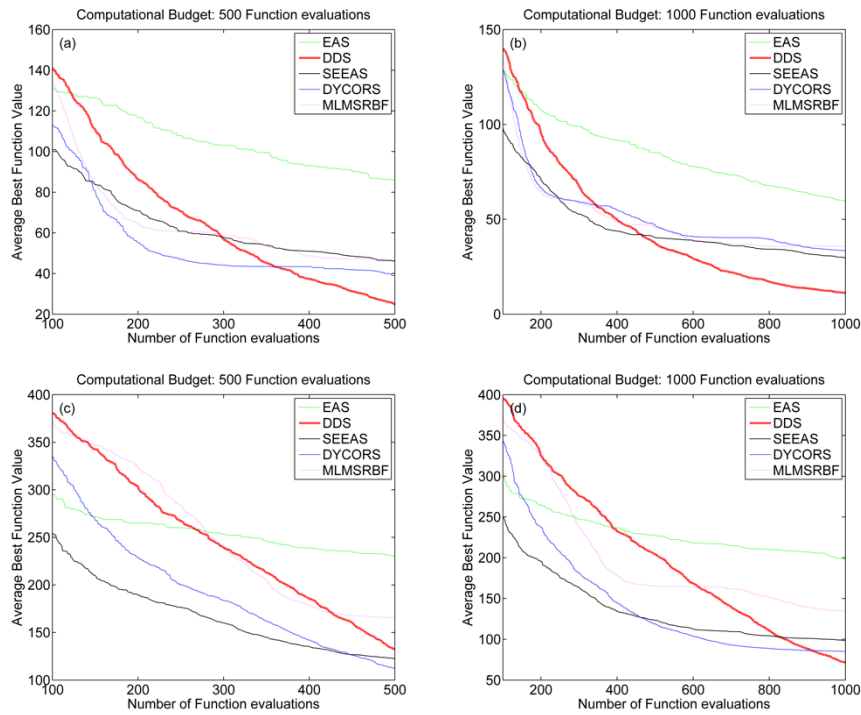
**Figure 9.4** | Convergence curves for test function OF2 (Ackley) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d).



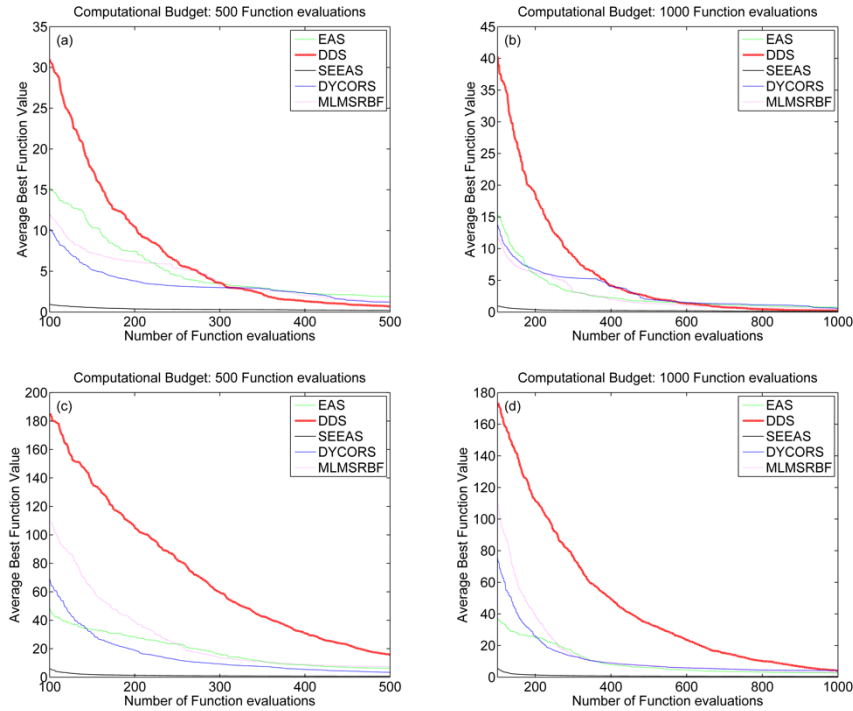
**Figure 9.5** | Convergence curves for test function OF3 (Griewank) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d).



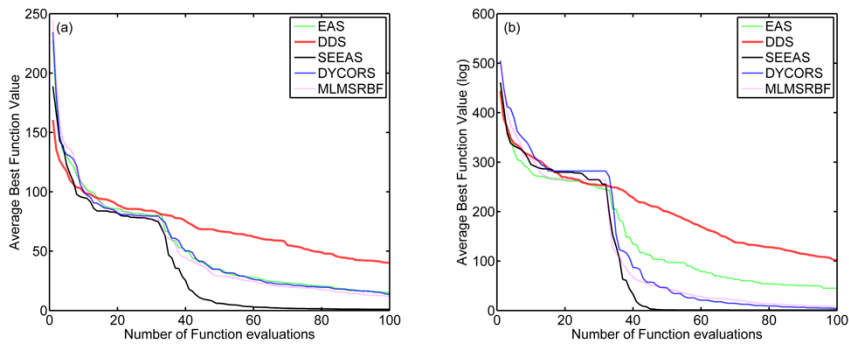
**Figure 9.6** | Convergence curves for test function OF<sub>4</sub> (Zakharov) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d).



**Figure 9.7** | Convergence curves for test function OF<sub>5</sub> (Rastrigin) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d).



**Figure 9.8** | Convergence curves for test function OF6 (Levy) with 15 (a, b) and 30 variables (c, d), with MFE=500 (a, c) and MFE=1000 (b, d).



**Figure 9.9** | Initial part of convergence curves up to 100 function evaluations, for test functions Levy (a) and Griewank (b), for the case of MFE=500 and 15 variables.

#### 9.4.4 Sensitivity analysis against input parameters of SEEAS

As mentioned in section 9.2.2.4, SEEAS requires determining several input arguments, in terms of step parameters  $N_r$ ,  $N_e$ ,  $N_c$  and  $N_u$ , mutation probability  $p_m$ , and adjusting factor  $\xi$  of the annealing cooling schedule. In order to investigate the sensitivity of SEEAS against the default values adopted so far (i.e.,  $N_r = N_e = N_c = N_u = 20$ ,  $p_m = 0.10$  and  $\xi = 2$ ), we employed 30 independent runs of the tests functions (for  $n = 15$  variables and MFE = 500), assigning different values to its input parameters. The configurations and summary statistics, in terms of means and standard deviation of the optimal solution of each set of optimizations, are given in to **Table 9-5-Table 9-7**.

The analysis justifies our recommendations for the input parameters of SEEAS. As shown in **Table 9-5**, the performance of the algorithm is significantly improved by increasing the common value of the step parameters from 5 to 20, while it is slightly improved by further increasing this value to 50. Actually, the simplex transitions are considerably assisted by using



the outcomes of the surrogate model within local search; however, it does not make sense calling the SM too many times, which introduces unnecessary computations with marginally only benefit. Regarding the mutation probability (Table 9-6), the algorithm provides almost identical results for  $p_m$  values as low as 0.05 or 0.10, yet its performance is deteriorated by increasing this probability up to 0.30. This is also a non-surprising conclusion, since it is well-known that in evolutionary algorithms the mutation operator should be occasionally called in order to avoid making search too random. Finally, the setup with  $\xi = 2$  provides systematically better results compared to  $\xi = 1$ , while it exhibits either better or similar performance when the annealing cooling parameter increases up to  $\xi = 4$  (Table 9-7). Nevertheless, a common outcome from the above investigations is the relatively low sensitivity of SEEAS against the examined configurations, for most of test problems.

**Table 9-5** | Mean and standard deviation of best solutions in 15-D test problems for MFE = 500, for different values of the four step parameters of SEEAS (for  $p_m = 0.10$  and  $\xi = 2$ ).

| Test function | $N_r = N_e = N_c = N_u = 5$ |        | $N_r = N_e = N_c = N_u = 20$ |        | $N_r = N_e = N_c = N_u = 50$ |        |
|---------------|-----------------------------|--------|------------------------------|--------|------------------------------|--------|
|               | Mean                        | StDev  | Mean                         | StDev  | Mean                         | StDev  |
| OF1           | 0.002                       | 0.001  | 0.002                        | 0.001  | 0.001                        | 0.001  |
| OF2           | 1.724                       | 0.641  | 0.812                        | 0.233  | 0.806                        | 0.258  |
| OF3           | 0.714                       | 0.15   | 0.538                        | 0.118  | 0.504                        | 0.133  |
| OF4           | 87.043                      | 32.027 | 59.144                       | 28.023 | 58.030                       | 28.911 |
| OF5           | 51.299                      | 21.579 | 46.268                       | 15.359 | 45.101                       | 13.634 |
| OF6           | 0.266                       | 0.228  | 0.203                        | 0.105  | 0.192                        | 0.114  |

**Table 9-6** | Mean and standard deviation of best solutions in 15-D test problems for MFE = 500, for different values of mutation probability  $p_m$  (for  $N_r = N_e = N_c = N_u = 20$  and  $\xi = 2$ ).

| Test function | $p_m = 0.05$ |        | $p_m = 0.10$ |        | $p_m = 0.30$ |        |
|---------------|--------------|--------|--------------|--------|--------------|--------|
|               | Mean         | StDev  | Mean         | StDev  | Mean         | StDev  |
| OF1           | 0.002        | 0.001  | 0.002        | 0.001  | 0.002        | 0.001  |
| OF2           | 0.895        | 0.345  | 0.812        | 0.233  | 0.994        | 0.480  |
| OF3           | 0.534        | 0.125  | 0.538        | 0.118  | 0.654        | 0.149  |
| OF4           | 61.071       | 21.639 | 59.144       | 28.023 | 61.876       | 29.762 |
| OF5           | 46.176       | 15.088 | 46.268       | 15.359 | 49.930       | 15.636 |
| OF6           | 0.226        | 0.088  | 0.203        | 0.105  | 0.277        | 0.095  |

**Table 9-7** | Mean and standard deviation of best solutions in 15-D test problems for MFE = 500, for different values of mutation probability  $p_m$  and cooling parameter  $\xi$  (for  $N_r = N_e = N_c = N_u = 20$  and  $p_m = 0.10$ ).

| Test function | $\xi = 1$ |        | $\xi = 2$ |        | $\xi = 4$ |        |
|---------------|-----------|--------|-----------|--------|-----------|--------|
|               | Mean      | StDev  | Mean      | StDev  | Mean      | StDev  |
| OF1           | 0.003     | 0.002  | 0.002     | 0.001  | 0.002     | 0.002  |
| OF2           | 0.978     | 0.349  | 0.812     | 0.233  | 0.896     | 0.381  |
| OF3           | 0.759     | 0.141  | 0.538     | 0.118  | 0.894     | 0.172  |
| OF4           | 69.603    | 29.140 | 59.144    | 28.023 | 62.896    | 35.297 |
| OF5           | 66.730    | 15.871 | 46.268    | 15.359 | 45.403    | 17.132 |
| OF6           | 0.423     | 0.113  | 0.203     | 0.105  | 0.206     | 0.151  |

#### 9.4.5 Suitability assessment based on stochastic dominance

For each problem and each algorithm, we illustrate the empirical CDFs, using the sample of 30 best solutions found after the termination of the corresponding search procedures. Based on them, we calculated the medians of the CDFs (Table 9-8), and next employed the MWU test between the algorithms providing the best (lower) medians, to assess whether the obtained differences are significant. The results of all tests are summarized in Table 9-9.

The outcomes of the MWU test are in full accordance with previous conclusions, and prove the statistical suitability of SEEAS. Considering the full set of problems, SEEAS is evaluated as *preferred* or *equally good* in 18 out of 24 cases. Next best method is DYCORs, which is preferred or equally good in 6 out of 24 cases. If we isolate the less beneficial subset, i.e., the formulation

with 30 decision variables under the lower computational budget (lower left panel of [Table 9-9](#)), the superiority of SEEAS is even more evident.

**Table 9-8** | Median of best function values obtained from all algorithms.

| n  | Problem | MFE = 500      |               |              |                |         | MFE = 1000    |               |              |               |         |
|----|---------|----------------|---------------|--------------|----------------|---------|---------------|---------------|--------------|---------------|---------|
|    |         | EAS            | DDS           | SEEAS        | DYCORDS        | MLMSRBF | EAS           | DDS           | SEEAS        | DYCORDS       | MLMSRBF |
| 15 | OF1     | 1.457          | 0.684         | <b>0.002</b> | 0.002          | 0.012   | 0.380         | 0.131         | <b>0.001</b> | 0.001         | 0.008   |
|    | OF2     | 7.367          | 5.942         | 0.838        | <b>0.745</b>   | 2.353   | 3.519         | 3.877         | 0.410        | 0.574         | 1.629   |
|    | OF3     | 7.446          | 2.312         | <b>0.513</b> | 0.921          | 1.088   | 2.211         | 1.400         | <b>0.360</b> | 0.819         | 1.027   |
|    | OF4     | <b>34.205</b>  | 133.574       | 53.874       | 154.151        | 147.998 | <b>25.224</b> | 98.089        | 34.413       | 127.557       | 110.313 |
|    | OF5     | 85.223         | <b>24.714</b> | 45.061       | 37.912         | 37.696  | 58.926        | <b>10.813</b> | 31.808       | 32.644        | 34.522  |
|    | OF6     | 1.592          | 0.616         | <b>0.198</b> | 0.681          | 0.488   | 0.765         | 0.216         | 0.114        | <b>0.069</b>  | 0.191   |
| 30 | OF1     | 4.391          | 9.828         | <b>0.018</b> | 0.073          | 0.590   | 2.516         | 1.860         | <b>0.005</b> | 0.009         | 0.270   |
|    | OF2     | 9.844          | 13.110        | <b>1.918</b> | 3.144          | 4.725   | 6.579         | 7.831         | 1.170        | <b>1.108</b>  | 3.438   |
|    | OF3     | 17.758         | 36.453        | <b>0.807</b> | 1.249          | 2.974   | 8.741         | 7.920         | <b>0.554</b> | 1.025         | 2.507   |
|    | OF4     | <b>114.878</b> | 540.070       | 168.695      | 456.956        | 570.266 | <b>95.274</b> | 386.140       | 147.120      | 409.986       | 465.057 |
|    | OF5     | 232.766        | 130.090       | 121.973      | <b>112.009</b> | 156.834 | 200.952       | 71.160        | 97.994       | <b>85.728</b> | 127.299 |
|    | OF6     | 5.264          | 15.496        | <b>0.630</b> | 2.075          | 2.302   | 2.458         | 2.918         | <b>0.431</b> | 2.762         | 1.412   |

**Table 9-9** | Summary results of MWU test to infer about the preferred algorithm. H-value indicates the rejection or not of the null hypothesis, i.e., if  $H = 0$ , the null hypothesis is not rejected.

| n  | Problem | MFE = 500 |             |           |   | MFE = 1000 |             |           |   |
|----|---------|-----------|-------------|-----------|---|------------|-------------|-----------|---|
|    |         | Preferred | Alternative | p-Value   | H | Preferred  | Alternative | p-Value   | H |
| 15 | OF1     | SEEAS     | DYCORDS     | 5.298E-01 | 0 | SEEAS      | MLMSRBF     | 3.020E-11 | 1 |
|    | OF2     | DYCORDS   | SEEAS       | 3.478E-01 | 0 | SEEAS      | DYCORDS     | 1.492E-06 | 1 |
|    | OF3     | SEEAS     | DYCORDS     | 9.919E-11 | 1 | SEEAS      | DYCORDS     | 1.206E-10 | 1 |
|    | OF4     | EAS       | SEEAS       | 3.034E-03 | 1 | EAS        | SEEAS       | 9.883E-03 | 1 |
|    | OF5     | DDS       | DYCORDS     | 9.514E-06 | 1 | SEEAS      | DYCORDS     | 2.028E-07 | 1 |
|    | OF6     | SEEAS     | MLMSRBF     | 3.644E-02 | 1 | DYCORDS    | SEEAS       | 6.952E-01 | 0 |
| 30 | OF1     | SEEAS     | DYCORDS     | 3.020E-11 | 1 | SEEAS      | DYCORDS     | 2.154E-06 | 1 |
|    | OF2     | SEEAS     | DYCORDS     | 3.474E-10 | 1 | DYCORDS    | SEEAS       | 9.926E-02 | 0 |
|    | OF3     | SEEAS     | DYCORDS     | 3.020E-11 | 1 | SEEAS      | DYCORDS     | 3.020E-11 | 1 |
|    | OF4     | EAS       | SEEAS       | 6.526E-07 | 1 | EAS        | SEEAS       | 3.157E-05 | 1 |
|    | OF5     | DYCORDS   | SEEAS       | 5.746E-02 | 0 | DYCORDS    | DDS         | 1.988E-02 | 1 |
|    | OF6     | SEEAS     | DYCORDS     | 1.094E-10 | 1 | SEEAS      | MLMSRBF     | 2.572E-07 | 1 |

## 9.5 HYDROLOGICAL CALIBRATION

### 9.5.1 Study area, simulation model and calibration setup

Hydrological calibration is probably the most typical global optimization problem in water resources. Numerous studies have been published dealing with calibration and its shortcomings, arising from the multiple sources of uncertainty that govern all aspects of the parameter estimation procedure [[Efstratiadis and Koutsoyiannis, 2010](#)]. Here we investigated the calibration of a lumped simulation model, applied to Boeoticos Kephisos river basin, in Eastern Greece (1850 km<sup>2</sup>). The basin extends over a heavily-modified karst system with multiple peculiarities, as result of complex interactions between surface and groundwater processes as well as human interventions, by means of surface and groundwater abstractions. This hydrosystem has been subject of comprehensive research, through alternative [[Rozos et al., 2004](#); [Efstratiadis et al., 2008](#); [Nalbantis et al., 2011](#)]. Monthly precipitation, potential evapotranspiration, runoff and groundwater abstraction data are available for a 77-year period (Oct. 1907 to Sep. 1984), to be used as inputs in simulations.

For the representation of the basin processes we applied a lumped version of Hydrogeios model [[Efstratiadis et al., 2008](#)]. The basin is vertically subdivided into three storage elements that represent interception, soil moisture and groundwater. The model estimates the main responses of the basin, i.e., actual evapotranspiration, surface and groundwater runoff and groundwater losses, using nine parameters and two initial conditions, i.e., the water levels of soil and groundwater tanks at the beginning of simulation. A brief description of the model

parameters and their feasible bounds assigned is given in **Table 9-10**. Based on the above data and tools, we formulated two optimization problems, using as objective function the well-known Nash-Sutcliffe efficiency metric (NSE). The first one follows the typical calibration paradigm (i.e., inverse modelling), in which the model parameters are unknown and the model is fitted to the observed runoff of the basin. In the second formulation, also referred to as *toy* calibration, we considered the (arbitrary) parameter set shown in **Table 9-10**, which ensures a relatively high NSE value. Next, we run the model forward to obtain synthetic runoff time series, for the given parameters and the same hydrological inputs, and finally we used these synthetic runoff data to infer the model parameters. The key difference of the second approach is that since the theoretical values of model parameters are *a priori* known, the theoretical optimum is by definition one. On the contrary, in real-data calibrations both the value and the location of the global optimum are unknown. A plausible (but not certain) approximation of the maximum NSE is 0.775, which was estimated by running EAS for multiple initial populations, allowing a reasonably large number of function evaluations (MFE = 5000). The key advantage of toy calibration is that the search procedure is not affected by structural and observation errors (i.e., both the model and the data are considered perfect), which allows fairly evaluating the performance of the optimization methods. In addition, since the value of the global optimum is by definition higher than is the case of real data, the optimization problem itself becomes harder to solve. Similarly to test functions, we assessed the performance of SEEAS against the other four algorithms assuming 30 independent runs and the typical computational budgets of 500 and 1000 function evaluations (each evaluation involves the implementation of a full simulation, for given parameters). We underline that the current suite of problems is only regarded as a computational exercise, aiming to test the algorithms against challenging problems of real-world type. In an operational context, hydrological calibration is far from a blind optimization game, since it should also account for issues such as the model predictive capacity and the physical interpretation of the optimized parameters [Efstathiadis and Koutsoyiannis, 2010].

**Table 9-10** | Model parameters, feasible bounds and values assigned for toy calibrations.

| Parameter  | Description and units                              | Lower value | Upper value | Toy value |
|------------|--|-------------|-------------|-----------|
| $r$        | Interception capacity (mm)                         | 0.010       | 100.0       | 13.0      |
| $c$        | Recession coefficient for direct runoff (-)        | 0.010       | 1.000       | 0.098     |
| $k$        | Soil capacity (mm)                                 | 5.0         | 600.0       | 506.7     |
| $l$        | Recession coefficient for interflow (-)            | 0.010       | 1.000       | 0.922     |
| $\kappa$   | Interflow threshold, as ratio of soil capacity (-) | 0.010       | 1.000       | 0.945     |
| $m$        | Recession coefficient for percolation (-)          | 0.010       | 1.000       | 0.064     |
| $\varphi$  | Recession coefficient for baseflow (-)             | 0.010       | 1.000       | 0.031     |
| $y_b$      | Threshold for baseflow generation (mm)             | 5.0         | 300.0       | 35.9      |
| $\xi$      | Recession coefficient for underground losses (-)   | 0.010       | 1.000       | 0.068     |
| $s_0$      | Initial soil moisture storage (mm)                 | 0.0         | 600.0       | 5.1       |
| $\gamma_0$ | Initial groundwater storage (mm)                   | 5.0         | 300.0       | 111.2     |

### 9.5.2 Model calibration with unknown parameters

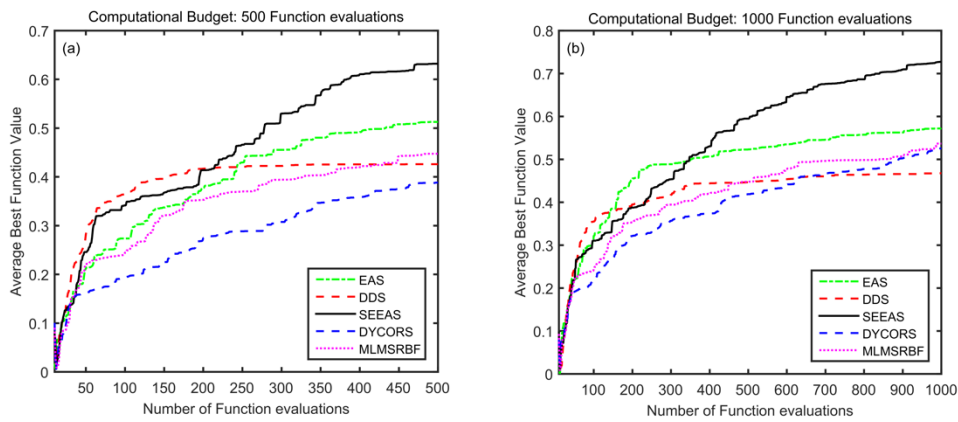
**Table 9-11** summarizes the statistical characteristics of the set of 30 optimal solutions found under the two budgets. SEEAS outperforms all other algorithms in terms of mean and median values of NSE. In particular, the mean optimal efficiency is 0.632 and 0.727, for MFE = 500 and 1000, respectively, while the medians are even higher (0.714 and 0.747, respectively). In addition, the variability of NSE values is the lowest among all algorithms. For comparison, EAS reaches a median efficiency of only 0.448, for MFE = 500, but it is considerably increased up to 0.719, for MFE = 1000. In this last case, the mean NSE is only 0.572, due to the existence of

some quite low values in the sample of 30 optimal solutions, which converge to a local optimum far from the global one. Finally, the statistical performance of the other three schemes (DDS, DYCORS, and MLMSRBF) is much less satisfactory, particularly under the restricted budget of 500 simulations.

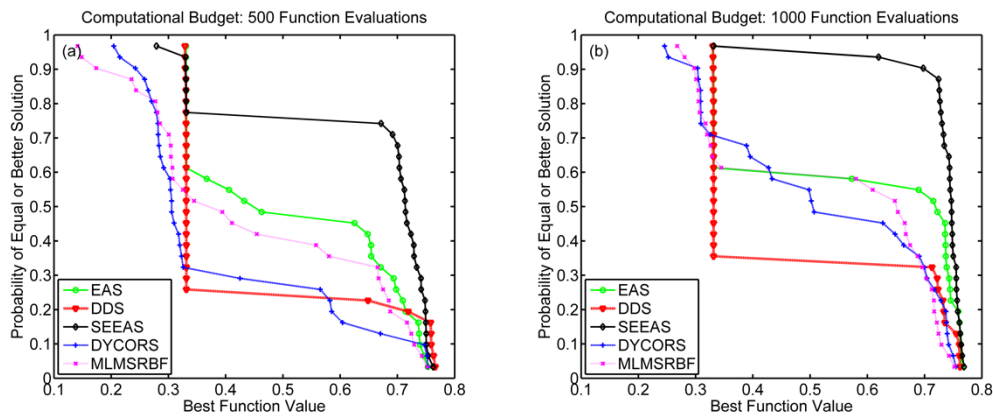
**Table 9-11** | Statistical characteristics of NSE values obtained from all algorithms.

| Budget / Statistics | MFE = 500 |       |              |        |         | MFE = 1000 |       |              |        |         |
|---------------------|-----------|-------|--------------|--------|---------|------------|-------|--------------|--------|---------|
|                     | EAS       | DDS   | SEEAS        | DYCORS | MLMSRBF | EAS        | DDS   | SEEAS        | DYCORS | MLMSRBF |
| Min                 | 0.331     | 0.329 | 0.279        | 0.204  | 0.141   | 0.331      | 0.330 | 0.331        | 0.246  | 0.268   |
| Average             | 0.513     | 0.426 | <b>0.632</b> | 0.389  | 0.447   | 0.572      | 0.467 | <b>0.727</b> | 0.525  | 0.537   |
| StDev               | 0.176     | 0.174 | 0.172        | 0.171  | 0.205   | 0.200      | 0.193 | 0.078        | 0.185  | 0.190   |
| Median              | 0.448     | 0.331 | <b>0.714</b> | 0.306  | 0.369   | 0.719      | 0.331 | <b>0.747</b> | 0.505  | 0.651   |
| Max                 | 0.753     | 0.766 | 0.763        | 0.753  | 0.752   | 0.764      | 0.762 | 0.769        | 0.755  | 0.752   |

The above conclusions are further justified when comparing the convergence curves (Figure 9.10) and the CDFs (Figure 9.11) of the five algorithms. It is shown that after 300 (for MFE = 500) or 400 (for MFE = 1000) simulations, SEEAS evolves much faster, thus locating much higher NSE values than other algorithms. The performance of EAS is also very satisfactory, given that it outperforms the other three state-of-the-art algorithms, two of which are also surrogate-assisted. Similarly, in terms of CDFs, in the low-budget scenario, SEEAS ensures NSE values greater than 0.65 in 23 out of 30 calibrations (Figure 9.11a). At the same problem, EAS performs better than other algorithms, particularly DDS, which is usually trapped to a remote local optimum. By increasing the computational budget to MFE = 1000, SEEAS systematically dominates all other schemes, ensuring NSE values greater than 0.70 in 28/30 independent calibrations (Figure 9.11b).



**Figure 9.10** | Convergence curves for MFE = 500 (a) and MFE = 1000 (b).



**Figure 9.11** | Empirical CDFs of best NSE values for MFE= 500 (a) and MFE = 1000 (b).

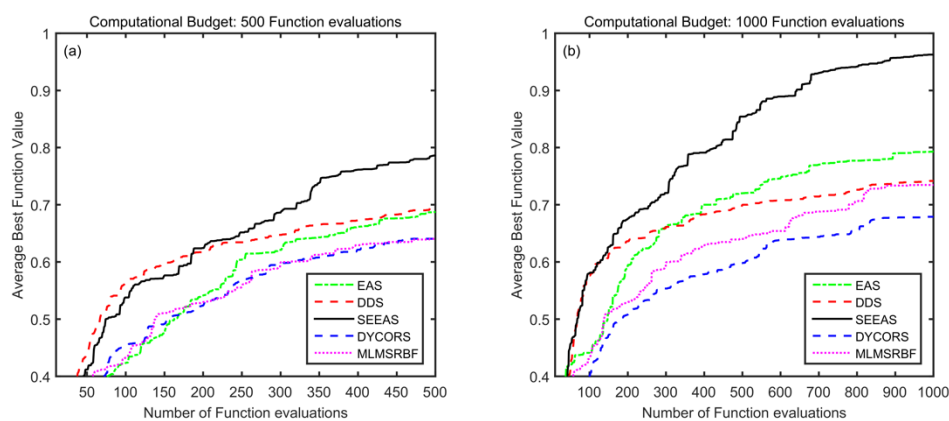
### 9.5.3 Toy calibration with synthetic runoff

As explained above, toy calibrations are more challenging, in the sense that the theoretical values of the model parameters are known, thus corresponding to unit efficiency. The outcomes of all associated tests, in terms of statistical characteristics of the best NSE value found so far, convergence curves and CDFs are shown in **Table 9-12**, **Figure 9.12** and **Figure 9.13**, respectively.

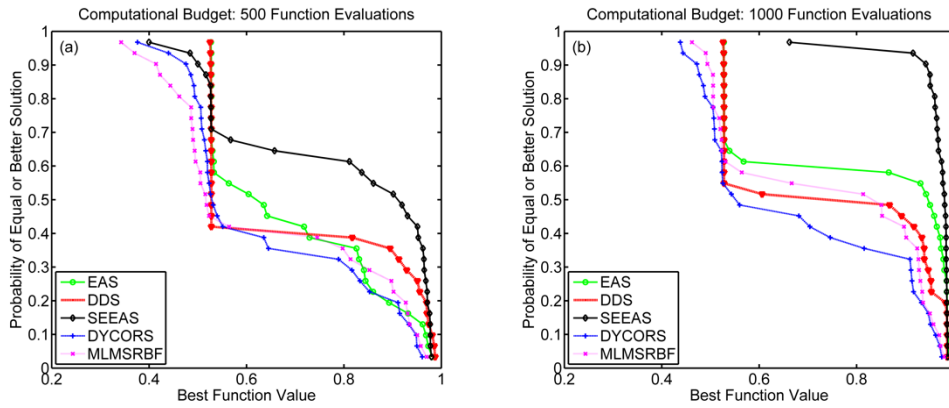
**Table 9-12** | Statistical characteristics of NSE values obtained from all algorithms.

| Budget / Statistics | MFE = 500 |       |              |        |       | MFE = 1000 |       |              |        |       |
|---------------------|-----------|-------|--------------|--------|-------|------------|-------|--------------|--------|-------|
|                     | EAS       | DDS   | SEEAS        | DYCORS | SRBF  | EAS        | DDS   | SEEAS        | DYCORS | SRBF  |
| Min                 | 0.527     | 0.525 | 0.400        | 0.376  | 0.343 | 0.527      | 0.526 | 0.662        | 0.438  | 0.462 |
| Average             | 0.688     | 0.694 | <b>0.787</b> | 0.641  | 0.640 | 0.793      | 0.742 | <b>0.963</b> | 0.679  | 0.735 |
| StDev               | 0.172     | 0.206 | 0.207        | 0.186  | 0.213 | 0.215      | 0.211 | 0.058        | 0.202  | 0.204 |
| Median              | 0.620     | 0.528 | <b>0.910</b> | 0.529  | 0.517 | 0.947      | 0.737 | <b>0.981</b> | 0.551  | 0.832 |
| Max                 | 0.981     | 0.987 | 0.978        | 0.961  | 0.970 | 0.989      | 0.986 | 0.987        | 0.975  | 0.980 |

The configuration of the calibration problem with synthetic runoff data further highlights the superiority of SEEAS against other algorithms. Specifically, the median NSE value found after only 500 simulations is 0.910, while the next best value is only 0.620, which is obtained through EAS. The increased computational budget ensures almost perfect calibrations (mean NSE = 0.963, median = 0.981), with minimal variability (standard deviation 0.058). For this budget, the median of EAS is also remarkably high (NSE = 0.947). Furthermore, SEEAS outperforms all other algorithms from the early search steps. Actually, for MFE = 500, until the first ~200 simulations DDS is competent, but then its improvement rate is significantly restricted. For the increased budget of 1000 simulations, SEEAS is arguably the best option, while EAS remains very competent. At the 2/3 of the budget, SEEAS achieves efficiency values up to 0.90, while EAS reaches values around 0.75. Even more exciting are the CDF charts, particularly for MFE = 1000; in this case, SEEAS achieves NSE values greater than 0.95 in 29 out of 30 calibration trials, and the original EAS approach also provides NSE values greater than 0.95 in 27 out of 30 trials. This indicates the remarkable reliability and robustness of the two algorithms, in contrast to other methods that generally fail to reach the known optimum in reasonable time, thus requiring multiple independent runs to ensure statistically good calibrations.



**Figure 9.12** | Convergence curves for MFE = 500 (a) and MFE = 1000 (b).



**Figure 9.13** | Empirical CDFs of best NSE values for MFE= 500 (a) and MFE = 1000 (b).

## 9.6 OPTIMIZATION OF MULTI-RESERVOIR SYSTEM PERFORMANCE

### 9.6.1 Problem statement

The second real-world test application involves the optimization of the operation of a multi-reservoir system in North-Eastern Greece. The objective was the development of uncertainty-aware operational rules that maximize the mean annual economic benefit of the system from energy production. The operation model of the hydrosystem is driven by synthetic hydrological data of 500 years length, thus drastically encumbering the computational time of simulation. In contrast to previous benchmarking tests, this problem is fully representative of real-world optimizations on a budget, given that a single function evaluation (i.e., a 500-year simulation) required  $\sim 90$  s. Due to time limitations, we compared SEEAS only against the two other surrogate-assisted algorithms, i.e., DYCORS and MLMSRBF. For each algorithm, we employed 10 independent optimizations, allowing 500 function evaluations (thus each optimization run required about 12.5 hours).

### 9.6.2 The parameterization-simulation-optimization scheme

The reservoir system extends along the downstream branch of Nestos, a transboundary river shared by Bulgaria and Greece. It comprises three serially-connected hydroelectric reservoirs (Thysavros 381 MW; Platanovryssi 116 MW; Temenos 19 MW) and a small irrigation reservoir at the outlet. The first two power plants are reversible, thus employing pumped-storage to maximize energy efficiency. The river flows are mostly regulated in the most upstream reservoir (Thysavros), while the rest of projects have limited storage capacity.

The monthly operation of the system is represented by the well-known modelling tool WEAP21 [Yates et al., 2005]. Hydrological inputs are inflows to Thysavros, as well as rainfall and evapotranspiration over all reservoir areas. The configuration of the simulation problem is explained in the recent articles by Tsoukalas and Makropoulos [2015a, 2015b], where are also provided further details about the study area and associated data.

Since the size of historical hydrological data (1968-1982; 1991-1995) is not sufficient to extract safe conclusions about the long-term performance of the system, we used instead synthetic time series of 500 years length that were generated through Castalia software [Efstratiadis et al., 2014a]<sup>12</sup>. Castalia employs a multivariate stochastic simulation scheme to generate synthetic

<sup>12</sup> See also the R language implementation, i.e., *CastaliaR* package [Tsoukalas et al., 2018c].

time series that reproduce the statistical properties of the parent historical data, at multiple temporal scales. In the specific study, in which the time step of simulation is monthly, the model preserves the observed means, standard deviations, skewness coefficients, first order autocorrelations and cross-correlations at the monthly and annual scales; it also reproduces the long-term persistence (Hurst-Kolmogorov dynamics) at the annual and over-annual scales, thus accounting for the changing behavior of hydroclimatic processes [Koutsoyiannis, 2011b]. We note that the use of Castalia in this case study is supported by the fact that the focus is in the investigation of SEEAS (and two other algorithms) performance to handle computationally expensive simulation-optimization problems. In an operational context, it could be preferable to employ the stochastic modelling and simulation approach of Chapter 4-7, since it overcomes many of limitations of the current synthetic data generation schemes.

The model decision variables were expressed in terms of energy targets, which are assigned to the associated system components, i.e., the three power plants (the two reversible). The targets refer to power production (forward operation of turbines) and consumption (backward operation, i.e., pumping). All targets were seasonally varying, considering four seasons per year, but they did not change over time (steady-state simulation). In this context, we parameterized the operation of the reservoir system through  $(3 + 2) \times 4 = 20$  energy targets. The upper bound of target values was set equal to the installed capacity of the corresponding machine (turbine or pump), while all lower bounds were set zero. At each simulation step, for given (i.e., provided by the stochastic model) inflows and known initial conditions (reservoir storages), the model transforms energy targets to equivalent minimum flow constraints, thus forcing the model to pass the required amount of water to produce (or consume) the desired amount of energy.

In the formulation of the optimization problem, we assessed the long-term performance of the system in terms of mean energy benefit from the three energy components. Based on a slight modification of the expression introduced by [Efstratiadis et al., 2012; Tsoukalas and Makropoulos, 2015b], we evaluated the monthly benefit  $b_i$  gained from each component  $i$  by:

$$b_i = c_B + e_i^* + c_S \max(e_i - e_i^*, 0) + c_D \min(e_i - e_i^*, 0) - c_P p_i \quad (9.13)$$

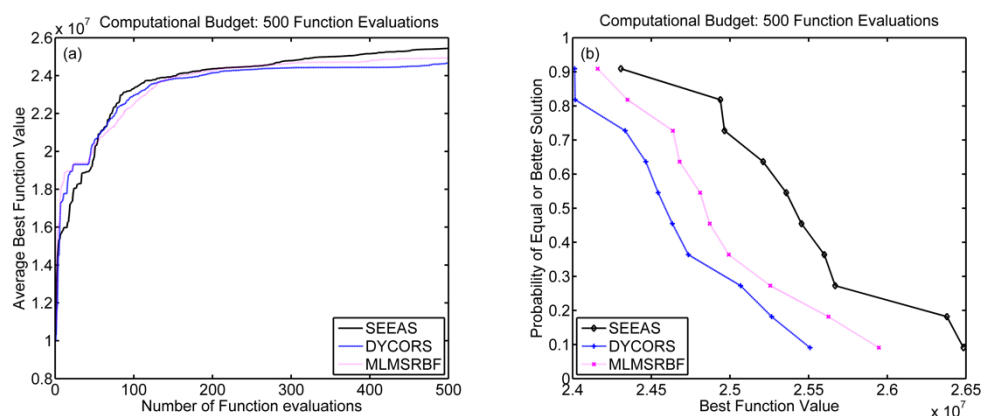
where  $e_i^*$  is the target energy that corresponds to the specific season of the year,  $e_i$  and  $p_i$  are the actual energy production and consumption (in the case of pumped-storage), which are estimated through the simulation model,  $c_B$  and  $c_S$  are unit profits for firm and secondary energy production, respectively,  $c_D$  is a unit penalty cost for energy deficits, and  $c_P$  is the unit pumping cost. The unit profit or cost values were set 0.43, 0.23, 0.80 and 0.23 €/kWh, respectively.

### 9.6.3 Results

In **Figure 9.14a** are plotted the average convergence curves of the three algorithms, while in **Figure 9.14b** are illustrated the corresponding CDFs, estimated on the basis of optimal results obtained from 10 independent trials. Once again, SEEAS outperforms both DYCORS and MLMSRBF, considering the budget of 500 trials, although their differences are relatively small. Algorithms have almost similar behavior until  $\sim 300$  FE, but then SEEAS evolves faster. In terms CDFs, SEEAS stochastically dominates MLMSRBF which, in turn, dominates DYCORS.

The two figures reveal the key peculiarity of reservoir optimization problems, which is the formulation of flat response surfaces, indicating low sensitivity of the system performance against the associated parameters. This is due to the existence of numerous constraints, physical and operational, which significantly restrict the flexibility of decisions. Generally, the decision

variables of water management models represent desirable quantities (by means of target storages, target abstractions, target flows, etc.) that may be infeasible across a wide range of the decision space. In such cases, the actual (i.e., simulated) decisions and the system performance are mainly determined by the system constraints, which in turn results to the formulation of extended valleys across the response surface.



**Figure 9.14** | Convergence curves (a) and empirical CDFs (b) for MFE = 500.

## 9.7 SUMMARY

This Chapter introduces a surrogate-enhanced extension of the evolutionary annealing-simplex (EAS) method. This new scheme, called SEEAS, uses the RBF metamodel to assist the generating mechanisms of EAS and identify promising solutions with low computational cost.

The effectiveness and efficiency of SEEAS have been demonstrated on the basis of a benchmarking suite comprising a variety of optimization problems, both theoretical and real-world. All problems were examined with alternative formulations and different budgets. The performance characteristics of SEEAS (statistical characteristics of best solution found so far, convergence behavior, stochastic dominance), were compared against three other state-of-the-art algorithms, as well as the original EAS algorithm. SEEAS outperformed all other methods in 18 out of 24 of theoretical problems (six functions with alternative configurations). Moreover, SEEAS was found superior in all real-world applications. Specifically, in hydrologic calibrations SEEAS showed consistency and robustness in locating near optimal solutions. In the first sub-case, using real runoff data, SEEAS located parameter sets with efficiency values larger than 0.70 in 28 out of 30 independent runs. Similarly, in the toy application with synthetic runoff, where the location of the optimum is *a priori* known, SEEAS ensured efficiency values larger than 0.90 in 29 out of 30 runs. Finally, in the optimization of the multi-reservoir system, SEEAS also exhibited the best behavior.

It is interesting mentioning that the two real-world applications are representative of the most typical global optimization problems in water resources. Both problems are very demanding, due to the complexity of their search space geometry. In particular, the goodness-of-fit measures used in calibrations generate highly irregular response surfaces with many local optima at all scales, in contrast to performance measures employed in water management problems, which usually compose extended smooth areas. A common characteristic of the two problems is the interactions between the model variables (or subsets of them), which is a major reason of multimodality (i.e., existence of multiple local optima with almost similar performance). The variety of generating mechanisms and quasi-stochastic transitions of



SEEAS provide flexibility to handle search spaces with such peculiarities and so diverse geometry, while other algorithms seem to be less generic.

A well-known shortcoming of hybrid optimization algorithms (also including SEEAS) is the need for defining a number of input arguments which may confuse and even discourage non-experienced users. However, in the case of SEEAS, we recommend the use of generic values for the associated inputs, which have been specified after extended investigations. In fact, our analyses indicated that the algorithm is little sensitive against its input parameters, provided that reasonable values are assigned to them. This is also a strong evidence of the robustness of SEEAS.

Current research focuses on further improving the performance of SEEAS, by testing new simplex transformations and investigating other metamodels. Moreover, the authors are working towards extending SEEAS to handle noisy functions and developing a multi-objective version of the algorithm

## CONCLUSIONS AND DISCUSSION

---

The main aim of this Thesis is to provide innovative tools and methodologies for the realistic modelling and simulation of hydrometeorological processes (i.e., the generation of synthetic hydrometeorological time series with the desirable probabilistic and stochastic properties), and simultaneously tackle the additional computational effort, which arises when long synthetic time series are used to represent the input uncertainty in simulation-optimization frameworks. Thereby, eventually ensuring the practical implementation of uncertainty-aware water-system optimization problems.

More specifically, the main objectives of this PhD Thesis are twofold:

- a) The development of novel non-Gaussian stochastic simulation models, able to account also for the other peculiarities typically encountered in hydrometeorological processes, such as, intermittency, auto- and cross- dependence, periodicity, as well as their scale-varying probabilistic and stochastic behavior (Chapter 4 to 7).
- b) The development of surrogate-based optimization methodologies and algorithms that can efficiently and effectively confront water-system simulation-optimization problems under uncertainty, i.e., when using stochastic inputs to drive the simulation-optimization procedure (Chapter 8 and 9).

As K. Pearson remark in *Notes on the History of Correlation* [1920], *the mathematics are not there for the joy of the analyst but because they are essential to the solution*. In analogy to Pearson's statement, in this Thesis, stochastic modelling and simulation, as well as optimization methods regard the *mathematics* and *water-system problems* are those requiring a *solution*. A solution that due to the critical nature of such systems, for human life and security, owes to be both uncertainty-aware and optimum.

### 10.1 STOCHASTIC MODELLING AND SIMULATION OF HYDROMETEOROLOGICAL PROCESSES

This Thesis, identified critical problems and constraints in existing simulation schemes which in turn motivated the quest for alternative simulation schemes. In this respect, the major contributions of this work are:

- a) The identification of an important flaw of linear stochastic models with non-Gaussian white noise; which can lead to bounded, and thus unrealistic and non-natural dependencies.
- b) The formal introduction in hydrology of the so-called Nataf's joint distribution model (NDM); which to the best of author's knowledge has been unknown to the hydrological community for years. NDM provides the theoretical basis for the description of the multivariate joint distribution of non-Gaussian random variables, as well as act as a main building block for the establishment of non-Gaussian conditional distribution and processes.

- c) The extension of NDM for the derivation of non-Gaussian conditional distributions.
- d) The development of a computationally simple and efficient algorithm, based on a hybrid Monte Carlo procedure, that is used to approximate the so-called equivalent correlation coefficients; which are essential for any Nataf-based model.
- e) The formulation of general guidelines that regard the development of non-Gaussian Nataf-based stochastic models, for univariate and multivariate, stationary and cyclostationary processes.
- f) The development of two Nataf-based models, termed Symmetric Moving Average (nearLy) To Anything (SMARTA) and Contemporaneous Multivariate Autoregressive (nearLy) to Anything (CMARTA), that are able to simulate univariate and multivariate stationary processes with any marginal distribution and autocorrelation structure.
- g) The extension of the notion of Nataf-based stationary processes to the cyclostationary case, and the introduction of Stochastic Periodic AutoRegressive To Anything (SPARTA) model, which hold outs the promise of simulating univariate and multivariate cyclostationary processes with arbitrary seasonally varying marginal distributions and correlations.
- h) The integration of the developed Nataf-based models within a multivariate, multi-scale disaggregation-based scheme, termed Nataf-based disaggregation to Anything (NDA), allowing the development of a modular stochastic simulation framework. This framework, enables the development of various configurations that can reproduce the desirable distributions and correlation structures at multiple time scales (e.g., as shown herein via two configurations, from annual to daily and from daily to hourly), and also cope for the unique peculiarities encountered in different scales (e.g., periodicity and intermittency in monthly and daily time scale respectively).

***What is the added value offered by Nataf-based stochastic models?***

The generation of long synthetic (hydrometeorological) time series that ideally resemble the marginal and joint properties of the parent information (e.g., observed records) is a prerequisite in many uncertainty-related hydrological studies, since they can be used as inputs and hence allow the propagation of natural variability and uncertainty to the typically deterministic water-system models. For this reason, it has been for years, one of the main research topics in the field of stochastic hydrology.

It can be argued, that the overall question is not just a technical issue, i.e., providing better stochastic models, but, in a more general context, revisiting the *essentials* of synthetic data generation. In particular, it is suggested moving from the preservation of a specific set of statistical characteristics, which are exclusively inferred from the observed data, to the preservation of *a priori* specified theoretical distributions and correlation structures that are hypothesized to be consistent with the anticipated probabilistic and stochastic behavior of the underlying processes.

Distribution functions fully describe the behavior of random variables, hence their use within stochastic simulation models is a reasonably a more precise modelling approach. Theoretical correlation structures, allow modelling and description of the dependence (temporal or spatial) in a parsimonious manner and additionally provide advantages, such as enhanced model stability and incorporation of estimator's bias. In both cases, it is also possible to take advantage of the numerous available large-scale regional studies and identify appropriate models for data-scarce regions.

The flexibility offered by the developed Nataf-based simulation schemes (particularly when integrated in a multi-scale simulation configurations, i.e., through NDA), can facilitate the

preservation of the typically non-Gaussian distribution of hydrometeorological processes and simultaneously cope for other common peculiarities (i.e., intermittency, auto- and cross-dependence, periodicity, as well as their scale-varying probabilistic and stochastic behavior).

Specifically, Nataf-based methods, allow modelling processes with continuous, discrete or mixed-type distributions (provided that their variance exists), as well as allow the selection of any the fitting method for the identification of their parameters; this in turn offers the means to exploit years of research and advances in statistical analysis of hydrometeorological variables/processes. Further to this, these models, can and should, be coupled with theoretical correlation structures thus parsimoniously identifying target dependencies to preserve, in time and space.

It is stressed that blind use of stochastic models, with overconfidence on historical data, may create a distorted *reality*, thus feeding operational hydrological and water management studies with inconsistent synthetic inputs. In this vein, it is recommended to turn our efforts into the selection of the suitable distribution model, as well as the careful assessment of the sample statistics, with emphasis to high order moments and correlations that are prone to uncertainties. Therefore, the flexibility of the proposed schemes can contribute towards the establishment of a new paradigm in hydrological stochastics.

Of course, the need and utility of non-Gaussian processes spans beyond the realm of hydrology and engineering, since it is widely acknowledged that such processes are omnipresent in many other scientific domains, such as, finance, biology, communication networks and operations research. The proposed non-Gaussian stochastic process models may find fertile ground of application also in such domains, and hopefully resolve existing issues and trigger new developments. It is also interesting to note that the developed schemes, after minimal modifications, can also be used for forecasting purposes, an arguably interesting topic for future research.

## 10.2 OPTIMIZATION OF WATER-SYSTEM PROBLEMS UNDER UNCERTAINTY

Increasing model requirements, in order to allow process descriptions at fine spatial and temporal resolutions, as well as incorporation of uncertainty (e.g., using stochastic inputs though the methods described in Chapter 4-7 of this Thesis), have substantially increased hardware requirements, in terms of computational resources and time (e.g., water-system models, and especially flood models are famous *time-expensive* simulation models). In this respect, surrogate-modelling techniques have gained significant attention, since they promise handling high-demanding optimization problems with a limited computational budget.

This Thesis contributions on the field of water-system optimization under uncertainty can be summarized as follows,

- a) The extension of the parameterization-simulation-optimization (PSO) framework for water-systems management, to handle multiple objectives, as well its effective and efficient implementation *on a budget* through the use of multi-objective surrogate-based algorithms.
- b) The development of a surrogate-enhanced evolutionary optimization algorithm, termed SEEAS, capable of handling a variety of time-expensive, water-resources, global optimization problems.

***What is the added value of employing surrogate-modelling techniques in typical simulation-optimization problems?***

This type of methodologies and algorithms are specifically designed to confront optimization problems, which are omnipresent in engineering sciences, in a fraction of time that is required by other state-of-the-art methods (e.g., evolutionary algorithms), while not relying on increasing the hardware's performance. Their utility is highlighted by the fact that simulation models requirements in computational time increase with a fast rate, therefore unwittingly pose a barrier to real-world applications of typical optimization methods, particularly when uncertainty needs to be explicitly embedded (i.e., using stochastic inputs). Such methods, ensure the practical implementation of research developments in the realm of stochastic modelling and simulation for uncertainty embedding within real-world engineering works, which this Thesis also contributes to. Beyond hydrology, fruitful applications domains are those of aerospace engineering and computational fluid dynamics, where a single simulation run of the model may require several hours or even days; a fact that prohibits the use of classical optimization methods.

### **10.3 OVERALL CONCLUSIONS AND FUTURE RESEARCH**

Incorporating uncertainty within decision making (regardless of its origin) is, and unfortunately will probably remain, a fruitful topic of research for the foreseeable future. In principle, this is achieved by formulating Monte Carlo simulation-optimization experiments driven by stochastic inputs. Such constructs enable the conversion of deterministic systems (e.g., physical or conceptual) to stochastic ones, and hence allow the analysis of the system's behavior in a probabilistic and risk-aware manner.

The contributions of this Thesis to this paradigm are twofold: (i) developing novel stochastic modelling and simulation approaches (Chapter 4-7) for the input (hydrometeorological) processes, thus allowing their more accurate representation, and hence eventually improve the quality of the deterministic model's outputs; (ii) developing novel surrogate-based methods (Chapter 8-9) to handle time expensive simulation-optimization problems, thus ensuring the operational and practical implementation of such frameworks without requiring extensive and expensive hardware infrastructure nor sacrificing the identification of optimum solutions.

The combination of these new developments, can further contribute to the wide-spread implementation of uncertainty-aware frameworks, for the design, management and operation of complex systems, aiming to identify reliable and optimal engineering solutions for the protection of human life and society from low-frequency high-impact extreme events (e.g., floods and droughts). Currently, such frameworks, are typically employed within the domains of water resources and multi-reservoir systems, yet their use and utility in other systems is relatively unexplored. As such, future research, apart from further exploring, extending and improving the new developments presented herein (see the summary of each Chapter), may enable the development and application of uncertainty-aware frameworks for similarly critical (hence requiring uncertainty embedding) and, arguably more complex systems (typically simulated by time-expensive models), such as, urban water-systems and renewable energy systems.

## REFERENCES

---

- Acreman, M. C. (1990), A simple stochastic model of hourly rainfall for Farnborough, England, *Hydrol. Sci. J.*, 35(2), 119–148, doi:10.1080/02626669009492414.
- Adeloye, A. J., B.-S. Soundharajan, J. N. Musto, and C. Chiamsathit (2015), Stochastic assessment of Phien generalized reservoir storage–yield–probability models using global runoff data records, *J. Hydrol.*, 529, 1433–1441, doi:10.1016/j.jhydrol.2015.08.038.
- Ahmad, A., A. El-Shafie, S. F. Mohd Razali, and Z. S. Mohamad (2014), Reservoir optimization in water resources: A review, *Water Resour. Manag.*, 28(11), 3391–3405, doi:10.1007/s11269-014-0700-5.
- Ailliot, P., D. Allard, V. Monbet, and P. Naveau (2015), Stochastic weather generators: an overview of weather type models, *J. la Société Française Stat.*, 156(1), 101–113.
- Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, 19(6), 716–723.
- Allard, D., and M. Bourotte (2015), Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process, *Stoch. Environ. Res. Risk Assess.*, 29(2), 453–462, doi:10.1007/s00477-014-0913-4.
- Anderson, P. L., and M. M. Meerschaert (1998), Modeling river flows with heavy tails, *Water Resour. Res.*, 34(9), 2271–2280.
- Anscombe, F. J. (1973), Graphs in Statistical Analysis, *Am. Stat.*, 27(1), 17–21, doi:10.1080/00031305.1973.10478966.
- Apipattanavis, S., G. Podestá, B. Rajagopalan, and R. W. Katz (2007), A semiparametric multivariate and multisite weather generator, *Water Resour. Res.*, 43(11), doi:10.1029/2006WR005714.
- Asadzadeh, M., and B. Tolson (2013), Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization, *Eng. Optim.*, 45(12), 1489–1509, doi:10.1080/0305215X.2012.748046.
- Australian Government Bureau of Meteorology (2015), *Hydrologic reference stations*, Bureau of Meteorology. [Available at: [www.bom.gov.au/water/hrs/](http://www.bom.gov.au/water/hrs/)].
- Bader, J., K. Deb, and E. Zitzler (2010), Faster Hypervolume-Based Search Using Monte Carlo Sampling, in *Multiple Criteria Decision Making for Sustainable Energy and Transportation Systems*, vol. 634, edited by M. Ehrgott, B. Naujoks, T. J. Stewart, and J. Wallenius, pp. 313–326, Springer Berlin Heidelberg.
- Baigorria, G. A., and J. W. Jones (2010), GiST: A Stochastic Model for Generating Spatially and Temporally Correlated Daily Rainfall Data, *J. Clim.*, 23(22), 5990–6008, doi:10.1175/2010JCLI3537.1.
- Baltas, E. A. (2007), Impact of Climate Change on the Hydrological Regime and Water Resources in the Basin of Siatista, *Int. J. Water Resour. Dev.*, 23(3), 501–518, doi:10.1080/07900620701485980.
- Baltas, E. A., and M. C. Karaliolidou (2008), Hydrological effects of land use and climate changes in northern Greece, *J. Land Use Sci.*, 2(4), 225–241,

- doi:10.1080/17474230701622908.
- Bardossy, A., and E. J. Plate (1992), Space-time model for daily rainfall using atmospheric circulation patterns, *Water Resour. Res.*, 28(5), 1247–1259, doi:10.1029/91WR02589.
- Bárdossy, A. (1998), Generating precipitation time series using simulated annealing, *Water Resour. Res.*, 34(7), 1737–1744, doi:10.1029/98WR00981.
- Bárdossy, A., and G. Pegram (2009), Copula based multisite model for daily precipitation simulation, *Hydrol. Earth Syst. Sci. Discuss.*, 6(3), 4485–4534, doi:10.5194/hessd-6-4485-2009.
- Bárdossy, A., and G. G. S. Pegram (2016), Space-time conditional disaggregation of precipitation at high resolution via simulation, *Water Resour. Res.*, 52(2), 920–937, doi:10.1002/2015WR018037.
- Barnes, F. B. (1954), Storage required for a city water supply, *J. Inst. Eng. Aust.*, 26(9), 198–203.
- Bartolini, P., J. D. Salas, and J. T. B. Obeysekera (1988), Multivariate Periodic ARMA(1,1) Processes, *Water Resour. Res.*, 24(8), 1237–1246, doi:10.1029/WR024i008p01237.
- Basso, S., M. Schirmer, and G. Botter (2015), On the emergence of heavy-tailed streamflow distributions, *Adv. Water Resour.*, 82, 98–105.
- Beersma, J. J., and T. A. Buishand (2003), Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation, *Clim. Res.*, 25(2), 121–133.
- Behzadian, K., Z. Kapelan, D. Savic, and A. Ardeshir (2009), Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks, *Environ. Model. Softw.*, 24(4), 530–541, doi:http://dx.doi.org/10.1016/j.envsoft.2008.09.013.
- Bell, T. L. (1987), A space-time stochastic model of rainfall for satellite remote-sensing studies, *J. Geophys. Res.*, 92(D8), 9631, doi:10.1029/JD092iD08p09631.
- Beran, J. (1994), *Statistics for long-memory processes*, CRC press.
- Beran, J., Y. Feng, S. Ghosh, and R. Kulik (2013), *Long-Memory Processes*, Springer Berlin Heidelberg, Berlin, Heidelberg, Heidelberg.
- Beume, N. (2009), S-metric calculation by considering dominated hypervolume as klee's measure problem, *Evol. Comput.*, 17(4), 477–492, doi:10.1162/evco.2009.17.4.17402.
- Beume, N., B. Naujoks, and M. Emmerich (2007), SMS-EMOA: Multiobjective selection based on dominated hypervolume, *Eur. J. Oper. Res.*, 181(3), 1653–1669, doi:http://dx.doi.org/10.1016/j.ejor.2006.08.008.
- Beume, N., C. M. Fonseca, I. Lopez M., L. Paquete, and J. Vahrenhold (2009), On the Complexity of Computing the Hypervolume Indicator, *Evol. Comput. IEEE Trans.*, 13(5), 1075–1082, doi:10.1109/TEVC.2009.2015575.
- Biller, B., and B. L. Nelson (2003), Modeling and generating multivariate time-series input processes using a vector autoregressive technique, *ACM Trans. Model. Comput. Simul.*, 13(3), 211–237, doi:10.1145/937332.937333.
- Blanning, R. W. (1975), The construction and implementation of metamodels, *Trans. Soc. Model. Simul. Int.*, 24(6), 177–184, doi:10.1177/003754977502400606.
- Blum, A. G., S. A. Archfield, and R. M. Vogel (2017), On the probability distribution of daily

## BIBLIOGRAPHY

- streamflow in the United States, *Hydrol. Earth Syst. Sci.*, 21(6), 3093–3103, doi:10.5194/hess-21-3093-2017.
- Bo, Z., S. Islam, and E. A. B. Eltahir (1994), Aggregation-disaggregation properties of a stochastic rainfall model, *Water Resour. Res.*, 30(12), 3423–3435, doi:10.1029/94WR02026.
- Borgomeo, E., C. L. Farmer, and J. W. Hall (2015), Numerical rivers: A synthetic streamflow generator for water resources vulnerability assessments, *Water Resour. Res.*, 51(7), 5382–5405, doi:10.1002/2014WR016827.
- Bowers, M. C., W. W. Tung, and J. B. Gao (2012), On the distributions of seasonal river flows: Lognormal or power law?, *Water Resour. Res.*, 48(5), 1–12, doi:10.1029/2011WR011308.
- Box, G. E., and G. M. Jenkins (1970), Time series analysis, forecasting and control, ed San Fr. CA Holden Day.
- Box, G. E. P., and N. R. Draper (1987), *Empirical model-building and response surfaces*, Wiley New York.
- Brandsma, T., and T. A. Buishand (1998), Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling, *Hydrol. Earth Syst. Sci.*, 2(2/3), 195–209, doi:10.5194/hess-2-195-1998.
- Bras, R. L., and I. Rodríguez-Iturbe (1985), *Random functions and hydrology*, Addison-Wesley, Reading, Mass.
- Breinl, K., T. Turkington, and M. Stowasser (2013), Stochastic generation of multi-site daily precipitation for applications in risk management, *J. Hydrol.*, 498, 23–35, doi:10.1016/j.jhydrol.2013.06.015.
- Breinl, K., T. Turkington, and M. Stowasser (2015), Simulating daily precipitation and temperature: a weather generation framework for assessing hydrometeorological hazards, *Meteorol. Appl.*, 22(3), 334–347, doi:10.1002/met.1459.
- Breinl, K., G. Di Baldassarre, M. Girons Lopez, M. Hagenlocher, G. Vico, and A. Rutgersson (2017), Can weather generation capture precipitation patterns across different climates, spatial scales and under data scarcity?, *Sci. Rep.*, 7(1), 5449, doi:10.1038/s41598-017-05822-y.
- Bringmann, K., and T. Friedrich (2009), Approximating the Least Hypervolume Contributor: NP-Hard in General, But Fast in Practice, in *Evolutionary Multi-Criterion Optimization*, vol. 5467, edited by M. Ehrgott, C. Fonseca, X. Gandibleux, J.-K. Hao, and M. Sevaux, pp. 6–20, Springer Berlin Heidelberg.
- Brissette, F. P., M. Khalili, and R. Leconte (2007), Efficient stochastic generation of multi-site synthetic precipitation data, *J. Hydrol.*, 345(3–4), 121–133, doi:10.1016/j.jhydrol.2007.06.035.
- Broad, D., G. Dandy, and H. Maier (2005), Water Distribution System Optimization Using Metamodels, *J. Water Resources Plan. Manag.*, 131(3), 172–180.
- Brockhoff, D., and E. Zitzler (2009), Objective Reduction in Evolutionary Multiobjective Optimization: Theory and Applications, *Evol. Comput.*, 17(2), 135–166, doi:10.1162/evco.2009.17.2.135.
- Brockwell, P. J., and R. A. Davis (2006), *Time series : theory and methods*.



- Buhmann, M. D. (2003), *Radial Basis Functions*, Cambridge University Press.
- Buishand, T. A. (1978), Some remarks on the use of daily rainfall models, *J. Hydrol.*, 36(3–4), 295–308, doi:10.1016/0022-1694(78)90150-6.
- Buishand, T. A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine Basin by nearest-neighbor resampling, *Water Resour. Res.*, 37(11), 2761–2776, doi:10.1029/2001WR000291.
- Burr, I. W. (1942), Cumulative Frequency Functions, *Ann. Math. Stat.*, 13(2), 215–232, doi:10.1214/aoms/1177731607.
- Burton, A., C. G. Kilsby, H. J. Fowler, P. S. P. Cowpertwait, and P. E. O’Connell (2008), RainSim: A spatial–temporal stochastic rainfall modelling system, *Environ. Model. Softw.*, 23(12), 1356–1369, doi:10.1016/j.envsoft.2008.04.003.
- Camacho, F., A. I. McLeod, and K. W. Hipel (1985), Contemporaneous autoregressive-moving average (CARMA) modeling in water resources, *J. Am. Water Resour. Assoc.*, 21(4), 709–720, doi:10.1111/j.1752-1688.1985.tb05384.x.
- Camacho, F., A. I. McLeod, and K. W. Hipel (1987), Multivariate contemporaneous ARMA model with hydrological applications, *Stoch. Hydrol. Hydraul.*, 1(2), 141–154, doi:10.1007/BF01543810.
- Cannon, A. J. (2008), Probabilistic Multisite Precipitation Downscaling by an Expanded Bernoulli–Gamma Density Network, *J. Hydrometeorol.*, 9(6), 1284–1300, doi:10.1175/2008JHM960.1.
- Cario, M. C. (1996), Modeling and simulating time series input processes with ARTAFACTS and ARTAGEN, in *Proceedings of the 28th conference on Winter simulation - WSC ’96*, edited by J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, pp. 207–213, ACM Press, New York, New York, USA.
- Cario, M. C., and B. L. Nelson (1996), Autoregressive to anything: Time-series input processes for simulation, *Oper. Res. Lett.*, 19(2), 51–58, doi:10.1016/0167-6377(96)00017-X.
- Cario, M. C., and B. L. Nelson (1997), *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix*, Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.
- Cavanaugh, N. R., A. Gershunov, A. K. Panorska, and T. J. Kozubowski (2015), The probability distribution of intense daily precipitation, *Geophys. Res. Lett.*, 42(5), 1560–1567.
- Celeste, A. B., and M. Billib (2009), Evaluation of stochastic reservoir operation optimization models, *Adv. Water Resour.*, 32(9), 1429–1443, doi:10.1016/j.advwatres.2009.06.008.
- Chen, H. (2001), Initialization for NORTA: Generation of Random Vectors with Specified Marginals and Correlations, *INFORMS J. Comput.*, 13(4), 312–331, doi:10.1287/ijoc.13.4.312.9736.
- Chen, J., F. P. Brissette, and X. J. Zhang (2014), A multi-site stochastic weather generator for daily precipitation and temperature, *Trans. ASABE*, 57(5), 1375–1391.
- Chen, L., V. P. Singh, S. Guo, J. Zhou, and J. Zhang (2015), Copula-based method for multisite monthly and daily streamflow simulation, *J. Hydrol.*, 528, 369–384, doi:10.1016/j.jhydrol.2015.05.018.

## BIBLIOGRAPHY

- Chen, L., V. Singh, and F. Xiong (2017), An Entropy-Based Generalized Gamma Distribution for Flood Frequency Analysis, *Entropy*, 19(12), 239, doi:10.3390/e19060239.
- Cheng, C.-T., X.-Y. Wu, and K. W. Chau (2005), Multiple criteria rainfall–runoff model calibration using a parallel genetic algorithm in a cluster of computers / Calage multi-critères en modélisation pluie–débit par un algorithme génétique parallèle mis en œuvre par une grappe d'ordinateurs, *Hydrol. Sci. J.*, 50(6), 1069–1087, doi:10.1623/hysj.2005.50.6.1069.
- Chilès, J.-P., and P. Delfiner (1999), Geostatistics: modeling spatial uncertainty, *Jhon Wiley Sons Inc., New York*, 695.
- Chin, E. H. (1977), Modeling daily precipitation occurrence process with Markov chain, *Water Resour. Res.*, 13(6), 949–956.
- Clark, M. P., S. Gangopadhyay, D. Brandon, K. Werner, L. Hay, B. Rajagopalan, and D. Yates (2004), A resampling procedure for generating conditioned daily weather sequences, *Water Resour. Res.*, 40(4), 1–15, doi:10.1029/2003WR002747.
- Coello Coello, C. A., G. B. Lamont, and D. A. Van Veldhuizen (2007), *Evolutionary Algorithms for Solving Multi-Objective Problems (Genetic and Evolutionary Computation)*, 2nd ed., Springer, New York.
- Cortes, C., and V. Vapnik (1995), Support-vector networks, *Mach. Learn.*, 20(3), 273–297, doi:10.1007/BF00994018.
- Couckuyt, I., D. Deschrijver, and T. Dhaene (2012), Towards Efficient Multiobjective Optimization: Multiobjective statistical criteria, in *IEEE Congress on Evolutionary Computation (CEC), 2012*, pp. 1–8.
- Couckuyt, I., D. Deschrijver, and T. Dhaene (2013), Fast calculation of multiobjective probability of improvement and expected improvement criteria for Pareto optimization, *J. Glob. Optim.*, 1–20, doi:10.1007/s10898-013-0118-2.
- Cowpertwait, P. S. P. (1991), Further developments of the neyman-scott clustered point process for modeling rainfall, *Water Resour. Res.*, 27(7), 1431–1438, doi:10.1029/91WR00479.
- Crestaux, T., O. Le Maître, and J.-M. Martinez (2009), Polynomial chaos expansion for sensitivity analysis, *Reliab. Eng. Syst. Saf.*, 94(7), 1161–1172, doi:10.1016/j.res.2008.10.008.
- Crouse, M., and R. G. Baraniuk (1999), Fast, exact synthesis of gaussian and nongaussian long-range-dependent processes, *IEEE Trans. Inf. Theory*.
- Cryer, J. D., and K.-S. Chan (2008), *Time Series Analysis. With Applications to R*.
- Dall'Aglio, G. (1959), *Sulla compatibilità delle funzioni di ripartizione doppia*.
- Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan (2002), A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Trans. Evol. Comput.*, 6(2), 182–197, doi:10.1109/4235.996017.
- Deidda, R., R. Benzi, and F. Siccardi (1999), Multifractal modeling of anomalous scaling laws in rainfall, *Water Resour. Res.*, 35(6), 1853–1867, doi:10.1029/1999WR900036.
- Demirtas, H., and D. Hedeker (2011), A Practical Way for Computing Approximate Lower and Upper Correlation Bounds, *Am. Stat.*, 65(2), 104–109, doi:10.1198/tast.2011.10090.
- Deodatis, G., and R. C. Micaletti (2001), Simulation of highly skewed non-Gaussian stochastic

- processes, *J. Eng. Mech.*, 127(12), 1284–1295.
- Detzel, D. H. M., and M. R. M. Mine (2017), Comparison between Deseasonalized Models for Monthly Streamflow Generation in a Hurst–Kolmogorov Process Framework, *J. Hydrol. Eng.*, 22(4), 05016040, doi:10.1061/(ASCE)HE.1943-5584.0001488.
- Dias, B. H., M. A. Tomim, A. L. M. Marcato, T. P. Ramos, R. B. S. Brandi, I. C. da S. Junior, and J. A. P. Filho (2013), Parallel computing applied to the stochastic dynamic programming for long term operation planning of hydrothermal power systems, *Eur. J. Oper. Res.*, 229(1), 212–222, doi:http://dx.doi.org/10.1016/j.ejor.2013.02.024.
- Dibike, Y., S. Velickov, D. Solomatine, and M. Abbott (2001), Model Induction with Support Vector Machines: Introduction and Applications, *J. Comput. Civ. Eng.*, 15(3), 208–216, doi:doi:10.1061/(ASCE)0887-3801(2001)15:3(208).
- Dimitriadis, P., and D. Koutsoyiannis (2015), Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes, *Stoch. Environ. Res. Risk Assess.*, 29(6), 1649–1669, doi:10.1007/s00477-015-1023-7.
- Ditlevsen, O. (1971), *Extremes and first passage times with applications in civil engineering*, Technical University of Denmark.
- Ditlevsen, O., and H. O. Madsen (2007), *Structural Reliability Methods*.
- Drosou, A., P. Dimitriadis, A. Lykou, P. Kossieris, I. Tsoukalas, A. Efstratiadis, and N. Mamassis (2015), Assessing and optimising flood control options along the Arachthos river floodplain (Epirus, Greece), in *European Geosciences Union General Assembly 2015, Geophysical Research Abstracts, Vol. 17, Vienna, EGU2015-9148, European Geosciences Union*.
- Duan, Q. (2013), Global Optimization for Watershed Model Calibration, in *Calibration of Watershed Models*, pp. 89–104, American Geophysical Union.
- Dunn, P. K. (2004), Occurrence and quantity of precipitation can be modelled simultaneously, *Int. J. Climatol.*, 24(10), 1231–1239.
- Eaton, M. L. (1983), *Multivariate statistics: a vector space approach.*, JOHN WILEY SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 512.
- Efstratiadis, A., and D. Koutsoyiannis (2002), An evolutionary annealing-simplex algorithm for global optimisation of water resource systems Overview of nonlinear optimisation algorithms, *Proc. Fifth Int. Conf. Hydroinformatics*, 3.
- Efstratiadis, A., and D. Koutsoyiannis (2008), Fitting Hydrological Models on Multiple Responses Using the Multiobjective Evolutionary Annealing-Simplex Approach, in *Practical Hydroinformatics*, pp. 259–273, Springer Berlin Heidelberg, Berlin, Heidelberg.
- Efstratiadis, A., and D. Koutsoyiannis (2010), One decade of multi-objective calibration approaches in hydrological modelling: a review, *Hydrol. Sci. J.*, 55(1), 58–78, doi:10.1080/02626660903526292.
- Efstratiadis, A., D. Koutsoyiannis, and D. Xenos (2004), Minimizing water cost in water resource management of Athens, *Urban Water J.*, 1(1), 3–15, doi:10.1080/15730620410001732099.
- Efstratiadis, A., I. Nalbantis, A. Koukouvinos, E. Rozos, and D. Koutsoyiannis (2008), HYDROGEIOS: a semi-distributed GIS-based hydrological model for modified river

## BIBLIOGRAPHY

- basins, *Hydrol. Earth Syst. Sci.*, 12(4), 989–1006, doi:10.5194/hess-12-989-2008.
- Efstratiadis, A., D. Bouziotas, and D. Koutsoyiannis (2012), The parameterization-simulation-optimisation framework for the management of hydroelectric reservoir systems, in *Hydrology and Society, EGU Leonardo Topical Conference Series on the hydrological cycle 2012*, Torino, European Geosciences Union.
- Efstratiadis, A., Y. G. Dialynas, S. Kozanis, and D. Koutsoyiannis (2014a), A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence, *Environ. Model. Softw.*, 62(July), 139–152, doi:10.1016/j.envsoft.2014.08.017.
- Efstratiadis, A., I. Nalbantis, and D. Koutsoyiannis (2014b), Hydrological modelling of temporally-varying catchments: facets of change and the value of information, *Hydrol. Sci. J.*, 60(7–8), null-null, doi:10.1080/02626667.2014.982123.
- Efstratiadis, A., I. Tsoukalas, P. Kossieris, G. Karavokiros, A. Christofides, A. Siskos, N. Mamassis, and D. Koutsoyiannis (2015), Computational issues in complex water-energy optimization problems: Time scales, parameterizations, objectives and algorithms, in *EGU General Assembly Conference Abstracts*, vol. 17.
- Embrechts, P., A. J. McNeil, and D. Straumann (1999), Correlation and Dependence in Risk Management: Properties and Pitfalls, in *Risk Management*, edited by M. A. H. Dempster, pp. 176–223, Cambridge University Press, Cambridge.
- Embrechts, P., F. Lindskog, and A. Mcneil (2003), Modelling Dependence with Copulas and Applications to Risk Management, in *Handbook of Heavy Tailed Distributions in Finance*, pp. 329–384.
- Emmerich, M., K. Giannakoglou, and B. Naujoks (2006), Single-and multiobjective evolutionary optimization assisted by gaussian random field metamodels, *IEEE Trans. Evol. Comput.*, 10(4), 421–439, doi:10.1109/TEVC.2005.859463.
- Emmerich, M., A. Deutz, and J. Klinkenberg (2011), Hypervolume-based expected improvement: Monotonicity properties and exact computation, in *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pp. 2147–2154.
- Esscher, F. (1924), On a method of determining correlation from the ranks of the variates, *Scand. Actuar. J.*, 1924(1), 201–219.
- Evin, G., and A.-C. Favre (2008), A new rainfall model based on the Neyman-Scott process using cubic copulas, *Water Resour. Res.*, 44(3), 1–18, doi:10.1029/2007WR006054.
- Evin, G., A. Favre, and B. Hingray (2018), Stochastic generation of multi-site daily precipitation focusing on extreme events, *Hydrol. Earth Syst. Sci.*, 22(1), 655–672, doi:10.5194/hess-22-655-2018.
- Fatichi, S., V. Y. Ivanov, and E. Caporali (2011), Simulation of future climate scenarios with a weather generator, *Adv. Water Resour.*, 34(4), 448–467, doi:10.1016/j.advwatres.2010.12.013.
- Favre, A., S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée (2004), Multivariate hydrological frequency analysis using copulas, *Water Resour. Res.*, 40(1), doi:10.1029/2003WR002456.
- Feng, M., P. Liu, S. Guo, Z. Gui, X. Zhang, W. Zhang, and L. Xiong (2017), Identifying changing

- patterns of reservoir operating rules under various inflow alteration scenarios, *Adv. Water Resour.*, 104, 23–36, doi:10.1016/j.advwatres.2017.03.003.
- Fernandez, B., and J. D. Salas (1986), Periodic Gamma Autoregressive Processes for Operational Hydrology, *Water Resour. Res.*, 22(10), 1385–1396.
- Féron, R. (1956), Sur les tableaux de corrélation dont les marges sont données, cas de l'espace à trois dimensions, *Publ. Inst. Stat. Univ. Paris*, 5, 3–12.
- Feyen, L., J. a Vrugt, B. Ó. Nualláin, J. van der Knijff, and A. De Roo (2007), Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model, *J. Hydrol.*, 332(3–4), 276–289, doi:10.1016/j.jhydrol.2006.07.004.
- Fiering, B., and B. Jackson (1971), *Synthetic Streamflows*, Water Resources Monograph, American Geophysical Union, Washington, D. C.
- Fiering, M. B. (1964), Multivariate technique for synthetic hydrology, *J. Hydraul. Div.*, 90(5), 43–60.
- Fiering, M. B. (1967), Streamflow synthesis, CAMBRIDGE, HARVARD Univ. Press. 1967. 139 P.
- Flecher, C., P. Naveau, D. Allard, and N. Brisson (2010), A stochastic daily weather generator for skewed data, *Water Resour. Res.*, 46(7), doi:10.1029/2009WR008098.
- Fleischer, M. (2003), The Measure of Pareto Optima Applications to Multi-objective Metaheuristics, in *Evolutionary Multi-Criterion Optimization*, vol. 2632, edited by C. Fonseca, P. Fleming, E. Zitzler, L. Thiele, and K. Deb, pp. 519–533, Springer Berlin Heidelberg.
- Fonseca, C., J. Knowles, and L. Thiele (2005), A tutorial on the performance assessment of stochastic multiobjective optimizers,
- Fonseca, C., A. P. Guerreiro, M. López-Ibáñez, and L. Paquete (2011), On the Computation of the Empirical Attainment Function, in *Evolutionary Multi-Criterion Optimization*, vol. 6576, edited by R. C. Takahashi, K. Deb, E. Wanner, and S. Greco, pp. 106–120, Springer Berlin Heidelberg.
- da Fonseca, V. G., C. Fonseca, and A. O. Hall (2001), Inferential performance assessment of stochastic optimisers and the attainment function, *Springer*, 213–225, doi:10.1007/3-540-44719-9\_15.
- Forrester, A., A. Sobester, and A. Keane (2008), *Engineering Design via Surrogate Modelling: A Practical Guide*, John Wiley & Sons.
- Forrester, A. I. J., and A. J. Keane (2009), Recent advances in surrogate-based optimization, *Prog. Aerosp. Sci.*, 45(1–3), 50–79, doi:10.1016/j.paerosci.2008.11.001.
- Foufoula-Georgiou, E., and D. P. Lettenmaier (1987), A Markov Renewal Model for rainfall occurrences, *Water Resour. Res.*, 23(5), 875–884, doi:10.1029/WR023i005p00875.
- Fowler, H. J., C. G. Kilsby, and P. E. O'Connell (2000), A stochastic rainfall model for the assessment of regional water resource systems under changed climatic condition, *Hydrol. Earth Syst. Sci.*, 4(2), 263–281, doi:10.5194/hess-4-263-2000.
- Fowler, K. R. et al. (2008), Comparison of derivative-free optimization methods for groundwater supply and hydraulic capture community problems, *Adv. Water Resour.*,

## BIBLIOGRAPHY

- 31(5), 743–757, doi:<http://dx.doi.org/10.1016/j.advwatres.2008.01.010>.
- Fréchet, M. (1951), Sur les tableaux de corrélation dont les marges sont données, *Ann. Univ. Lyon, 3<sup>e</sup> Ser. Sci. Sect. A, 14*, 53–77.
- Fréchet, M. (1957), Les tableaux de corrélation et les programmes linéaires, *Rev. l'Institut Int. Stat. / Rev. Int. Stat. Inst.*, 25(1/3), 23, doi:10.2307/1401672.
- Fu, G., C. Makropoulos, and D. Butler (2010), Simulation of urban wastewater systems using artificial neural networks: embedding urban areas in integrated catchment modelling, *J. Hydroinformatics*, 12(2), 140–149.
- Fu, G., Z. Kapelan, and P. Reed (2012), Reducing the Complexity of Multiobjective Water Distribution System Optimization through Global Sensitivity Analysis, *J. Water Resour. Plan. Manag.*, 138(3), 196–207, doi:10.1061/(ASCE)WR.1943-5452.0000171.
- Furrer, E. M., and R. W. Katz (2008), Improving the simulation of extreme precipitation events by stochastic weather generators, *Water Resour. Res.*, 44(12), 1–13, doi:10.1029/2008WR007316.
- Gabriel, K. R., and J. Neumann (1962), A Markov chain model for daily rainfall occurrence at Tel Aviv, *Q. J. R. Meteorol. Soc.*, 88(375), 90–95.
- Gardner, W. A., A. Napolitano, and L. Paura (2006), Cyclostationarity: Half a century of research, *Signal Processing*, 86(4), 639–697, doi:10.1016/j.sigpro.2005.06.016.
- Gates, P., and H. Tong (1976), On Markov chain modeling to some weather data, *J. Appl. Meteorol.*, 15(11), 1145–1151.
- Gaver, D. P., and P. A. W. Lewis (1980), First-order autoregressive gamma sequences and point processes, *Adv. Appl. Probab.*, 12(3), 727–745, doi:DOI: 10.1017/S0001867800035473.
- Genest, C., and A.-C. Favre (2007), Everything you always wanted to know about copula modeling but were afraid to ask, *J. Hydrol. Eng.*, 12(4), 347–368, doi:10.1061/(ASCE)1084-0699(2007)12:4(347).
- Genton, M. G. (2007), Separable approximations of space-time covariance matrices, *Environmetrics*, 18(7), 681–695, doi:10.1002/env.854.
- Genton, M. G., and W. Kleiber (2015), Cross-Covariance Functions for Multivariate Geostatistics, *Stat. Sci.*, 30(2), 147–163, doi:10.1214/14-STS487.
- Georgescu, D. I., N. Higham, and G. W. Peters (2017), Explicit Solutions to Correlation Matrix Completion Problems, with an Application to Risk Management and Insurance,
- Giuliani, M., J. D. Herman, A. Castelletti, and P. Reed (2014), Many-objective reservoir policy identification and refinement to reduce policy inertia and myopia in water management, *Water Resour. Res.*, 50(4), 3355–3377, doi:10.1002/2013WR014700.
- Giunta, A. A., S. F. Wojtkiewicz Jr, and M. S. Eldred (2003), Overview of Modern Design of Experiments Methods for Computational Simulations, in *Proceedings of the 41st AIAA Aerospace Sciences Meeting and Exhibit*, Reno, NV.
- Glasbey, C. A., G. Cooper, and M. B. McGechan (1995), Disaggregation of daily rainfall by conditional simulation from a point-process model, *J. Hydrol.*, 165(1–4), 1–9, doi:10.1016/0022-1694(94)02598-6.
- Gneiting, T. (2000), Power-law correlations, related models for long-range dependence and

- their simulation, *J. Appl. Probab.*, 37(4), 1104–1109.
- Gneiting, T., and M. Schlather (2004), Stochastic Models That Separate Fractal Dimension and the Hurst Effect, *SIAM Rev.*, 46(2), 269–282.
- Gneiting, T., W. Kleiber, and M. Schlather (2010), Matérn Cross-Covariance Functions for Multivariate Random Fields, *J. Am. Stat. Assoc.*, 105(491), 1167–1177, doi:10.1198/jasa.2010.tm09420.
- Gorissen, D., I. Couckuyt, P. Demeester, T. Dhaene, and K. Crombecq (2010), A surrogate modeling and adaptive sampling toolbox for computer based design, *J. Mach. Learn. Res.*, 11, 2051–2055.
- Granger, C. W. J., and R. Joyeux (1980), An introduction to long-memory time series and fractional differencing, *J. Time Ser. Anal.*, 1(1), 15–29, doi:10.1111/j.1467-9892.1980.tb00297.x.
- Greenwood, J. A., J. M. Landwehr, N. C. Matalas, and J. R. Wallis (1979), Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form, *Water Resour. Res.*, 15(5), 1049–1054, doi:10.1029/WR015i005p01049.
- Grigoriu, M. (1998), Simulation of stationary non-Gaussian translation processes, *J. Eng. Mech.*, 124(2), 121–126.
- Grygier, J. C., and J. R. Stedinger (1988), Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, 24(10), 1574–1584, doi:10.1029/WR024i010p01574.
- Grygier, J. C., and J. R. Stedinger (1990), SPIGOT, A synthetic streamflow generation software package, *Tech. Descr. version*, 2.
- Gupta, V. K., and E. Waymire (1990), Multiscaling properties of spatial rainfall and river flow distributions, *J. Geophys. Res.*, doi:10.1029/JD095iD03p01999.
- Gupta, V. K., and E. C. Waymire (1993), A Statistical Analysis of Mesoscale Rainfall as a Random Cascade, *J. Appl. Meteorol.*, 32(2), 251–267, doi:10.1175/1520-0450(1993)032<0251:ASAOMR>2.0.CO;2.
- Gyasi-Agyei, Y. (2011), Copula-based daily rainfall disaggregation model, *Water Resour. Res.*, 47(7), 1–17, doi:10.1029/2011WR010519.
- Gyasi-Agyei, Y., and C. S. Melching (2012), Modelling the dependence and internal structure of storm events for continuous rainfall simulation, *J. Hydrol.*, 464–465, 249–261, doi:10.1016/j.jhydrol.2012.07.014.
- Haberlandt, U., Y. Hundecha, M. Pahlow, and A. H. Schumann (2011), Rainfall generators for application in flood studies, in *Flood Risk Assessment and Management*, pp. 117–147, Springer.
- Haimes, Y. Y. (1977), Sensitivity, Responsivity, Stability and Irreversibility as Multiple Objectives in Civil Systems, *Adv. Water Resour.*, 1(2), 71–81, doi:10.1016/0309-1708(77)90025-2.
- Hamlet, A., D. Huppert, and D. Lettenmaier (2002), Economic Value of Long-Lead Streamflow Forecasts for Columbia River Hydropower, *J. Water Resour. Plan. Manag.*, 128(2), 91–101, doi:doi:10.1061/(ASCE)0733-9496(2002)128:2(91).

## BIBLIOGRAPHY

- Hao, Z., and V. P. Singh (2011), Single-site monthly streamflow simulation using entropy theory, *Water Resour. Res.*, 47(9), n/a-n/a, doi:10.1029/2010WR010208.
- Hao, Z., and V. P. Singh (2013), Modeling multisite streamflow dependence with maximum entropy copula, *Water Resour. Res.*, 49(10), 7139–7143, doi:10.1002/wrcr.20523.
- Hao, Z., and V. P. Singh (2016), Review of dependence modeling in hydrology and water resources, *Prog. Phys. Geogr.*, 40(4), 549–578, doi:10.1177/0309133316632460.
- Harms, A. A., and T. H. Campbell (1967), An extension to the Thomas-Fiering Model for the sequential generation of streamflow, *Water Resour. Res.*, 3(3), 653–661, doi:10.1029/WR003i003p00653.
- Hasan, M. M., and P. K. Dunn (2011), Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia, *Int. J. Climatol.*, 31(9), 1389–1397.
- Hazen, A. (1914), Storage to be provided in impounded reservoirs for municipal water supply, *Trans. ASCE*, 77: 1539.
- He, K., Z. Li, D. Shoubin, T. Liqun, W. Jianfeng, and Z. Chunmiao (2007), PGO: A parallel computing platform for global optimization based on genetic algorithm, *Comput. Geosci.*, 33(3), 357–366, doi:10.1016/j.cageo.2006.09.002.
- Herman, J. D., H. B. Zeff, J. R. Lamontagne, P. M. Reed, and G. W. Characklis (2016), Synthetic Drought Scenario Generation to Support Bottom-Up Water Supply Vulnerability Assessments, *J. Water Resour. Plan. Manag.*, 04016050, doi:10.1061/(ASCE)WR.1943-5452.0000701.
- Herr, H. D., and R. Krzysztofowicz (2005), Generic probability distribution of rainfall in space: The bivariate model, *J. Hydrol.*, 306(1–4), 234–263, doi:10.1016/j.jhydrol.2004.09.011.
- Higham, N. J. (2002), Computing the nearest correlation matrix--a problem from finance, *IMA J. Numer. Anal.*, 22(3), 329–343, doi:10.1093/imanum/22.3.329.
- Hipel, K. W., and A. I. McLeod (1994), *Time series modelling of water resources and environmental systems*, Elsevier.
- Hirsch, R. M. (1979), Synthetic hydrology and water supply reliability, *Water Resour. Res.*, 15(6), 1603–1615, doi:10.1029/WR015i006p01603.
- Hoeffding, W. (1994), Scale—invariant correlation theory, in *The collected works of Wassily Hoeffding*, pp. 57–107, Springer.
- Hosking, J. R. . M. (1984), Modeling persistence in hydrological time series using fractional differencing, *Water Resour. Res.*, 20(12), 1898–1908, doi:10.1029/WR020i012p01898.
- Hosking, J. R. M. (1990), L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics, *J. R. Stat. Soc. Ser. B*, 52(1), 105–124.
- Hsu, K., H. V. Gupta, and S. Sorooshian (1995), Artificial Neural Network Modeling of the Rainfall-Runoff Process, *Water Resour. Res.*, 31(10), 2517–2530, doi:10.1029/95WR01955.
- Hurst, H. E. (1951), Long-term storage capacity of reservoirs, *Trans. Amer. Soc. Civ. Eng.*, 116, 770–808.
- Iliopoulou, T., S. M. Papalexiou, Y. Markonis, and D. Koutsoyiannis (2016), Revisiting long-range dependence in annual precipitation, *J. Hydrol.*, 6(4), 399–401,



- doi:10.1016/j.jhydrol.2016.04.015.
- Jackson, B. B. (1975), The use of streamflow models in planning, *Water Resour. Res.*, 11(1), 54–63, doi:10.1029/WR011i001p00054.
- Jain, A., and A. M. Kumar (2007), Hybrid neural network models for hydrologic time series forecasting, *Appl. Soft Comput. J.*, doi:10.1016/j.asoc.2006.03.002.
- Jeong, C., and T. Lee (2015), Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions, *J. Hydro-Environment Res.*, 9(4), 604–613, doi:10.1016/j.jher.2014.06.001.
- Jin, Y. (2005), A comprehensive survey of fitness approximation in evolutionary computation, *Soft Comput.*, 9(1), 3–12, doi:10.1007/s00500-003-0328-5.
- Jin, Y. (2011), Surrogate-assisted evolutionary computation: Recent advances and future challenges, *Swarm Evol. Comput.*, 1(2), 61–70, doi:10.1016/j.swevo.2011.05.001.
- Joe, H. (2014), *Dependence modeling with copulas*, CRC Press.
- Johnson, M. E. (1987), *Multivariate Statistical Simulation*, John Wiley, New York, NY, USA.
- Jones, D. R., M. Schonlau, and W. J. Welch (1998), Efficient Global Optimization of Expensive Black-Box Functions, *J. Glob. Optim.*, 13(4), 455–492, doi:10.1023/A:1008306431147.
- Jorgensen, B. (1987), Exponential Dispersion Models, *J. R. Stat. Soc. Ser. B*, 49(2), 127–162.
- Jothiprakash, V., and G. Shanthi (2009), Comparison of Policies Derived from Stochastic Dynamic Programming and Genetic Algorithm Models, *Water Resour. Manag.*, 23(8), 1563–1580, doi:10.1007/s11269-008-9341-x.
- Kaczmarek, J., V. Isham, and C. Onof (2014), Point process models for fine-resolution rainfall, *Hydrol. Sci. J.*, 59(11), 1972–1991, doi:10.1080/02626667.2014.925558.
- Kantelhardt, J. W., E. Koscielny-Bunde, D. Rybski, P. Braun, A. Bunde, and S. Havlin (2006), Long-term persistence and multifractality of precipitation and river runoff records, *J. Geophys. Res.*, 111(D1), D01106, doi:10.1029/2005JD005881.
- Karatzas, G. P., and G. F. Pinder (1993), Groundwater management using numerical simulation and the outer approximation method for global optimization, *Water Resour. Res.*, 29(10), 3371–3378, doi:10.1029/93WR01388.
- Karatzas, G. P., and G. F. Pinder (1996), The Solution of Groundwater Quality Management Problems with a Nonconvex Feasible Region Using a Cutting Plane Optimization Technique, *Water Resour. Res.*, 32(4), 1091–1100, doi:10.1029/95WR03812.
- Katz, R. W. (1977), Precipitation as a Chain-Dependent Process, *J. Appl. Meteorol.*, 16(7), 671–676, doi:10.1175/1520-0450(1977)016<0671:PAACDP>2.0.CO;2.
- Katz, R. W., and M. B. Parlange (1995), Generalizations of Chain-Dependent Processes: Application to Hourly Precipitation, *Water Resour. Res.*, 31(5), 1331–1341, doi:10.1029/94WR03152.
- Katz, R. W., and M. B. Parlange (1998), Overdispersion Phenomenon in Stochastic Modeling of Precipitation, *J. Clim.*, 11(4), 591–601, doi:10.1175/1520-0442(1998)011<0591:OPISMO>2.0.CO;2.
- Keane, A. J. (2006), Statistical improvement criteria for use in multiobjective design

## BIBLIOGRAPHY

- optimization, *AIAA J.*, 44(4), 879–891, doi:10.2514/1.16875.
- Keating, E. H., J. Doherty, J. A. Vrugt, and Q. Kang (2010), Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality, *Water Resour. Res.*, 46(10), W10517, doi:10.1029/2009WR008584.
- Kelly, K. S., and R. Krzysztofowicz (1997), A bivariate meta-Gaussian density for use in hydrology, *Stoch. Hydrol. Hydraul.*, 11(1), 17–31, doi:10.1007/BF02428423.
- Kennedy, J., and R. Eberhart (1995), Particle swarm optimization, *Neural Networks, 1995. Proceedings., IEEE Int. Conf.*, 4, 1942–1948 vol.4, doi:10.1109/ICNN.1995.488968.
- Khalili, M., F. Brissette, and R. Leconte (2009), Stochastic multi-site generation of daily weather data, *Stoch. Environ. Res. Risk Assess.*, 23(6), 837–849, doi:10.1007/s00477-008-0275-x.
- Khu, S. T., D. A. Savic, and Z. Kapelan (2007), Evolutionary-based Meta-modelling: The relevance of using approximate models in Hydroinformatics, in *Hydroinformatics in Practice: Computational Intelligence and Technological Developments in Water Applications*, edited by R. J. Abraham, See, L.M. and Solomatine, D.P., Springer DE: Water Science and Technology Library, 2007.
- Kilsby, C. G., P. D. Jones, A. Burton, A. C. Ford, H. J. Fowler, C. Harpham, P. James, A. Smith, and R. L. Wilby (2007), A daily weather generator for use in climate change studies, *Environ. Model. Softw.*, 22(12), 1705–1719, doi:10.1016/j.envsoft.2007.02.005.
- Kim, U., J. J. Kaluarachchi, and V. U. Smakhtin (2008), Generation of Monthly Precipitation Under Climate Change for the Upper Blue Nile River Basin, Ethiopia 1, *JAWRA J. Am. Water Resour. Assoc.*, 44(5), 1231–1247, doi:10.1111/j.1752-1688.2008.00220.x.
- Kirby, W. (1972), Computer-oriented Wilson-Hilferty transformation that preserves the first three moments and the lower bound of the Pearson type 3 distribution, *Water Resour. Res.*, 8(5), 1251–1254, doi:10.1029/WR008i005p01251.
- Kisiel, C. C. (1967), Transformation of deterministic and stochastic processes in hydrology, Der Kiureghian, A., and P.-L. Liu (1986), Structural reliability under incomplete probability information, *J. Eng. Mech.*, 112(1), 85–104.
- Kleijnen, J. P. C. (2009), Kriging metamodeling in simulation: A review, *Eur. J. Oper. Res.*, 192(3), 707–716, doi:10.1016/j.ejor.2007.10.013.
- Klemeš, V. (1981), Applied stochastic theory of storage in evolution, in *Advances in hydroscience*, vol. 12, pp. 79–141, Elsevier.
- Klemeš, V. (1997), Water storage: Source of inspiration and desperation, in *Reflections on Hydrology: Science and Practice*, pp. 286–314, American Geophysical Union, Washington, D. C.
- Klemeš, V., and L. Borůvka (1974), Simulation of Gamma-Distributed First-Order Markov Chain, *Water Resour. Res.*, 10(1), 87–91, doi:10.1029/WR010i001p00087.
- Klemeš, V., R. Srikanthan, and T. A. McMahon (1981), Long-memory flow models in reservoir analysis: What is their practical value?, *Water Resour. Res.*, 17(3), 737–751, doi:10.1029/WR017i003p00737.
- Knowles, J. (2002), Local-Search and Hybrid Evolutionary Algorithms for Pareto Optimization, University of Reading.

- Knowles, J. (2005), ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multi-objective optimization problems, *IEEE Trans. Evol. Comput.*, 10(1), 50–66, doi:10.1109/TEVC.2005.851274.
- Knowles, J., and D. Corne (2003), Properties of an adaptive archiving algorithm for storing nondominated vectors, *IEEE Trans. Evol. Comput.*, 7, doi:10.1109/TEVC.2003.810755.
- Knowles, J., and H. Nakayama (2008), Meta-Modeling in Multiobjective Optimization, in *Multiobjective Optimization*, vol. 5252, edited by J. Branke, K. Deb, K. Miettinen, and R. Słowiński, pp. 245–284, Springer Berlin Heidelberg.
- Kolmogorov, A. N. (1940), Wienersche Spiralen und einige andere interessante Kurven im Hilbertschen Raum, in *CR (Dokl.) Acad. Sci. URSS*, vol. 26, pp. 115–118.
- Kossieris, P., A. Efstratiadis, and D. Koutsoyiannis (2013), The use of stochastic objective functions in water resource optimization problems, in *5th EGU Leonardo Conference – Hydrofractals 2013 – STAHY '13, Kos Island, Greece, European Geosciences Union, International Association of Hydrological Sciences, International Union of Geodesy and Geophysics*.
- Kossieris, P., A. Efstratiadis, I. Tsoukalas, and D. Koutsoyiannis (2015), Assessing the performance of Bartlett-Lewis model on the simulation of Athens rainfall, in *EGU General Assembly Conference Abstracts*, vol. 17.
- Kossieris, P., C. Makropoulos, C. Onof, and D. Koutsoyiannis (2016), A rainfall disaggregation scheme for sub-hourly time scales: Coupling a Bartlett-Lewis based model with adjusting procedures, *J. Hydrol.*, doi:10.1016/j.jhydrol.2016.07.015.
- Kottegoda, N. T. (1980), *Stochastic water resources technology*, Springer.
- Kourakos, G., and A. Mantoglou (2009), Pumping optimization of coastal aquifers based on evolutionary algorithms and surrogate modular neural network models, *Adv. Water Resour.*, 32(4), 507–521, doi:10.1016/j.advwatres.2009.01.001.
- Koutsoyiannis, D. (1999), Optimal decomposition of covariance matrices for multivariate stochastic models in hydrology, *Water Resour. Res.*, 35(4), 1219–1229, doi:10.1029/1998WR900093.
- Koutsoyiannis, D. (2000), A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series, *Water Resour. Res.*, 36(6), 1519–1533, doi:10.1029/2000WR900044.
- Koutsoyiannis, D. (2001), Coupling stochastic models of different timescales, *Water Resour. Res.*, 37(2), 379–391, doi:10.1029/2000WR900200.
- Koutsoyiannis, D. (2002), The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrol. Sci. J.*, 47(4), 573–595, doi:10.1080/02626660209492961.
- Koutsoyiannis, D. (2003), Climate change, the Hurst phenomenon, and hydrological statistics, *Hydrol. Sci. J.*, 48(1), 3–24, doi:10.1623/hysj.48.1.3.43481.
- Koutsoyiannis, D. (2005a), Reliability Concepts in Reservoir Design, in *Water Encyclopedia*, John Wiley & Sons, Inc.
- Koutsoyiannis, D. (2005b), Stochastic Simulation of Hydrosystems, in *Water Encyclopedia*, John Wiley & Sons, Inc.

## BIBLIOGRAPHY

- Koutsoyiannis, D. (2005c), Uncertainty, entropy, scaling and hydrological stochasticity. 1. Marginal distributional properties of hydrological processes and state scaling / Incertitude, entropie, effet d'échelle et propriétés stochastiques hydrologiques. 1. Propriétés distributionnelles, *Hydrol. Sci. J.*, 50(3), 381–404, doi:10.1623/hysj.50.3.381.65031.
- Koutsoyiannis, D. (2006), An entropic-stochastic representation of rainfall intermittency: The origin of clustering and persistence, *Water Resour. Res.*, 42(1), n/a-n/a, doi:10.1029/2005WR004175.
- Koutsoyiannis, D. (2010), A random walk on water, *Hydrol. Earth Syst. Sci.*, 14(3), 585–601, doi:10.5194/hess-14-585-2010.
- Koutsoyiannis, D. (2011a), Hurst-Kolmogorov Dynamics and Uncertainty, *JAWRA J. Am. Water Resour. Assoc.*, 47(3), 481–495, doi:10.1111/j.1752-1688.2011.00543.x.
- Koutsoyiannis, D. (2011b), Hurst-Kolmogorov dynamics as a result of extremal entropy production, *Phys. A Stat. Mech. its Appl.*, 390(8), 1424–1432, doi:10.1016/j.physa.2010.12.035.
- Koutsoyiannis, D. (2016), Generic and parsimonious stochastic modelling for hydrology and beyond, *Hydrol. Sci. J.*, 61(2), 225–244, doi:10.1080/02626667.2015.1016950.
- Koutsoyiannis, D. (2017), Entropy Production in Stochastics, *Entropy*, 19(11), 581, doi:10.3390/e19110581.
- Koutsoyiannis, D., and A. Economou (2003), Evaluation of the parameterization-simulation-optimization approach for the control of reservoir systems, *Water Resour. Res.*, 39(6), n/a-n/a, doi:10.1029/2003WR002148.
- Koutsoyiannis, D., and E. Foufoula-Georgiou (1993), A scaling model of a storm hyetograph, *Water Resour. Res.*, 29(7), 2345–2361, doi:10.1029/93WR00395.
- Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, 32(7), 2105–2117, doi:10.1029/96WR00488.
- Koutsoyiannis, D., and A. Montanari (2007), Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resour. Res.*, 43(5), 1–9, doi:10.1029/2006WR005592.
- Koutsoyiannis, D., and C. Onof (2001), Rainfall disaggregation using adjusting procedures on a Poisson cluster model, *J. Hydrol.*, 246(1–4), 109–122, doi:10.1016/S0022-1694(01)00363-8.
- Koutsoyiannis, D., and S. M. Papalexiou (2016), Extreme rainfall: Global perspective, in *Chow's handbook of applied hydrology, 2nd Ed.*, McGraw-Hill, New York.
- Koutsoyiannis, D., A. Efstratiadis, and G. Karavokiros (2002), A decision support tool for the management of multi-reservoir systems, *J. Am. Water Resour. Assoc.*, 38(4), 945–958, doi:10.1111/j.1752-1688.2002.tb05536.x.
- Koutsoyiannis, D., G. Karavokiros, A. Efstratiadis, N. Mamassis, A. Koukouvinos, and A. Christofides (2003a), A decision support system for the management of the water resource system of Athens, *Phys. Chem. Earth, Parts A/B/C*, 28(14–15), 599–609, doi:10.1016/S1474-7065(03)00106-2.
- Koutsoyiannis, D., C. Onof, and H. S. Wheater (2003b), Multivariate rainfall disaggregation at a fine timescale, *Water Resour. Res.*, 39(7), 1–62, doi:10.1029/2002WR001600.

- Koutsoyiannis, D., H. Yao, and A. Georgakakos (2008), Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods, *Hydrol. Sci. Journal-Journal Des Sci. Hydrol.*, 53(1), 142–164, doi:10.1623/hysj.53.1.142.
- Koutsoyiannis, D., P. Dimitriadis, F. Lombardo, and S. Stevens (2018), From Fractals to Stochastics: Seeking Theoretical Consistency in Analysis of Geophysical Data, in *Advances in Nonlinear Geosciences*, pp. 237–278, Springer International Publishing, Cham.
- Krige, D. G. (1951), A statistical approach to some basic mine valuation problems on the Witwatersrand, *J. Chem. Metall. Min. Eng. Soc. South Africa*, 52(6), 119–139.
- Kroese, D. P., T. Taimre, and Z. I. Botev (2011), *Handbook of Monte Carlo Methods*, John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Kroese, D. P., T. Brereton, T. Taimre, and Z. I. Botev (2014), Why the Monte Carlo method is so important today, *Wiley Interdiscip. Rev. Comput. Stat.*, 6(6), 386–392, doi:10.1002/wics.1314.
- Kroll, C. N., and R. M. Vogel (2002), Probability Distribution of Low Streamflow Series in the United States, *J. Hydrol. Eng.*, 7(2), 137–146, doi:10.1061/(ASCE)1084-0699(2002)7:2(137).
- Kruskal, W. H. (1958), Ordinal measures of association, *J. Am. Stat. Assoc.*, 53(284), 814–861.
- Kuzmin, V., D.-J. Seo, and V. Koren (2008), Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search, *J. Hydrol.*, 353(1–2), 109–128, doi:10.1016/j.jhydrol.2008.02.001.
- Labadie, J. W. (2004), Optimal Operation of Multireservoir Systems: State-of-the-Art Review, *J. Water Resour. Plan. Manag.*, 130(2), 93–111, doi:10.1061/(ASCE)0733-9496(2004)130:2(93).
- Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrologic time series, *Water Resour. Res.*, 32(3), 679–693, doi:10.1029/95WR02966.
- Langousis, A., and D. Koutsoyiannis (2006), A stochastic methodology for generation of seasonal time series reproducing overyear scaling behaviour, *J. Hydrol.*, 322(1–4), 138–154, doi:10.1016/j.jhydrol.2005.02.037.
- Larson, S., and S. Larson (2007), Index-based tool for preliminary ranking of social and environmental impacts of hydropower and storage reservoirs, *Energy*, 32(6), 943–947, doi:10.1016/j.energy.2006.09.007.
- Lawrance, A. J., and N. T. Kottegoda (1977), Stochastic Modelling of Riverflow Time Series, *J. R. Stat. Soc. Ser. A*, 140(1), 1, doi:10.2307/2344516.
- Lawrance, A. J., and P. A. W. Lewis (1981a), A new autoregressive time series model in exponential variables (NEAR (1)), *Adv. Appl. Probab.*, 13(04), 826–845.
- Lawrance, A. J., and P. A. W. Lewis (1981b), *Generation of some first-order autoregressive Markovian sequences of positive random variables with given marginal distributions*, Monterey, California. Naval Postgraduate School.
- Lebrun, R., and A. Dutfoy (2009), An innovating analysis of the Nataf transformation from the copula viewpoint, *Probabilistic Eng. Mech.*, 24(3), 312–320, doi:10.1016/j.probengmech.2008.08.001.

## BIBLIOGRAPHY

- Lee, T. (2016), Stochastic simulation of precipitation data for preserving key statistics in their original domain and application to climate change analysis, *Theor. Appl. Climatol.*, 124(1–2), 91–102, doi:10.1007/s00704-015-1395-0.
- Lee, T. (2017), Multisite stochastic simulation of daily precipitation from copula modeling with a gamma marginal distribution, *Theor. Appl. Climatol.*, doi:10.1007/s00704-017-2147-0.
- Lee, T., and C. Jeong (2014), Nonparametric statistical temporal downscaling of daily precipitation to hourly precipitation and implications for climate change scenarios, *J. Hydrol.*, 510, 182–196.
- Lee, T., and J. D. Salas (2011), Copula-based stochastic simulation of hydrological data applied to Nile River flows, *Hydrol. Res.*, 42(4), 318–330, doi:10.2166/nh.2011.085.
- Lee, T., J. D. Salas, and J. Prairie (2010), An enhanced nonparametric streamflow disaggregation model with genetic algorithm, *Water Resour. Res.*, 46(8), 1–14, doi:10.1029/2009WR007761.
- Lekkas, D. F., C. E. Imrie, and M. J. Lees (2001), Improved non-linear transfer function and neural network methods of flow routing for real-time forecasting, *J. Hydroinformatics*, doi:doi: 10.1103/physrevb.53.3764.
- Lekkas, D. F., C. Onof, M. J. Lee, and E. A. Baltas (2004), Application of artificial neural networks for flood forecasting, *Glob. Nest J.*, 6(3), 205–211.
- Leonov, S., and B. Qaqish (2017), Correlated endpoints: simulation, modeling, and extreme correlations, *Stat. Pap.*, doi:10.1007/s00362-017-0960-2.
- Lettenmaier, D. P., and S. J. Burges (1977), An operational approach to preserving skew in hydrologic models of long-term persistence, *Water Resour. Res.*, 13(2), 281–290, doi:10.1029/WR013i002p00281.
- Levy, H. (1992), Stochastic dominance and expected utility: survey and analysis, *Manag. Sci.*, 38(4), 555–593, doi:10.1287/mnsc.38.4.555.
- Li, C., V. P. Singh, and A. K. Mishra (2012), Simulation of the entire range of daily precipitation using a hybrid probability distribution, *Water Resour. Res.*, 48(3), 1–17, doi:10.1029/2011WR011446.
- Li, C., V. P. Singh, and A. K. Mishra (2013), A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation, *Water Resour. Res.*, 49(2), 767–789, doi:10.1002/wrcr.20063.
- Li, H., Z. Lü, and X. Yuan (2008), Nataf transformation based point estimate method, *Chinese Sci. Bull.*, 53(17), 2586.
- Li, S. T., and J. L. Hammond (1975), Generation of Pseudorandom Numbers with Specified Univariate Distributions and Correlation Coefficients, *IEEE Trans. Syst. Man. Cybern.*, SMC-5(5), 557–561, doi:10.1109/TSMC.1975.5408380.
- Licznar, P., J. Łomotowski, and D. E. Rupp (2011), Random cascade driven rainfall disaggregation for urban hydrology: An evaluation of six models and a new generator, *Atmos. Res.*, 99(3–4), 563–578, doi:10.1016/j.atmosres.2010.12.014.
- Lindgren, G. (2013), *Stationary Stochastic Processes for Scientists and Engineers*, Chapman and Hall/CRC.

- Liu, P. L., and A. Der Kiureghian (1986), Multivariate distribution models with prescribed marginals and covariances, *Probabilistic Eng. Mech.*, 1(2), 105–112, doi:10.1016/0266-8920(86)90033-0.
- Lombardo, F., E. Volpi, and D. Koutsoyiannis (2012), Rainfall downscaling in time: theoretical and empirical comparison between multifractal and Hurst-Kolmogorov discrete random cascades, *Hydrol. Sci. J.*, 57(6), 1052–1066, doi:10.1080/02626667.2012.695872.
- Lombardo, F., E. Volpi, D. Koutsoyiannis, and S. M. Papalexiou (2014), Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology, *Hydrol. Earth Syst. Sci.*, 18(1), 243–255, doi:10.5194/hess-18-243-2014.
- Lombardo, F., E. Volpi, D. Koutsoyiannis, and F. Serinaldi (2017), A theoretically consistent stochastic cascade for temporal disaggregation of intermittent rainfall, *Water Resour. Res.*, doi:10.1002/2017WR020529.
- López-Ibáñez, M., L. Paquete, and T. Stützle (2010), Exploratory analysis of stochastic local search algorithms in biobjective optimization, *Springer*, 209–222, doi:10.1007/978-3-642-02538-9\_9.
- Loucks, D. P., and E. van Beek (2017), An Introduction to Probability, Statistics, and Uncertainty, in *Water Resource Systems Planning and Management*, pp. 213–300, Springer.
- Maass, A., M. M. Hufschmidt, R. Dorfman, H. A. Thomas, S. A. Marglin, G. M. Fair, B. T. Bower, W. W. Reedy, D. F. Manzer, and M. P. Barnett (1962), *Design of water-resource systems*, Cambridge: Harvard University Press.
- Macke, J. H., P. Berens, A. S. Ecker, A. S. Tolia, and M. Bethge (2009), Generating Spike Trains with Specified Correlation Coefficients, *Neural Comput.*, 21(2), 397–423, doi:10.1162/neco.2008.02-08-713.
- Maftai, C., A. Barbulescu, and A. A. Carsteanu (2016), Long-range dependence in the time series of Taița River discharges, *Hydrol. Sci. J.*, 61(9), 1740–1747, doi:10.1080/02626667.2016.1171869.
- Maier, H. R. et al. (2014), Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions, *Environ. Model. Softw.*, 62, 271–299, doi:10.1016/j.envsoft.2014.09.013.
- Makropoulos, C. et al. (2017), Sewer-mining: A water reuse option supporting circular economy, public service provision and entrepreneurship, *J. Environ. Manage.*, doi:10.1016/j.jenvman.2017.07.026.
- Makropoulos, C. K., and D. Butler (2005), A multi-objective evolutionary programming approach to the “object location” spatial analysis and optimisation problem within the urban water management domain, *Civ. Eng. Environ. Syst.*, 22(2), 85–108, doi:10.1080/10286600500126280.
- Mammas, K., and D. Lekkas (2018), Rainfall Generation Using Markov Chain Models; Case Study: Central Aegean Sea, *Water*, 10(7), 856, doi:10.3390/w10070856.
- Mandelbrot, B. (1971), A Fast Fractional Gaussian Noise Generator, *Water Resour. Res.*, 7(3), 543–553, doi:10.1029/WR007i003p00543.
- Mandelbrot, B., and J. R. Wallis (1969a), Computer Experiments With Fractional Gaussian

## BIBLIOGRAPHY

- Noises: Part 1, Averages and Variances, *Water Resour. Res.*, 5(1), 228–241, doi:10.1029/WR005i001p00228.
- Mandelbrot, B., and J. R. Wallis (1969b), Computer Experiments with Fractional Gaussian Noises: Part 2, Rescaled Ranges and Spectra, *Water Resour. Res.*, 5(1), 242–259, doi:10.1029/WR005i001p00242.
- Mandelbrot, B., and J. R. Wallis (1969c), Computer Experiments with Fractional Gaussian Noises: Part 3, Mathematical Appendix, *Water Resour. Res.*, 5(1), 260–267, doi:10.1029/WR005i001p00260.
- Mann, H. B., and D. R. Whitney (1947), On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *Ann. Math. Stat.*, 18(1), 50–60, doi:10.1214/aoms/1177730491.
- Mardia, K. V (1970), Sankhya: A Translation Family of Bivariate Distributions and Frechet's Bounds, *Sankhya Indian J. Stat. Ser. A*, 32(1), 119–122.
- Mardia, K. V, and C. R. Goodall (1993), Spatial-temporal analysis of multivariate environmental monitoring data, *Multivar. Environ. Stat.*, 6(76), 347–385.
- Marriott, F. H. C., and J. A. Pope (1954), Bias in the Estimation of Autocorrelations, *Biometrika*, 41(3/4), 390, doi:10.2307/2332719.
- Matalas, N. . C., and J. R. Wallis (1976), *Generation of synthetic flow sequences, Systems Approach to Water Management*, edited by A. K. Biswas, McGraw-Hill, New York, New York.
- Matalas, N. C. (1967), Mathematical assessment of synthetic hydrology, *Water Resour. Res.*, 3(4), 937–945, doi:10.1029/WR003i004p00937.
- Matalas, N. C. (1975), Developments in stochastic hydrology, *Rev. Geophys.*, 13(3), 67, doi:10.1029/RG013i003p00067.
- Matalas, N. C., and J. R. Wallis (1971), Statistical Properties of Multivariate Fractional Noise Processes, *Water Resour. Res.*, 7(6), 1460–1468, doi:10.1029/WR007i006p01460.
- Matejka, J., and G. Fitzmaurice (2017), Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1290–1294, ACM.
- Matott, L. S., B. A. Tolson, and M. Asadzadeh (2012), A benchmarking framework for simulation-based optimization of environmental models, *Environ. Model. Softw.*, 35(0), 19–30, doi:http://dx.doi.org/10.1016/j.envsoft.2012.02.002.
- May, R. J., G. C. Dandy, H. R. Maier, and J. B. Nixon (2008), Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems, *Environ. Model. Softw.*, 23(10–11), 1289–1299, doi:10.1016/j.envsoft.2008.03.008.
- McMahon, T. A., and A. J. Miller (1971), Application of the Thomas and Fiering Model to Skewed Hydrologic Data, *Water Resour. Res.*, 7(5), 1338–1340, doi:10.1029/WR007i005p01338.
- McMahon, T. A., R. M. Vogel, M. C. Peel, and G. G. S. Pegram (2007), Global streamflows - Part 1: Characteristics of annual streamflows, *J. Hydrol.*, 347(3–4), 243–259, doi:10.1016/j.jhydrol.2007.09.002.



- Mehrotra, R. (2005), A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences, *J. Geophys. Res.*, 110(D16), D16108, doi:10.1029/2004JD005677.
- Mehrotra, R., and A. Sharma (2007), A semi-parametric model for stochastic generation of multi-site daily rainfall exhibiting low-frequency variability, *J. Hydrol.*, 335(1–2), 180–193, doi:10.1016/j.jhydrol.2006.11.011.
- Mehrotra, R., R. Srikanthan, and A. Sharma (2006), A comparison of three stochastic multi-site precipitation occurrence generators, *J. Hydrol.*, 331(1–2), 280–292, doi:10.1016/j.jhydrol.2006.05.016.
- Mehrotra, R., J. Li, S. Westra, and A. Sharma (2015), A programming tool to generate multi-site daily rainfall using a two-stage semi parametric model, *Environ. Model. Softw.*, 63, 230–239, doi:10.1016/j.envsoft.2014.10.016.
- Mejia, J. M., and J. Rousselle (1976), Disaggregation models in hydrology revisited, *Water Resour. Res.*, 12(2), 185–186, doi:10.1029/WR012i002p00185.
- Mejia, J. M., I. Rodriguez-Iturbe, and D. R. Dawdy (1972), Streamflow simulation: 2. The broken line process as a potential model for hydrologic simulation, *Water Resour. Res.*, 8(4), 931–941, doi:10.1029/WR008i004p00931.
- Mejía, J. M., and I. Rodríguez-Iturbe (1974), Correlation links between normal and log normal processes, *Water Resour. Res.*, 10(4), 689–690, doi:10.1029/WR010i004p00689.
- Menabde, M., D. Harris, A. Seed, G. Austin, and D. Stow (1997), Multiscaling properties of rainfall and bounded random cascades, *Water Resour. Res.*, doi:10.1029/97WR02006.
- Mhanna, M., and W. Bauwens (2012), A stochastic space-time model for the generation of daily rainfall in the Gaza Strip, *Int. J. Climatol.*, 32(7), 1098–1112, doi:10.1002/joc.2305.
- De Michele, C., and G. Salvadori (2003), A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas, *J. Geophys. Res.*, 108(D2), 4067, doi:10.1029/2002JD002534.
- Mikosch, T. (2005), *Copulas: Tales and facts*, Laboratory of Actuarial Mathematics, University of Copenhagen.
- Molnar, P., and P. Burlando (2005), Preservation of rainfall properties in stochastic disaggregation by a simple random cascade model, *Atmos. Res.*, 77(1–4 SPEC. ISS.), 137–151, doi:10.1016/j.atmosres.2004.10.024.
- Molz, F. J., H. H. Liu, and J. Szulga (1997), Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions, *Water Resour. Res.*, 33(10), 2273–2286, doi:10.1029/97WR01982.
- Momtahn, S., and A. B. Dariane (2007), Direct search approaches using genetic algorithms for optimization of water reservoir operating policies, *J. Water Resour. Plan. Manag.*, 133(3), 202–209, doi:10.1061/(asce)0733-9496(2007)133:3(202).
- Montanari, A., R. Rosso, and M. S. Taqqu (1997), Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation, *Water Resour. Res.*, 33(5), 1035, doi:10.1029/97WR00043.
- Montanari, A., R. Rosso, and M. S. Taqqu (2000), A seasonal fractional ARIMA Model applied to the Nile River monthly flows at Aswan, *Water Resour. Res.*, 36(5), 1249–1259,

## BIBLIOGRAPHY

doi:10.1029/2000WR900012.

- Montaseri, M., B. Amirataee, and R. Nawaz (2017), A Monte Carlo Simulation-Based Approach to Evaluate the Performance of three Meteorological Drought Indices in Northwest of Iran, *Water Resour. Manag.*, 31(4), 1323–1342, doi:10.1007/s11269-017-1580-2.
- Moran, P. a. P. (1969), Statistical Inference with Bivariate Gamma Distributions, *Biometrika*, 56(3), 627, doi:10.2307/2334670.
- Moran, P. A. P. (1970), Simulation and Evaluation of Complex Water Systems Operations, *Water Resour. Res.*, 6(6), 1737–1742, doi:10.1029/WR006i006p01737.
- Moschopoulos, P. G. (1985), The distribution of the sum of independent gamma random variables, *Ann. Inst. Stat. Math.*, 37(1), 541–544, doi:10.1007/BF02481123.
- Mostafa, M. D., and M. W. Mahmoud (1964), On the problem of estimation for the bivariate lognormal distribution, *Biometrika*, 51(3–4), 522–527, doi:10.1093/biomet/51.3-4.522.
- Moustakis, Y., P. Kossieris, I. Tsoukalas, and A. Efstratiadis (2017), Quasi-continuous stochastic simulation framework for flood modelling, in *EGU General Assembly Conference Abstracts*, vol. 19, p. 534.
- Mugunthan, P., and C. A. Shoemaker (2006), Assessing the impacts of parameter uncertainty for computationally expensive groundwater models, *Water Resour. Res.*, 42(10), W10428, doi:10.1029/2005WR004640.
- Mugunthan, P., C. A. Shoemaker, and R. G. Regis (2005), Comparison of function approximation, heuristic, and derivative-based methods for automatic calibration of computationally expensive groundwater bioremediation models, *Water Resour. Res.*, 41(11), 1–17, doi:10.1029/2005WR004134.
- Müller, H., and U. Haberlandt (2015), Temporal rainfall disaggregation : from point disaggregation to spatial rainfall, *J. Hydrol. Eng.*, 20(11), 3973, doi:10.1061/(ASCE)HE.1943-5584.0001195.
- Müller, H., and U. Haberlandt (2018), Temporal rainfall disaggregation using a multiplicative cascade model for spatial application in urban hydrology, *J. Hydrol.*, 556, 847–864, doi:10.1016/j.jhydrol.2016.01.031.
- Müller, J., and C. Shoemaker (2014), Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems, *J. Glob. Optim.*, 60(2), 123–144, doi:10.1007/s10898-014-0184-0.
- Myers, R. H., and D. C. Montgomery (1995), *Response Surface Methodology: Process and Product in Optimization Using Designed Experiments*, John Wiley & Sons, Inc.
- Nalbantis, I., and D. Koutsoyiannis (1997), A parametric rule for planning and management of multiple-reservoir systems, *Water Resour. Res.*, 33(9), 2165–2177, doi:10.1029/97WR01034.
- Nalbantis, I., A. Efstratiadis, E. Rozos, M. Kopsiafti, and D. Koutsoyiannis (2011), Holistic versus monomeric strategies for hydrological modelling of human-modified hydrosystems, *Hydrol. Earth Syst. Sci.*, 15(3), 743–758, doi:10.5194/hess-15-743-2011.
- Nataf, A. (1962), Statistique mathématique-determination des distributions de probabilités

- dont les marges sont donnees, *C. R. Acad. Sci. Paris*, 255(1), 42–43.
- Nazemi, A., H. S. Wheater, K. P. Chun, and A. Elshorbagy (2013), A stochastic reconstruction framework for analysis of water resource system vulnerability to climate-induced changes in river flow regime, *Water Resour. Res.*, 49(1), 291–305, doi:10.1029/2012WR012755.
- Nelder, J. A., and R. Mead (1965), A Simplex Method for Function Minimization, *Comput. J.*, 7(4), 308–313, doi:10.1093/comjnl/7.4.308.
- Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.
- Neykov, N. M., P. N. Neytchev, and W. Zucchini (2014), Stochastic daily precipitation model with a heavy-tailed component, *Nat. Hazards Earth Syst. Sci.*, 14(9), 2321–2335, doi:10.5194/nhess-14-2321-2014.
- Nicklow, J. et al. (2010), State of the Art for Genetic Algorithms and Beyond in Water Resources Planning and Management, *J. Water Resour. Plan. Manag.*, 136(4), 412–432, doi:10.1061/(ASCE)WR.1943-5452.0000053.
- O’Connell, P. E. (1974), Stochastic modelling of long-term persistence in streamflow sequences,
- O’Connell, P. E., D. Koutsoyiannis, H. F. Lins, Y. Markonis, A. Montanari, and T. Cohn (2016), The scientific legacy of Harold Edwin Hurst (1880–1978), *Hydrol. Sci. J.*, 61(9), 1571–1590, doi:10.1080/02626667.2015.1125998.
- Obeysekera, J. T. B., and V. Yevjevich (1985), A Note on Simulation of Samples of Gamma-Autoregressive Variables, *Water Resour. Res.*, 21(10), 1569–1572, doi:10.1029/WR021i010p01569.
- Oliveira, R., and D. P. Loucks (1997), Operating rules for multireservoir systems, *Water Resour. Res.*, 33(4), 839–852.
- Olsson, J. (1998), Evaluation of a scaling cascade model for temporal rainfall disaggregation, *Hydrol. Earth Syst. Sci.*, doi:10.5194/hess-2-19-1998.
- Onof, C., and H. S. Wheater (1994a), Improved fitting of the bartlett-lewis rectangular pulse model for hourly rainfall, *Hydrol. Sci. J.*, 39(6), 663–680, doi:10.1080/02626669409492786.
- Onof, C., and H. S. Wheater (1994b), Improvements to the modelling of British rainfall using a modified Random Parameter Bartlett-Lewis Rectangular Pulse Model, *J. Hydrol.*, 157(1–4), 177–195, doi:10.1016/0022-1694(94)90104-X.
- Onof, C., and H. S. Wheater (1995), Modelling of rainfall time-series using the Barlett-Lewis model, *Proc. Inst. Civ. Eng. - Water Marit. Energy*, 112(4), 362–374, doi:10.1680/iwtme.1995.28116.
- Onof, C., R. E. Chandler, A. Kakou, P. Northrop, H. S. Wheater, and V. Isham (2000), Rainfall modelling using Poisson-cluster processes: a review of developments, *Stoch. Environ. Res. Risk Assess.*, 14(6), 0384–0411, doi:10.1007/s004770000043.
- Onof, C., J. Townend, and R. Kee (2005), Comparison of two hourly to 5-min rainfall disaggregators, *Atmos. Res.*, 77(1–4), 176–187, doi:10.1016/j.atmosres.2004.10.022.
- Ostfeld, A., and S. Salomons (2005), A hybrid genetic - instance based learning algorithm for CE-QUAL-W2 calibration, *J. Hydrol.*, 310(1–4), 122–142,

## BIBLIOGRAPHY

- doi:10.1016/j.jhydrol.2004.12.004.
- Pan, L., and L. Wu (1998), A hybrid global optimization method for inverse estimation of hydraulic parameters: Annealing-Simplex Method, *Water Resour. Res.*, 34(9), 2261–2269, doi:10.1029/98WR01672.
- Papalexiou, S. M. (2018), Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency, *Adv. Water Resour.*, doi:10.1016/j.advwatres.2018.02.013.
- Papalexiou, S. M., and D. Koutsoyiannis (2012), Entropy based derivation of probability distributions: A case study to daily rainfall, *Adv. Water Resour.*, 45, 51–57, doi:10.1016/j.advwatres.2011.11.007.
- Papalexiou, S. M., and D. Koutsoyiannis (2013), Battle of extreme value distributions : A global survey on extreme daily rainfall, *Water Resour. Res.*, 49(1), 187–201, doi:10.1029/2012WR012557.
- Papalexiou, S. M., and D. Koutsoyiannis (2016), A global survey on the seasonal variation of the marginal distribution of daily precipitation, *Adv. Water Resour.*, 94, 131–145, doi:10.1016/j.advwatres.2016.05.005.
- Papalexiou, S. M., D. Koutsoyiannis, and A. Montanari (2011), Can a simple stochastic model generate rich patterns of rainfall events?, *J. Hydrol.*, 411(3–4), 279–289, doi:10.1016/j.jhydrol.2011.10.008.
- Papalexiou, S. M., D. Koutsoyiannis, and C. Makropoulos (2013), How extreme is extreme? An assessment of daily rainfall distribution tails, *Hydrol. Earth Syst. Sci.*, 17(2), 851–862, doi:10.5194/hess-17-851-2013.
- Papoulis, A. (1991), *Probability, Random Variables, and Stochastic Processes*, Third edit., McGraw-Hill Series in Electrical Engineering. New York City, New York, USA: McGraw-Hill.
- Paraskevopoulos – Pangaea (1994), *Environmental Impact Assessment for the wider region of the Greek Nestos River Basin*.
- Paschalis, A., S. Fatichi, P. Molnar, S. Rimkus, and P. Burlando (2014), On the effects of small scale space–time variability of rainfall on basin flood response, *J. Hydrol.*, 514, 313–327, doi:10.1016/j.jhydrol.2014.04.014.
- Pattison, A. (1965), Synthesis of hourly rainfall data, *Water Resour. Res.*, 1(4), 489–498, doi:10.1029/WR0011004p00489.
- Pearson, K. (1920), Notes on the history of correlation, *Biometrika*, 13(1), 25–45.
- Pegram, G. G. S., and W. James (1972), Multilag multivariate autoregressive model for the generation of operational hydrology, *Water Resour. Res.*, 8(4), 1074–1076, doi:10.1029/WR0081004p01074.
- Ponweiser, W., T. Wagner, D. Biermann, and M. Vincze (2008), Multiobjective Optimization on a Limited Budget of Evaluations Using Model-Assisted  $\mathcal{S}$ -Metric Selection, in *Parallel Problem Solving from Nature – PPSN X*, vol. 5199, edited by G. Rudolph, T. Jansen, S. Lucas, C. Poloni, and N. Beume, pp. 784–794, Springer Berlin Heidelberg.
- Powell, M. J. D. (1992), *The theory of radial basis function approximation in 1990*, Light, Ed Advances in Numerical Analysis Advances in Numerical Analysis, vol. 2:

- wavelets, subdivision algorithms and radial basis functions. Oxford University Press, Oxford, pp. 105-210.
- Prairie, J., B. Rajagopalan, U. Lall, and T. Fulp (2007), A stochastic nonparametric technique for space-time disaggregation of streamflows, *Water Resour. Res.*, 43(3), 1-10, doi:10.1029/2005WR004721.
- Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (1992), *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge.
- Psarrou, E., I. Tsoukalas, and C. Makropoulos (2018), A Monte-Carlo-Based Method for the Optimal Placement and Operation Scheduling of Sewer Mining Units in Urban Wastewater Networks, *Water*, 10(2), 200, doi:10.3390/w10020200.
- Pui, A., A. Sharma, R. Mehrotra, B. Sivakumar, and E. Jeremiah (2012), A comparison of alternatives for daily to sub-daily rainfall disaggregation, *J. Hydrol.*, 470-471, 138-157, doi:10.1016/j.jhydrol.2012.08.041.
- Qian, B., J. Corte-Real, and H. Xu (2002), Multisite stochastic weather models for impact studies, *Int. J. Climatol.*, 22(11), 1377-1397, doi:10.1002/joc.808.
- Qin, X. S., and Y. Lu (2014), Study of Climate Change Impact on Flood Frequencies: A Combined Weather Generator and Hydrological Modeling Approach, *J. Hydrometeorol.*, 15(3), 1205-1219, doi:10.1175/JHM-D-13-0126.1.
- Queipo, N. V., R. T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, and P. Kevin Tucker (2005), Surrogate-based analysis and optimization, *Prog. Aerosp. Sci.*, 41(1), 1-28, doi:10.1016/j.paerosci.2005.02.001.
- Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res.*, 35(10), 3089-3101, doi:10.1029/1999WR900028.
- Rasmussen, P. F. (2013), Multisite precipitation generation using a latent autoregressive model, *Water Resour. Res.*, 49(4), 1845-1857, doi:10.1002/wrcr.20164.
- Rasmussen, P. F., J. D. Salas, L. Fagherazzi, J.-C. Rassam, and B. Bobée (1996), Estimation and validation of contemporaneous PARMA Models for streamflow simulation, *Water Resour. Res.*, 32(10), 3151-3160, doi:10.1029/96WR01528.
- Razavi, S., B. A. Tolson, L. S. Matott, N. R. Thomson, A. MacLean, and F. R. Seglenieks (2010), Reducing the computational cost of automatic calibration through model preemption, *Water Resour. Res.*, 46(11), W11523, doi:10.1029/2009WR008957.
- Razavi, S., B. A. Tolson, and D. H. Burn (2012a), Numerical assessment of metamodelling strategies in computationally intensive optimization, *Environ. Model. Softw.*, 34(0), 67-86, doi:http://dx.doi.org/10.1016/j.envsoft.2011.09.010.
- Razavi, S., B. A. Tolson, and D. H. Burn (2012b), Review of surrogate modeling in water resources, *Water Resour. Res.*, 48(7), doi:10.1029/2011WR011527.
- Reddy, P. J. R. (1997), *Stochastic hydrology*, Laxmi Publications, Ltd.
- Reed, P. M., D. Hadka, J. D. Herman, J. R. Kasprzyk, and J. B. Kollat (2013), Evolutionary multiobjective optimization in water resources: The past, present, and future, *Adv. Water Resour.*, 51, 438-456, doi:10.1016/j.advwatres.2012.01.005.

## BIBLIOGRAPHY

- Regis, R. ., and C. Shoemaker (2013), Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization, *Eng. Optim.*, 45(5), 529–555, doi:10.1080/0305215x.2012.687731.
- Regis, R. G. (2011), Stochastic radial basis function algorithms for large-scale optimization involving expensive black-box objective and constraint functions, *Comput. Oper. Res.*, 38(5), 837–853, doi:10.1016/j.cor.2010.09.013.
- Regis, R. G. (2014), Particle swarm with radial basis function surrogates for expensive black-box optimization, *J. Comput. Sci.*, 5(1), 12–23, doi:http://dx.doi.org/10.1016/j.jocs.2013.07.004.
- Regis, R. G., and C. A. Shoemaker (2004), Local function approximation in evolutionary algorithms for the optimization of costly functions, *IEEE Trans. Evol. Comput.*, 8(5), 490–505, doi:10.1109/tevc.2004.835247.
- Regis, R. G., and C. A. Shoemaker (2007a), A stochastic radial basis function method for the global optimization of expensive functions, *INFORMS J. Comput.*, 19(4), 497–509, doi:10.1287/ioc.1060.0182.
- Regis, R. G., and C. A. Shoemaker (2007b), Improved strategies for radial basis function methods for global optimization, *J. Glob. Optim.*, 37(1), 113–135, doi:10.1007/s10898-006-9040-1.
- Regis, R. G., and C. A. Shoemaker (2009), Parallel Stochastic Global Optimization Using Radial Basis Functions, *INFORMS J. Comput.*, 21(3), 411–426, doi:10.1287/ijoc.1090.0325.
- Renard, B., and M. Lang (2007), Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology, *Adv. Water Resour.*, 30(4), 897–912, doi:10.1016/j.advwatres.2006.08.001.
- Richardson, C. W., and D. A. Wright (1984), WGEN: A model for generating daily weather variables,
- Richardson, C. W. C. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resour. Res.*, 17(1), 182–190, doi:10.1029/WR017i001p00182.
- Rios, L. M., and N. V. Sahinidis (2013), Derivative-free optimization: A review of algorithms and comparison of software implementations, in *Journal of Global Optimization*, vol. 56, pp. 1247–1293.
- Robert, C., and G. Casella (2010), *Introducing Monte Carlo Methods with R*, Springer New York, New York, NY.
- Rodriguez-Iturbe, I., D. R. Cox, and V. Isham (1987), Some Models for Rainfall Based on Stochastic Point Processes, *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 410(1839), 269–288, doi:10.1098/rspa.1987.0039.
- Rodriguez-Iturbe, I., D. R. Cox, and V. Isham (1988), A Point Process Model for Rainfall: Further Developments, *Proc. R. Soc. A Math. Phys. Eng. Sci.*, 417(1853), 283–298, doi:10.1098/rspa.1988.0061.
- Rodríguez-Iturbe, I., and J. M. Mejía (1974), The design of rainfall networks in time and space, *Water Resour. Res.*, 10(4), 713–728, doi:10.1029/WR010i004p00713.
- Roldan, J., and D. A. Woolhiser (1982), Stochastic daily precipitation models: 1. A comparison of occurrence processes, *Water Resour. Res.*, 18(5), 1451–1459,

doi:10.1029/WR018i005p01451.

- Rozos, E., A. Efstratiadis, I. Nalbantis, and D. Koutsoyiannis (2004), Calibration of a semi-distributed model for conjunctive simulation of surface and groundwater flows / Calage d'un modèle semi-distribué pour la simulation conjointe d'écoulements superficiels et souterrains, *Hydrol. Sci. J.*, 49(5), null-842, doi:10.1623/hysj.49.5.819.55130.
- Rupp, D. E., R. F. Keim, M. Ossiander, M. Brugnach, and J. S. Selker (2009), Time scale and intensity dependency in multiplicative cascades for temporal rainfall disaggregation, *Water Resour. Res.*, doi:10.1029/2008WR007321.
- Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989), Design and Analysis of Computer Experiments, *Stat. Sci.*, 4(4), 409–423, doi:10.1214/ss/1177012413.
- Salas, J. D. (1993), Analysis and modeling of hydrologic time series, in *Handbook of hydrology*, edited by D. R. Maidment, p. Ch. 19.1-19.72, Mc-Graw-Hill, Inc.
- Salas, J. D., and M. W. Abdelmohsen (1993), Initialization for generating single-site and multisite low-order periodic autoregressive and moving average processes, *Water Resour. Res.*, 29(6), 1771–1776, doi:10.1029/93WR00371.
- Salas, J. D., and T. Lee (2010), Nonparametric simulation of single-site seasonal streamflows, *J. Hydrol. Eng.*, 15(4), 284–296, doi:10.1061/(ASCE)HE.1943-5584.0000189.
- Salas, J. D., and G. G. S. Pegram (1977), A seasonal multivariate multilag autoregressive model in hydrology, in *Proc. Third Int. Symp. on Theoretical and Applied Hydrology, Colorado State Univ., Fort Collins, CO, USA*.
- Salas, J. D., and R. a Pielke (2003), Stochastic characteristics modeling of hydroclimatic processes, in *Handbook of Weather, Climate, and Water: Atmospheric Chemistry, Hydrology, and Societal Impact*, vol. 2, edited by T. Potter and B. Colman, pp. 587–605, John Wiley & Sons, Hoboken, New Jersey.
- Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane (1980), *Applied modeling of hydrologic time series*, 2nd Print., Water Resources Publication, Littleton, Colorado.
- Salas, J. D., D. C. Boes, and R. A. Smith (1982), Estimation of ARMA Models with seasonal parameters, *Water Resour. Res.*, 18(4), 1006–1010, doi:10.1029/WR018i004p01006.
- Salas, J. D., G. Q. Tabios, and P. Bartolini (1985), Approaches to multivariate modeling of water resources time series, *J. Am. Water Resour. Assoc.*, 21(4), 683–708, doi:10.1111/j.1752-1688.1985.tb05383.x.
- Salas, J. D., O. G. Sveinsson, W. L. Lane, and D. K. Frevert (2006), Stochastic Streamflow Simulation Using SAMS-2003, *J. Irrig. Drain. Eng.*, 132(2), 112–122, doi:10.1061/(ASCE)0733-9437(2006)132:2(112).
- Salvadori, G., and C. De Michele (2004), Frequency analysis via copulas: Theoretical aspects and applications to hydrological events, *Water Resour. Res.*, 40(12), doi:10.1029/2004WR003133.
- Salvadori, G., and C. De Michele (2007), On the Use of Copulas in Hydrology: Theory and Practice, *J. Hydrol. Eng.*, 12(4), 369–380, doi:10.1061/(ASCE)1084-0699(2007)12:4(369).
- Samoradnitsky, G. (2017), *Stable non-Gaussian random processes: stochastic models with infinite variance*, Routledge.

## BIBLIOGRAPHY

- Santana-Quintero, L., A. Montaña, and C. Coello (2010), A Review of Techniques for Handling Expensive Functions in Evolutionary Multi-Objective Optimization, in *Computational Intelligence in Expensive Optimization Problems Adaptation Learning and Optimization*, vol. 2, edited by Y. Tenne and Goh, C., pp. 29–59, Springer, Berlin Heidelberg.
- Santner, T. J., B. Williams, and W. Notz (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag.
- Sasena, M., P. Papalambros, and P. Goovaerts (2002), Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization, *Eng. Optim.*, 34(3), 263–278, doi:10.1080/03052150211751.
- Savic, D. A., and G. A. Walters (1997), Genetic Algorithms for Least-Cost Design of Water Distribution Networks, *J. Water Resour. Plan. Manag.*, 123(2), 67–77, doi:10.1061/(ASCE)0733-9496(1997)123:2(67).
- Schutte, J. F., J. A. Reinbolt, B. J. Fregly, R. T. Haftka, and A. D. George (2004), Parallel global optimization with the particle swarm algorithm, *Int. J. Numer. Methods Eng.*, 61(13), 2296–2315, doi:10.1002/nme.1149.
- Segond, M.-L., C. Onof, and H. S. Wheeler (2006), Spatial–temporal disaggregation of daily rainfall from a generalized linear model, *J. Hydrol.*, 331(3–4), 674–689, doi:10.1016/j.jhydrol.2006.06.019.
- Semenov, M. A., and E. M. Barrow (1997), Use of a stochastic weather generator in the development of climate change scenarios, *Clim. Change*, 35(4), 397–414, doi:10.1023/A:1005342632279.
- Semenov, M. A., R. J. Brooks, E. M. Barrow, and C. W. Richardson (1998), Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates, *Clim. Res.*, 10(2), 95–107, doi:10.3354/cr010095.
- Serinaldi, F. (2009a), A multisite daily rainfall generator driven by bivariate copula-based mixed distributions, *J. Geophys. Res.*, 114(D10), D10103, doi:10.1029/2008JD011258.
- Serinaldi, F. (2009b), Copula-based mixed models for bivariate rainfall data: an empirical study in regression perspective, *Stoch. Environ. Res. Risk Assess.*, 23(5), 677–693, doi:10.1007/s00477-008-0249-z.
- Serinaldi, F., and C. G. Kilsby (2014), Simulating daily rainfall fields over large areas for collective risk estimation, *J. Hydrol.*, 512, 285–302, doi:10.1016/j.jhydrol.2014.02.043.
- Serinaldi, F., and F. Lombardo (2017), BetaBit: A fast generator of autocorrelated binary processes for geophysical research, *EPL (Europhysics Lett.)*, 118(3), 30007, doi:10.1209/0295-5075/118/30007.
- Shan, S., and G. G. Wang (2010), Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions, *Struct. Multidiscip. Optim.*, 41(2), 219–241, doi:10.1007/s00158-009-0420-2.
- Shao, Q., and R. Lund (2004), Computation and characterization of autocorrelations and partial autocorrelations in periodic arma models, *J. Time Ser. Anal.*, 25(3), 359–372, doi:10.1111/j.1467-9892.2004.00356.x.
- Shao, Q., L. Zhang, and Q. J. Wang (2016), A hybrid stochastic-weather-generation method for temporal disaggregation of precipitation with consideration of seasonality and within-



- month variations, *Stoch. Environ. Res. Risk Assess.*, 30(6), 1705–1724, doi:10.1007/s00477-015-1177-3.
- Sharma, A., D. G. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, 33(2), 291–308, doi:10.1029/96WR02839.
- Shoemaker, C. A., R. G. Regis, and R. C. Fleming (2007), Watershed calibration using multistart local optimization and evolutionary optimization with radial basis function approximation, *Hydrol. Sci. J.*, 52(3), 450–465, doi:10.1623/hysj.52.3.450.
- Shumway, R. H., and D. S. Stoffer (2017), *Time Series Analysis and Its Applications*, Springer Texts in Statistics, Springer International Publishing, Cham.
- Simonovic, S. P. (1992), RESERVOIR SYSTEMS-ANALYSIS - CLOSING GAP BETWEEN THEORY AND PRACTICE, *J. Water Resour. Plan. Manag.*, 118(3), 262–280, doi:10.1061/(asce)0733-9496(1992)118:3(262).
- Singh, S., and G. Maddala (1976), A function for size distribution of incomes, *Econometrica*, 44, 963–970, doi:10.2307/1910422.
- Sklar, A. (1973), Random variables, joint distribution functions, and copulas, *Kybernetika*, 9(6), 449–460.
- Sklar, M. (1959), Fonctions de repartition an dimensions et leurs marges, *Publ. Inst. Stat. Univ. Paris*, 8, 229–231.
- Skoulikaris, C., M. Monget, and J. Ganoulis (2008), Climate Change Impacts on Dams Projects on Transboundary River Basins. The Case of Mesta/Nestos River Basin, Greece, *IV Int. Symp. Transbound. Waters Manag.*
- Smithers, J. C., G. G. S. Pegram, and R. E. Schulze (2002), Design rainfall estimation in South Africa using Bartlett-Lewis rectangular pulse rainfall models, *J. Hydrol.*, 258(1–4), 83–99, doi:10.1016/S0022-1694(01)00571-6.
- Song, W. T., L. C. Hsiao, and Y. J. Chen (1996), Generating pseudo-random time series with specified marginal distributions, *Eur. J. Oper. Res.*, 94(1), 194–202, doi:10.1016/0377-2217(95)00206-5.
- Srikanthan, R., and T. a. McMahon (2001), Stochastic generation of annual, monthly and daily climate data: A review, *Hydrol. Earth Syst. Sci.*, 5(4), 653–670, doi:10.5194/hess-5-653-2001.
- Srikanthan, R., and G. G. S. Pegram (2009), A nested multisite daily rainfall stochastic generation model, *J. Hydrol.*, 371(1–4), 142–153, doi:10.1016/j.jhydrol.2009.03.025.
- Stacy, E. W. (1962), A Generalization of the Gamma Distribution, *Ann. Math. Stat.*, 33(3), 1187–1192, doi:10.1214/aoms/1177704481.
- Stedinger, J. R., and R. M. Vogel (1984), Disaggregation Procedures for Generating Serially Correlated Flow Vectors, *Water Resour. Res.*, 20(1), 47–56, doi:10.1029/WR020i001p00047.
- Stern, R. D., and R. Coe (1984), A Model Fitting Analysis of Daily Rainfall Data, *J. R. Stat. Soc. Ser. A*, 147(1), 1, doi:10.2307/2981736.
- Sudler, C. E. (1927), Storage required for the regulation of stream flow, in *Proceedings of the American Society of Civil Engineers*, vol. 52, pp. 1917–1955, ASCE.

## BIBLIOGRAPHY

- Sudret, B. (2008), Global sensitivity analysis using polynomial chaos expansions, *Reliab. Eng. Syst. Saf.*, 93(7), 964–979, doi:10.1016/j.ress.2007.04.002.
- Sveinsson, O., J. D. Salas, W. Lane, and D. Frevert (2007), Stochastic Analysis, Modeling, and Simulation (SAMS) Version 2007, User's Manual, *Comput. Hydrol. Lab. Dep. Civ. Environ. Eng. Color. State Univ. Fort Collins, Color.*, (11).
- Tadikamalla, P. R. (1980), A look at the Burr and related distributions, *Int. Stat. Rev. Int. Stat.*, 337–344.
- Tan, C.-C., C.-P. Tung, C.-H. Chen, and W. W. G. Yeh (2008), An integrated optimization algorithm for parameter structure identification in groundwater modeling, *Adv. Water Resour.*, 31(3), 545–560, doi:10.1016/j.advwatres.2007.11.007.
- Tang, Y., J. Chen, and J. Wei (2012), A surrogate-based particle swarm optimization algorithm for solving optimization problems with expensive black box functions, *Eng. Optim.*, 45(5), 557–576, doi:10.1080/0305215X.2012.690759.
- Tegos, A., N. Malamos, A. Efstratiadis, I. Tsoukalas, A. Karanasios, and D. Koutsoyiannis (2017), Parametric Modelling of Potential Evapotranspiration: A Global Survey, *Water*, 9(12), 795, doi:10.3390/w9100795.
- Tessier, Y., S. Lovejoy, P. Hubert, D. Schertzer, and S. Pecknold (1996), Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions, *J. Geophys. Res. Atmos.*, 101(D21), 26427–26440, doi:10.1029/96JD01799.
- Thomas, H. A., and R. P. Burden (1963), *Operations research in water quality management*, HARVARD UNIV CAMBRIDGE MASS DIV OF ENGINEERING AND APPLIED PHYSICS.
- Thomas, H. A., and M. B. Fiering (1962), Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, *Des. water Resour. Syst.*, 459–493.
- Thomas, H. A., and M. B. Fiering (1963), The nature of the storage yield function, *Oper. Res. Water Qual. Manag.*
- Thompson, C. S., P. J. Thomson, and X. Zheng (2007), Fitting a multisite daily rainfall model to New Zealand data, *J. Hydrol.*, 340(1–2), 25–39, doi:10.1016/j.jhydrol.2007.03.020.
- Tiao, G. C., and M. R. Grupe (1980), Hidden periodic autoregressive-moving average models in time series data, *Biometrika*, 67(2), 365–373.
- Todini, E. (1980), The preservation of skewness in linear disaggregation schemes, *J. Hydrol.*, 47(3–4), 199–214, doi:10.1016/0022-1694(80)90093-1.
- Todorovic, P., and D. A. Woolhiser (1975), A stochastic model of n-day precipitation, *J. Appl. Meteorol.*, 14(1), 17–24.
- Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, 43(1), doi:10.1029/2005WR004723.
- Tolson, B. A., M. Asadzadeh, H. R. Maier, and A. Zecchin (2009), Hybrid discrete dynamically dimensioned search (HD-DDS) algorithm for water distribution system design optimization, *Water Resour. Res.*, 45(12), W12416, doi:10.1029/2008WR007673.
- Troutman, B. M. (1979), Some results in periodic autoregression, *Biometrika*, 66(2), 219–228.

- Tsay, R. S. (2013), *Multivariate Time Series Analysis: with R and financial applications*, John Wiley & Sons, Hoboken, New Jersey.
- Tsoukalas, I., and C. Makropoulos (2013), Hydrosystem optimization with the use of evolutionary algorithms: The case of Nestos river, in *13th International Conference on Environmental Science and Technology*, Athens, Greece.
- Tsoukalas, I., and C. Makropoulos (2015a), A Surrogate Based Optimization Approach for the Development of Uncertainty-Aware Reservoir Operational Rules: the Case of Nestos Hydrosystem, *Water Resour. Manag.*, 29(13), 4719–4734, doi:10.1007/s11269-015-1086-8.
- Tsoukalas, I., and C. Makropoulos (2015b), Multiobjective optimisation on a budget: Exploring surrogate modelling for robust multi-reservoir rules generation under hydrological uncertainty, *Environ. Model. Softw.*, 69, 396–413, doi:10.1016/j.envsoft.2014.09.023.
- Tsoukalas, I., P. Kossieris, A. Efstratiadis, and C. Makropoulos (2015a), Handling time-expensive global optimization problems through the surrogate-enhanced evolutionary annealing-simplex algorithm, in *EGU General Assembly Conference Abstracts*, vol. 17.
- Tsoukalas, I., P. Dimas, and C. Makropoulos (2015b), Hydrosystem optimization on a budget: Investigating the potential of surrogate based optimization techniques, *J. | MESA; Vol 6 No 4*.
- Tsoukalas, I., P. Kossieris, A. Efstratiadis, and C. Makropoulos (2016), Surrogate-enhanced evolutionary annealing simplex algorithm for effective and efficient optimization of water resources problems on a budget, *Environ. Model. Softw.*, 77, 122–142, doi:10.1016/j.envsoft.2015.12.008.
- Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2017a), Stochastic simulation of periodic processes with arbitrary marginal distributions, in *15th International Conference on Environmental Science and Technology. CEST 2017.*, Rhodes, Greece.
- Tsoukalas, I., S. Papalexiou, A. Efstratiadis, and C. Makropoulos (2018a), A Cautionary Note on the Reproduction of Dependencies through Linear Stochastic Models with Non-Gaussian White Noise, *Water*, 10(6), 771, doi:10.3390/w10060771.
- Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2018b), Building a puzzle to solve a riddle: a new approach to multi-temporal stochastic simulation, *J. Hydrol.*, doi:(in review).
- Tsoukalas, I., P. Kossieris, A. Efstratiadis, C. Makropoulos, and D. Koutsoyiannis (2018c), CastaliaR: An R package for multivariate stochastic simulation at multiple temporal scales, in *European Geosciences Union General Assembly 2018, Geophysical Research Abstracts, Vol. 20, Vienna, EGU2018-18433, European Geosciences Union*.
- Tsoukalas, I., C. Makropoulos, and D. Koutsoyiannis (2018d), Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions, *Water Resour. Res.*, doi:10.1029/2017WR022462.
- Tsoukalas, I., A. Efstratiadis, and C. Makropoulos (2018e), Stochastic Periodic Autoregressive to Anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions, *Water Resour. Res.*, 54(1), 161–185, doi:10.1002/2017WR021394.
- Tsoukalas, I. K., C. K. Makropoulos, and S. N. Michas (2017b), Identification of potential sewer mining locations: a Monte-Carlo based approach, *Water Sci. Technol.*, 76(12), 3351–3357,

## BIBLIOGRAPHY

doi:10.2166/wst.2017.487.

- Tweedie, M. C. K. (1984), An index which distinguishes between some important exponential families, in *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, vol. 579, p. 604.
- Tyrallis, H., and D. Koutsoyiannis (2011), Simultaneous estimation of the parameters of the Hurst–Kolmogorov stochastic process, *Stoch. Environ. Res. Risk Assess.*, 25(1), 21–33, doi:10.1007/s00477-010-0408-x.
- Unal, N. E., H. Aksoy, and T. Akar (2004), Annual and monthly rainfall data generation schemes, *Stoch. Environ. Res. Risk Assess.*, 18(4), doi:10.1007/s00477-004-0186-4.
- Valencia, R. D. V., and J. C. Schakke (1973), Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, 9(3), 580–585, doi:10.1029/WR009i003p00580.
- Vecchia, A. V. (1985), Periodic AutoRegressive-Moving Average (PARMA) modeling with applications to water resources, *J. Am. Water Resour. Assoc.*, 21(5), 721–730, doi:10.1111/j.1752-1688.1985.tb00167.x.
- Veneziano, D., A. Langousis, and P. Furcolo (2006), Multifractality and rainfall extremes: A review, *Water Resour. Res.*, doi:10.1029/2005WR004716.
- Vink, K., and P. Schot (2002), Multiple-objective optimisation of drinking water production strategies using a genetic algorithm, *Water Resour. Res.*, 38(9), 1181.
- Vogel, R. M., and J. R. Stedinger (1988), The value of stochastic streamflow models in overyear reservoir design applications, *Water Resour. Res.*, 24(9), 1483–1490, doi:10.1029/WR024i009p01483.
- Vrugt, J. A., B. O Nuallain, B. A. Robinson, W. Bouten, S. C. Dekker, and P. M. A. Sloot (2006), Application of parallel computing to stochastic parameter estimation in environmental models, *Comput. Geosci.*, 32(8), 1139–1155, doi:10.1016/j.cageo.2005.10.015.
- Wagner, T., M. Emmerich, A. Deutz..., A. Deutz, and W. Ponweiser (2010), On Expected-Improvement Criteria for Model-based Multi-objective Optimization, in *Parallel Problem Solving from Nature, PPSN XI*, vol. 6238, edited by R. Schaefer, C. Cotta, J. Kołodziej, and G. Rudolph, pp. 718–727, Springer Berlin Heidelberg.
- Wang, W.-C., K.-W. Chau, C.-T. Cheng, and L. Qiu (2009), A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series, *J. Hydrol.*, 374(3–4), 294–306, doi:10.1016/j.jhydrol.2009.06.019.
- Wang, Y., C. Li, J. Liu, F. Yu, Q. Qiu, J. Tian, and M. Zhang (2017), Multivariate Analysis of Joint Probability of Different Rainfall Frequencies Based on Copulas, *Water*, 9(3), 198, doi:10.3390/w9030198.
- Wheater, H. S., R. E. Chandler, C. J. Onof, V. S. Isham, E. Bellone, C. Yang, D. Lekkas, G. Lourmas, and M.-L. Segond (2005), Spatial-temporal rainfall modelling for flood risk estimation, *Stoch. Environ. Res. Risk Assess.*, 19(6), 403–416, doi:10.1007/s00477-005-0011-8.
- While, L., L. Bradstreet, and L. Barone (2012), A Fast Way of Calculating Exact Hypervolumes, *Evol. Comput. IEEE Trans.*, 16(1), 86–95, doi:10.1109/TEVC.2010.2077298.
- Whitt, W. (1976), Bivariate Distributions with Given Marginals, *Ann. Stat.*, 4(6), 1280–1289, doi:10.1214/aos/1176343660.

- Wilby, R. L. (1994), Stochastic weather type simulation for regional climate change impact assessment, *Water Resour. Res.*, 30(12), 3395–3403, doi:10.1029/94WR01840.
- Wilby, R. L., T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks (1998), Statistical downscaling of general circulation model output: a comparison of methods, *Water Resour. Res.*, 34(11), 2995–3008.
- Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol.*, 210(1–4), 178–191, doi:10.1016/S0022-1694(98)00186-3.
- Wilks, D. S. (1999), Interannual variability and extreme-value characteristics of several stochastic daily precipitation models, *Agric. For. Meteorol.*, 93(3), 153–169, doi:10.1016/S0168-1923(98)00125-7.
- Wilks, D. S., and R. L. Wilby (1999), The weather generation game: a review of stochastic weather models, *Prog. Phys. Geogr.*, 23(3), 329–357, doi:10.1191/030913399666525256.
- Williams, P. (1998), Modelling seasonality and trends in daily rainfall data, *Adv. Neural Inf. Process. Syst.*, 10, 985–991.
- Wójcik, R., and T. A. Buishand (2003), Simulation of 6-hourly rainfall and temperature by two resampling schemes, *J. Hydrol.*, 273(1–4), 69–80, doi:10.1016/S0022-1694(02)00355-4.
- Xiao, Q. (2014), Evaluating correlation coefficient for Nataf transformation, *Probabilistic Eng. Mech.*, 37, 1–6, doi:10.1016/j.pro bengmech.2014.03.010.
- Yaglom, A. M. (1962), *An introduction to the theory of stationary random functions*, Courier Corporation.
- Yan, S., and B. Minsker (2006), Optimal groundwater remediation design using an Adaptive Neural Network Genetic Algorithm, *Water Resour. Res.*, 42(5), doi:10.1029/2005WR004303.
- Yan, S., and B. Minsker (2011), Applying Dynamic Surrogate Models in Noisy Genetic Algorithms to Optimize Groundwater Remediation Designs, *J. Water Resour. Plan. Manag.*, 137(3), 284–292, doi:10.1061/(ASCE)WR.1943-5452.0000106.
- Yates, D., J. Sieber, D. Purkey, A. Huber-Lee, Yates D., Sieber J., D. Purkey, and A. Huber-Lee (2005), WEAP21: A Demand, priority, and preference driver water planning model. Part 1: Model Characteristics, *Water Int.*, 30, 487–500.
- YDE (1954), *Nestos diversion dam, Macedonia, Greece. Basis of design on the Nestos diversion dam*, Library of Technical Chamber of Greece.
- Yeh, W. W. G. (1985), RESERVOIR MANAGEMENT AND OPERATIONS MODELS - A STATE-OF-THE-ART REVIEW, *Water Resour. Res.*, 21(12), 1797–1818, doi:10.1029/WR021i012p01797.
- Zaefferer, M., T. Bartz-Beielstein, B. Naujoks, T. Wagner, and M. Emmerich (2013), A Case Study on Multi-Criteria Optimization of an Event Detection Software under Limited Budgets, in *Evolutionary Multi-Criterion Optimization*, vol. 7811, edited by R. Purshouse, P. Fleming, C. Fonseca, S. Greco, and J. Shaw, pp. 756–770, Springer Berlin Heidelberg.
- Zhang, L., and V. P. Singh (2007), Gumbel–Hougaard Copula for Trivariate Rainfall Frequency Analysis, *J. Hydrol. Eng.*, 12(4), 409–419, doi:10.1061/(ASCE)1084-0699(2007)12:4(409).
- Zhang, L., V. P. Singh, and F. Asce (2006), Using the Copula Method, *Water*, 11(April), 150–

164.

- Zhang, X., R. Srinivasan, and M. Van Liew (2009), Approximating SWAT Model Using Artificial Neural Network and Support Vector Machine<sup>1</sup>, *JAWRA J. Am. Water Resour. Assoc.*, 45(2), 460–474, doi:10.1111/j.1752-1688.2009.00302.x.
- Zhou, A., B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang (2011), Multiobjective evolutionary algorithms: A survey of the state of the art, *Swarm Evol. Comput.*, 1(1), 32–49, doi:10.1016/j.swevo.2011.03.001.
- Zhou, J., and A. S. Nowak (1988), Integration formulas to evaluate functions of random variables, *Struct. Saf.*, 5(4), 267–284.
- Zitzler, E., and L. Thiele (1999), Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach, *IEEE Trans. Evol. Comput.*, 3(4), 257–271, doi:10.1109/4235.797969.
- Zitzler, E., L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. da Fonseca (2003), Performance assessment of multiobjective optimizers: an analysis and review, *IEEE Trans. Evol. Comput.*, 7(2), 117–132, doi:10.1109/TEVC.2003.810758.
- Zitzler, E., J. Knowles, and L. Thiele (2008), Quality assessment of pareto set approximations, *Berlin, Heidelb. Springer-Verlag*, 373–404, doi:10.1007/978-3-540-88908-3\_14.
- Zou, R., W.-S. Lung, and J. Wu (2007), An adaptive neural network embedded genetic algorithm approach for inverse water quality modeling, *Water Resour. Res.*, 43(8), doi:10.1029/2006wr005158.

# A

## APPENDIX A

---

### A.1 THE UNIVARIATE CYCLOSTATIONARY THOMAS-FIERING MODEL

Herein we present the mathematical background of the univariate cyclostationary Thomas-Fiering (TF) model, also known as the univariate periodic autoregressive model of order 1 (i.e., PAR(1)), with Pearson type-III ( $\mathcal{P}$ III) white noise. Let  $\underline{x}_{s,t}$  be a cyclostationary (i.e., periodic) process with each season denoted by,  $s = 1, \dots, S, 1, \dots, S, 1, \dots$ , and period  $t$ , where  $S$  denotes the total number of seasons (e.g., for a monthly model,  $S = 12$  and  $t$  denotes the year). The process can also be expressed as,  $\underline{x}_{s,n}$  where  $n \in \mathbb{Z}^+$  is the time index. In this form, the season  $s$  is obtained by,  $s = n \bmod(S)$ , while if  $n \bmod(S) = 0$ , then  $s = S$ . Furthermore, the period  $t$  can be obtained by,  $t = 1 + (n - s)/S$ . For convenience, we use the first formulation, also omitting the period index  $t$ . The generating mechanism of the model is:

$$\underline{x}_s = a_s \underline{x}_{s-1} + b_s \underline{\varepsilon}_s \quad (\text{A.1})$$

where  $a_s, b_s$  are seasonally-varying parameters and  $\underline{\varepsilon}_s$  denotes an i.i.d. variate. The parameter  $a_s = \text{Cov}[\underline{x}_s, \underline{x}_{s-1}] / \text{Var}[\underline{x}_{s-1}]$  and  $b_s = \sqrt{\text{Var}[\underline{x}_s] - a_s^2 \text{Var}[\underline{x}_{s-1}]}$ .

The statistical characteristics of the white noise  $\underline{\varepsilon}_s$  term, which is generated through  $\mathcal{P}$ III distribution, are related to those of the target process  $\underline{x}_s$  via the following relationships:

$$\mu_{\underline{\varepsilon}_s} = E[\underline{\varepsilon}_s] = b_s^{-1} \{E[\underline{x}_s] - a_s E[\underline{x}_{s-1}]\} \quad (\text{A.2})$$

$$\sigma_{\underline{\varepsilon}_s}^2 = \text{Var}[\underline{\varepsilon}_s] = 1 \quad (\text{A.3})$$

$$C_{s_{\underline{\varepsilon}_s}} = \mu_3[\underline{\varepsilon}_s] = b_s^{-3} \{\mu_3[\underline{x}_s] - a_s^3 \mu_3[\underline{x}_{s-1}]\} \quad (\text{A.4})$$

where  $\mu_3[\underline{\xi}]$  denotes the third central moment of an arbitrary random variable  $\underline{\xi}$ , which in the case of  $\underline{\varepsilon}_s$  coincides with its skewness coefficient since the model assumes unit variance. Furthermore, following the rationale of Chapter 3, the envelope function of the generation mechanism can be expressed as:

$$x_s \geq a_s x_{s-1} + b_s \ell_{\underline{\varepsilon}_s} \quad (\text{A.5})$$

for positive skewness (i.e.,  $\mathcal{P}$ III with  $b > 0$ ), hence forming a lower boundary, and:

$$x_s \leq a_s x_{s-1} + b_s \nu_{\underline{\varepsilon}_s} \quad (\text{A.6})$$

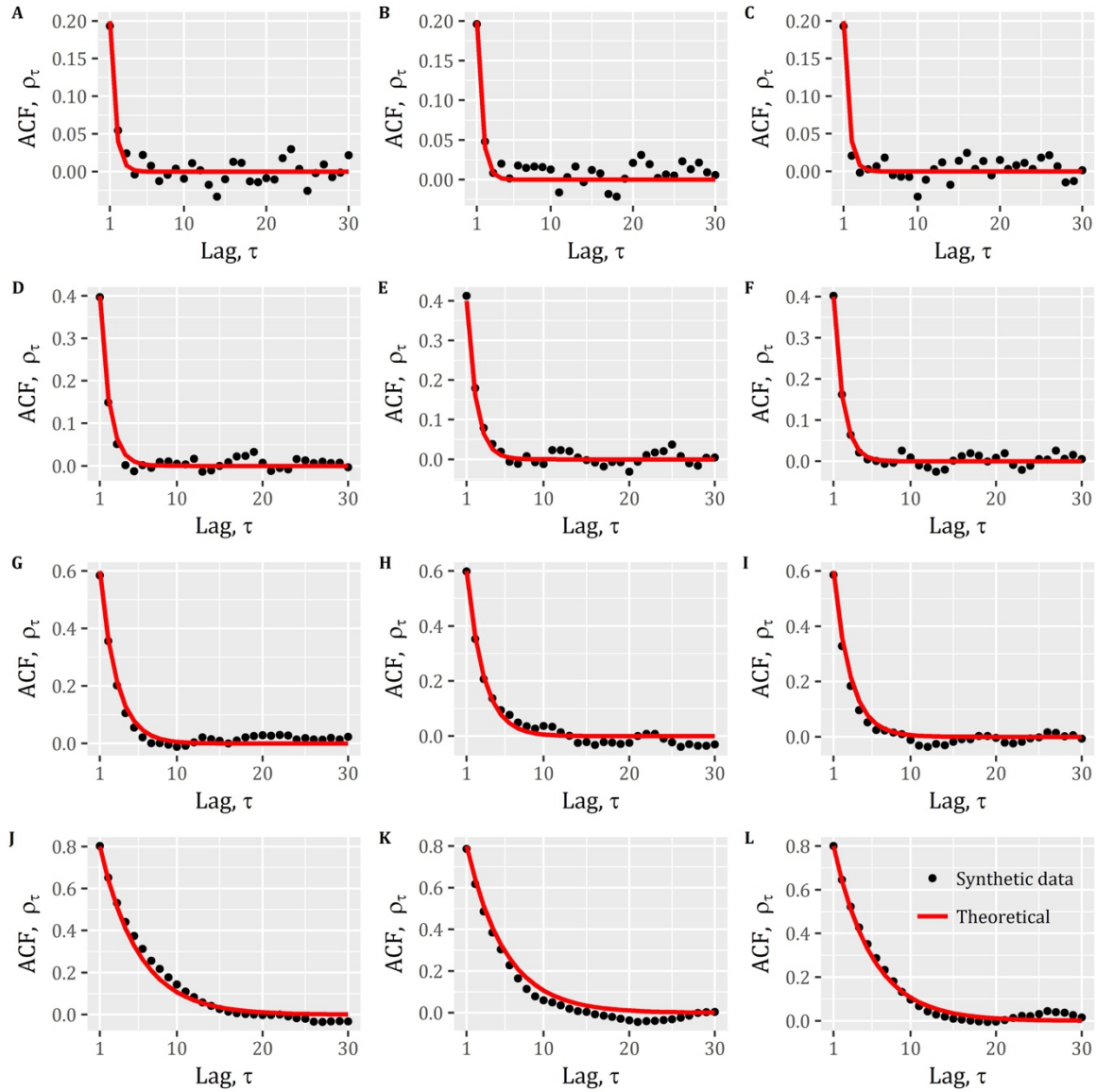
for negative skewness (i.e.,  $\mathcal{P}III$  with  $b < 0$ ), hence forming an upper boundary. In the above,  $\ell_s$  and  $u_s$  respectively denote the lower and upper supports of the distribution of the white noise at season  $s$ . We remark that similar derivations, yet much more complex, can be derived for other models that employ skewed white noise.

## A.2 SUPPLEMENTARY MATERIAL OF CHAPTER 3

**Table A-1** | Scenario-based summary of theoretical (see **Table 3-1** of the main manuscript; section 3.2—“*The envelope behavior in the classical univariate AR(1) model*”) and simulated (synthetically generated; using an AR(1) with  $\mathcal{P}III$  white noise) statistics.

| Scenario   | Type        | Mean ( $\mu$ ) | Variance ( $\sigma^2$ ) | Skewness ( $C_s$ ) | Autocorrelation ( $\rho_1$ ) |
|------------|-------------|----------------|-------------------------|--------------------|------------------------------|
| Scenario A | Theoretical | 0.50           | 1.00                    | 1.00               | 0.20                         |
|            | Simulated   | 0.46           | 0.93                    | 1.05               | 0.20                         |
| Scenario B | Theoretical | 0.50           | 1.00                    | 2.00               | 0.20                         |
|            | Simulated   | 0.54           | 1.06                    | 2.07               | 0.18                         |
| Scenario C | Theoretical | 0.50           | 1.00                    | 4.00               | 0.20                         |
|            | Simulated   | 0.50           | 0.91                    | 3.48               | 0.21                         |
| Scenario D | Theoretical | 0.50           | 1.00                    | 1.00               | 0.40                         |
|            | Simulated   | 0.46           | 0.97                    | 0.91               | 0.34                         |
| Scenario E | Theoretical | 0.50           | 1.00                    | 2.00               | 0.40                         |
|            | Simulated   | 0.49           | 1.11                    | 2.09               | 0.45                         |
| Scenario F | Theoretical | 0.50           | 1.00                    | 4.00               | 0.40                         |
|            | Simulated   | 0.46           | 1.01                    | 4.89               | 0.45                         |
| Scenario G | Theoretical | 0.50           | 1.00                    | 1.00               | 0.60                         |
|            | Simulated   | 0.42           | 0.97                    | 0.88               | 0.64                         |
| Scenario H | Theoretical | 0.50           | 1.00                    | 2.00               | 0.60                         |
|            | Simulated   | 0.48           | 1.04                    | 2.20               | 0.62                         |
| Scenario I | Theoretical | 0.50           | 1.00                    | 4.00               | 0.60                         |
|            | Simulated   | 0.48           | 0.93                    | 4.22               | 0.57                         |
| Scenario J | Theoretical | 0.50           | 1.00                    | 1.00               | 0.80                         |
|            | Simulated   | 0.50           | 1.09                    | 0.75               | 0.82                         |
| Scenario K | Theoretical | 0.50           | 1.00                    | 2.00               | 0.80                         |
|            | Simulated   | 0.45           | 0.97                    | 2.11               | 0.81                         |
| Scenario L | Theoretical | 0.50           | 1.00                    | 4.00               | 0.80                         |
|            | Simulated   | 0.55           | 1.08                    | 4.24               | 0.81                         |





**Figure A.1** | Scenario-based (see **Table 3-1** of the main manuscript; section 3.2—“*The envelope behavior in the classical univariate AR(1) model*”) comparison of synthetic (using the an AR(1) with  $\mathcal{P}$ III white noise) and theoretical autocorrelation function (ACF). The labels of each plot resemble the corresponding scenarios of the aforementioned table (see also **Table A-1**).

**Table A-2** | Summary of theoretical and simulated statistics for the first, zero-autocorrelated, bivariate AR(1) process with  $\mathcal{P}$ III white noise, employed in section 3.2.3—“*From the univariate to the multivariate AR(1) model*” of the main text.

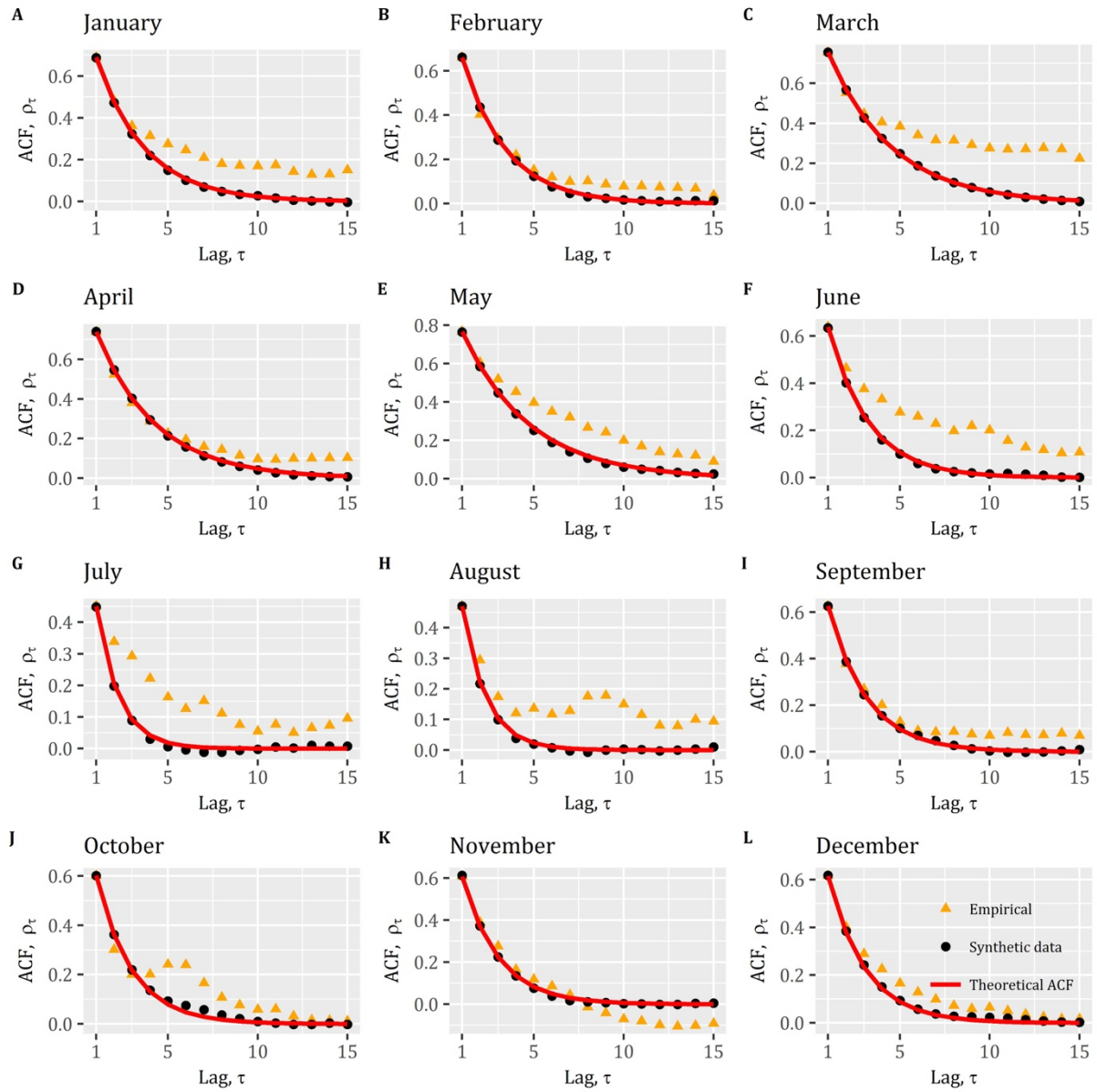
| Process   | Type        | Mean ( $\mu$ ) | Variance ( $\sigma^2$ ) | Skewness ( $C_s$ ) | Autocorrelation ( $\rho_1$ ) |
|---|-------------|----------------|-------------------------|--------------------|------------------------------|
| $x_t^1$   | Theoretical | 0.50           | 1.00                    | 2.00               | 0.00                         |
|   | Simulated   | 0.50           | 1.06                    | 2.39               | 0.00                         |
| $x_t^2$   | Theoretical | 0.50           | 1.00                    | 2.50               | 0.00                         |
|   | Simulated   | 0.51           | 1.14                    | 2.95               | 0.00                         |
| Theoretical cross-correlation ( $\rho_0$ ) = 0.80   Simulated cross-correlation ( $\rho_0$ ) = 0.79 |             |                |                         |                    |                              |

**Table A-3** | Summary of theoretical and simulated statistics for the second, autocorrelated, bivariate AR(1) process with  $\mathcal{P}$ III white noise, employed in section 3.2.3—“From the univariate to the multivariate AR(1) model” of the main text.

| Process   | Type        | Mean ( $\mu$ ) | Variance ( $\sigma^2$ ) | Skewness ( $C_s$ ) | Autocorrelation ( $\rho_1$ ) |
|---|-------------|----------------|-------------------------|--------------------|------------------------------|
| $x_t^1$   | Theoretical | 0.50           | 1.00                    | 2.00               | 0.70                         |
|   | Simulated   | 0.52           | 1.08                    | 2.00               | 0.70                         |
| $x_t^2$   | Theoretical | 0.50           | 1.00                    | 2.50               | 0.50                         |
|   | Simulated   | 0.52           | 1.11                    | 2.51               | 0.51                         |
| Theoretical cross-correlation ( $\rho_0$ ) = 0.80   Simulated cross-correlation ( $\rho_0$ ) = 0.80 |             |                |                         |                    |                              |

**Table A-4** | Monthly-based summary of historical and simulated (synthetically generated using an AR(1) with  $\mathcal{P}$ III white noise) statistics of the real-world case study employed in section 3.3—“Real world case study” of the main text.

| Month     | Type       | Mean ( $\mu$ ) | Variance ( $\sigma^2$ ) | Skewness ( $C_s$ ) | Autocorrelation ( $\rho_1$ ) |
|-----------|------------|----------------|-------------------------|--------------------|------------------------------|
| January   | Historical | 167.89         | 33,973.86               | 3.89               | 0.69                         |
|           | Simulated  | 166.12         | 35,044.58               | 3.92               | 0.70                         |
| February  | Historical | 179.50         | 32,317.25               | 3.95               | 0.66                         |
|           | Simulated  | 177.10         | 32,538.62               | 4.28               | 0.66                         |
| March     | Historical | 172.07         | 13,773.37               | 2.69               | 0.75                         |
|           | Simulated  | 173.37         | 13,608.23               | 2.68               | 0.75                         |
| April     | Historical | 172.47         | 10,253.59               | 4.04               | 0.74                         |
|           | Simulated  | 171.62         | 10,502.08               | 4.28               | 0.74                         |
| May       | Historical | 107.83         | 4055.14                 | 2.29               | 0.77                         |
|           | Simulated  | 110.20         | 4368.32                 | 2.31               | 0.77                         |
| June      | Historical | 50.86          | 591.95                  | 1.59               | 0.64                         |
|           | Simulated  | 51.26          | 604.55                  | 1.58               | 0.63                         |
| July      | Historical | 31.13          | 177.42                  | 2.19               | 0.45                         |
|           | Simulated  | 31.06          | 176.04                  | 2.17               | 0.45                         |
| August    | Historical | 24.00          | 96.04                   | 2.41               | 0.47                         |
|           | Simulated  | 23.96          | 94.83                   | 2.35               | 0.47                         |
| September | Historical | 24.86          | 492.39                  | 5.99               | 0.63                         |
|           | Simulated  | 24.42          | 432.84                  | 5.57               | 0.63                         |
| October   | Historical | 51.77          | 8883.06                 | 6.70               | 0.60                         |
|           | Simulated  | 50.71          | 7905.46                 | 6.26               | 0.60                         |
| November  | Historical | 114.63         | 24,332.88               | 3.49               | 0.61                         |
|           | Simulated  | 111.69         | 23,039.17               | 3.63               | 0.61                         |
| December  | Historical | 197.14         | 68,785.55               | 4.87               | 0.62                         |
|           | Simulated  | 193.85         | 63,948.33               | 4.53               | 0.61                         |



**Figure A.2** | Monthly-based comparison of empirical (historical), synthetic (using AR(1) with  $\mathcal{P}$ III white noise), and theoretical autocorrelation functions (ACFs) of the real-world case study employed in section 3.3—“Real-world case study” of the main text.

## APPENDIX B

---

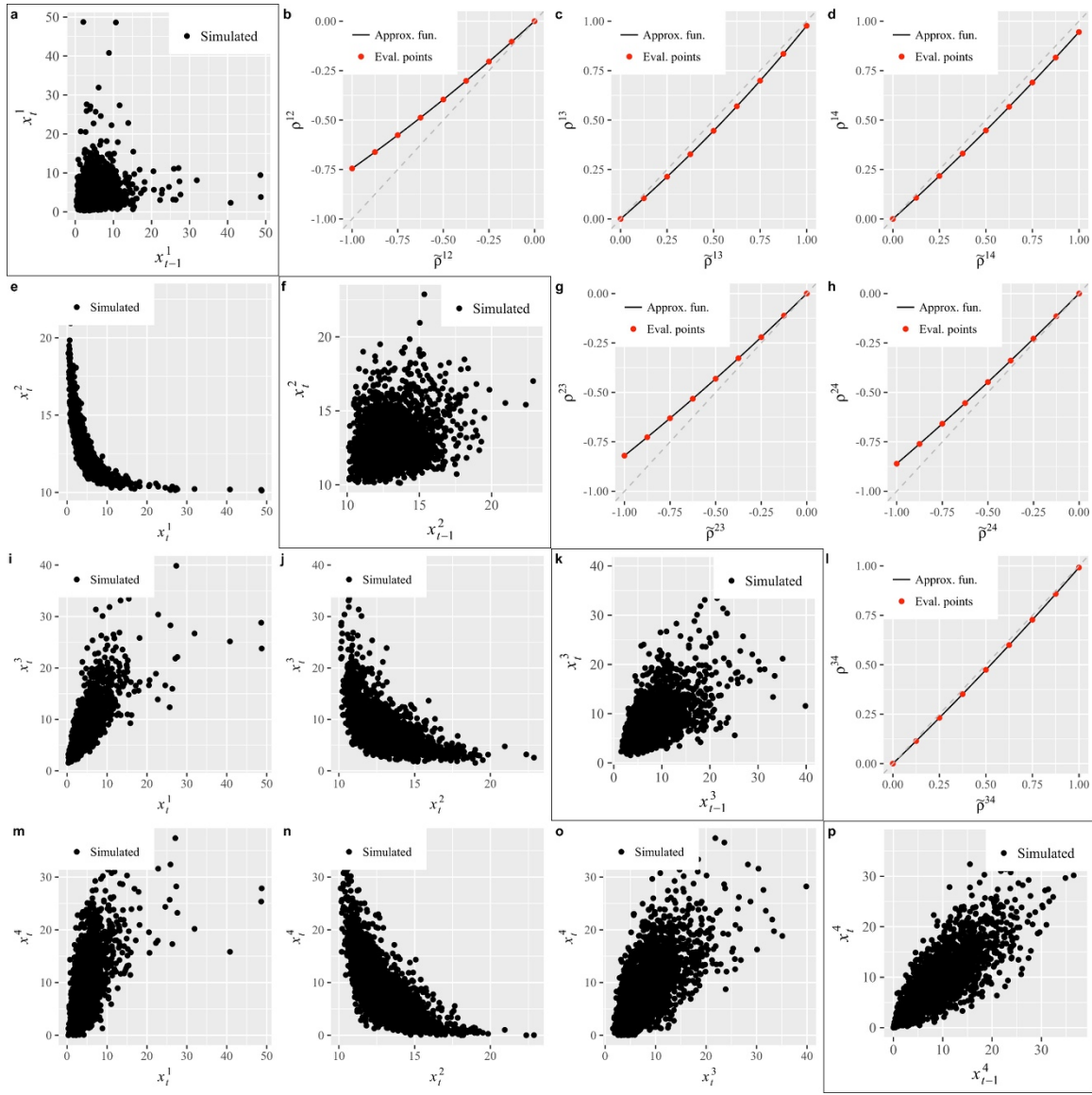
### B.1 SUPPLEMENTARY MATERIAL OF CHAPTER 5

The following figures (**Figure B.1** – **Figure B.6**) complement the simulation studies (hypothetical and real-world) presented in Chapter 5.

**Figure B.1** regards section 5.6.1.2 “*Simulation of multivariate processes*” and illustrates the established dependence patterns (for a randomly selected realization) among the 4 processes (referred to as sites A-D) for time lag 0 (**Figure B.1e**, i, j, m, n, o) and for each process for time lag 1 (**Figure B.1a**, f, k, p). Finally, the upper triangular panels (b, c, d, g, h, l) of **Figure B.1** illustrate the relationship between equivalent,  $\tilde{\rho}^{i,j}$  and target  $\rho^{i,j}$  correlation coefficients among the four processes (i.e., all pairs of sites A-D).

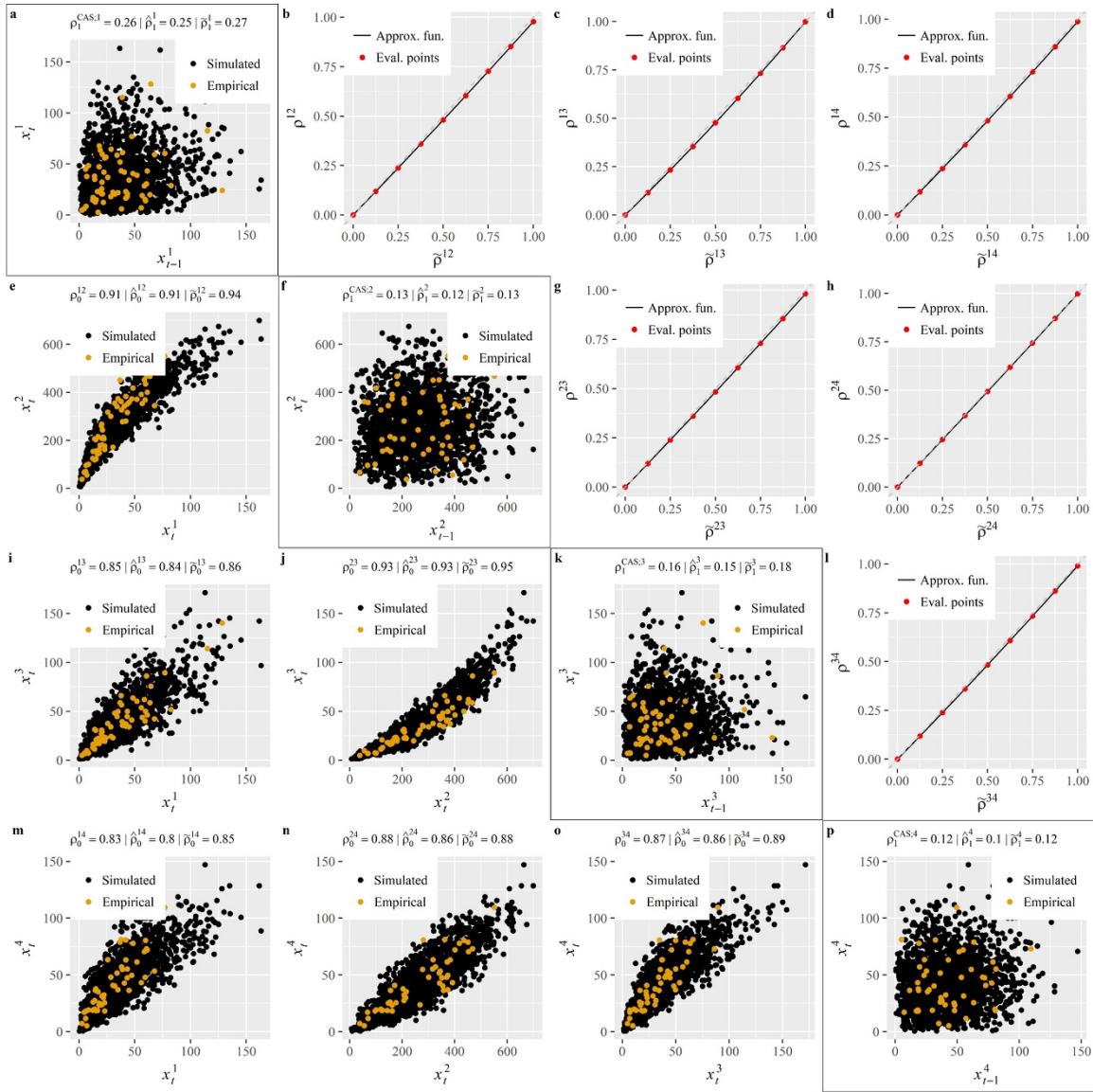
**Figure B.2** regards section 5.7.1 “*Simulation of multivariate annual streamflow processes*” and compares the historical and simulated dependence patterns among the 4 variables for time lag 0 (**Figure B.2e**, i, j, m, n, o) and for each variable for time lag 1 (**Figure B.2a**, f, k, p). This assessment highlights the good agreement between the patterns of observed and synthesized data. Finally, the upper triangular panels (b, c, d, g, h, l) of **Figure B.2** illustrate the relationship between equivalent,  $\tilde{\rho}^{i,j}$  and target  $\rho^{i,j}$  correlation coefficients among the four processes.

**Figure B.3 - Figure B.6** regards section 5.7.2 “*Simulation of univariate daily rainfall process*”, and provide supplementary information on the resemblance of the target marginal distributions and auto-dependence structures.

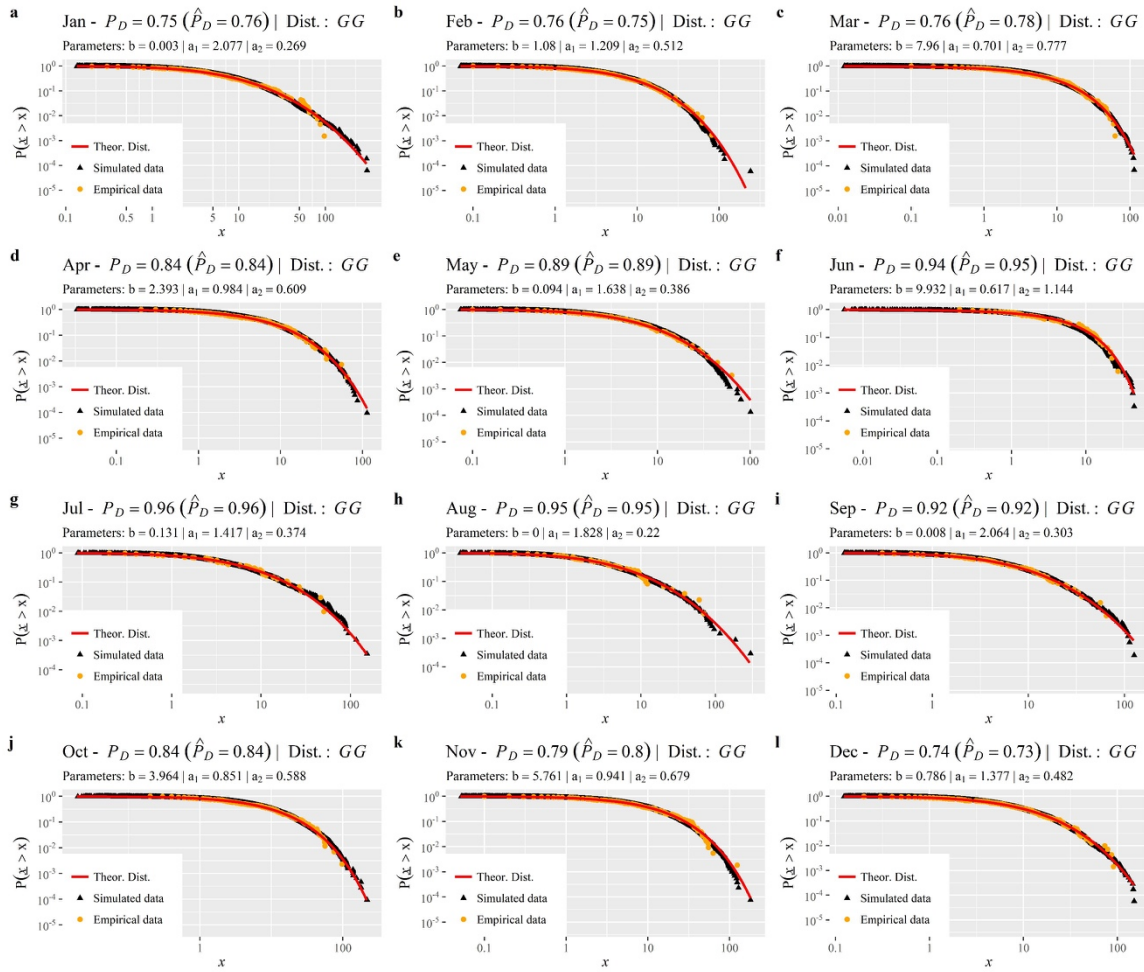


**Figure B.1** | The diagonal panels (a, f, k, p) depict, for a randomly selected realization, the dependence pattern of the synthetically generated data of each process (i.e., for each site) for time lag  $\tau = 1$ . The lower triangular panels (e, i, j, m, n, o) illustrate the dependence pattern of the synthetically generated data among the 4 processes (i.e., for each pair of sites A-D) for time lag  $\tau = 0$ . The upper triangular panels (b, c, d, g, h, l) present the established relationships between equivalent,  $\tilde{\rho}^{i,j}$  and target  $\rho^{i,j}$  correlation coefficients given the corresponding distributions of processes  $\underline{x}_t^i$  and  $\underline{x}_t^j$  (i.e., for each pair of sites A-D).

APPENDIX B

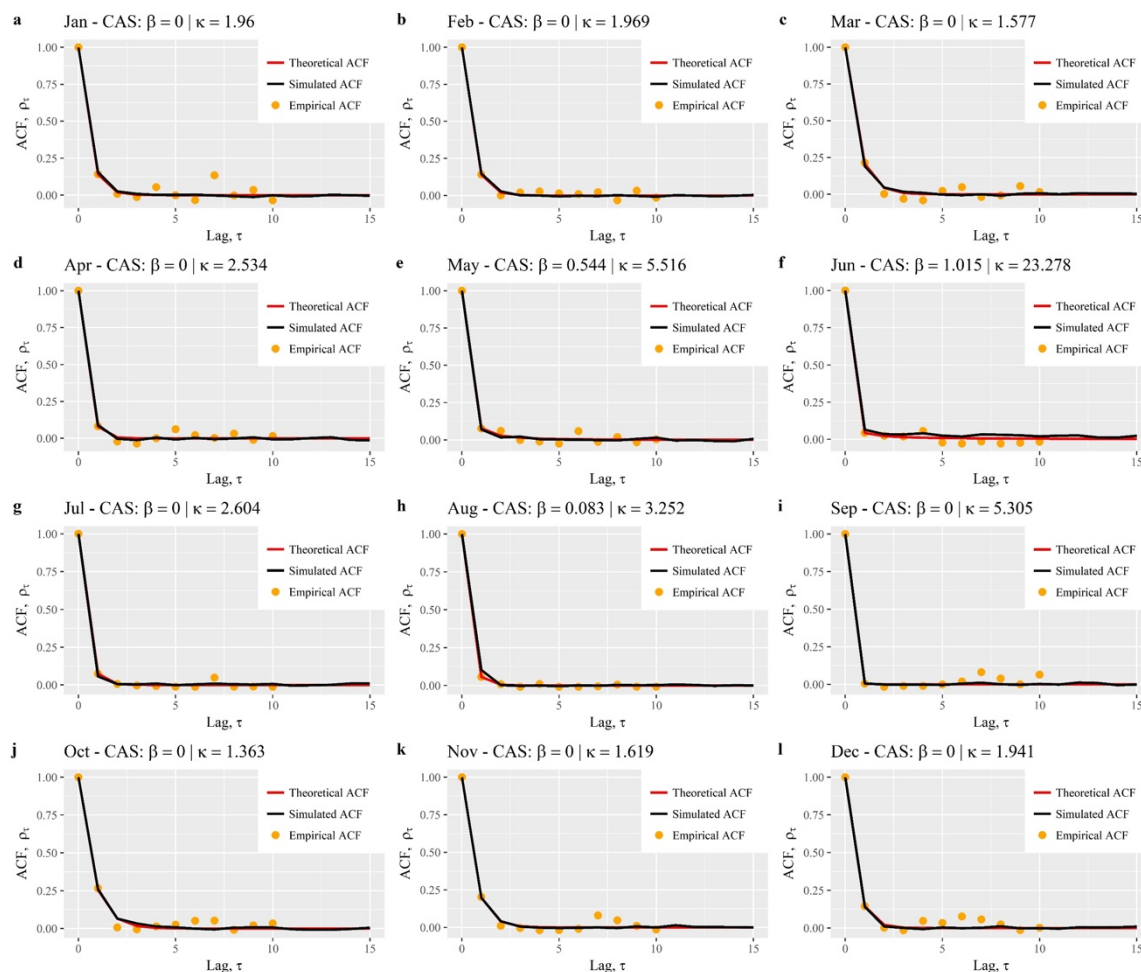


**Figure B.2** | The diagonal panels (a, f, k, p) depict the dependence pattern of the observed and synthetically generated data of each process (i.e., for each station ID1-4) for time lag  $\tau = 1$ . Furthermore, they depict the lag-1, target ( $\rho_1^{CAS;i}$ ), simulated ( $\hat{\rho}_1^i$ ), and equivalent ( $\tilde{\rho}_1^i$ ) autocorrelation coefficients. The lower triangular panels (e, i, j, m, n, o) illustrate the dependence pattern of the observed and synthetically generated data among the processes (i.e., for each pair of stations ID1-4) for time lag  $\tau = 0$ . Furthermore, they depict the lag-0, target ( $\rho_0^{i,j}$ ), simulated ( $\hat{\rho}_0^{i,j}$ ), and equivalent ( $\tilde{\rho}_0^{i,j}$ ) cross-correlation coefficients. The upper triangular panels (b, c, d, g, h, l) present the established relationships between equivalent,  $\tilde{\rho}^{i,j}$  and target  $\rho^{i,j}$  correlation coefficients given the corresponding marginal distributions of processes  $x_t^i$  and  $x_t^j$  (i.e., for each pair of stations ID1-4).



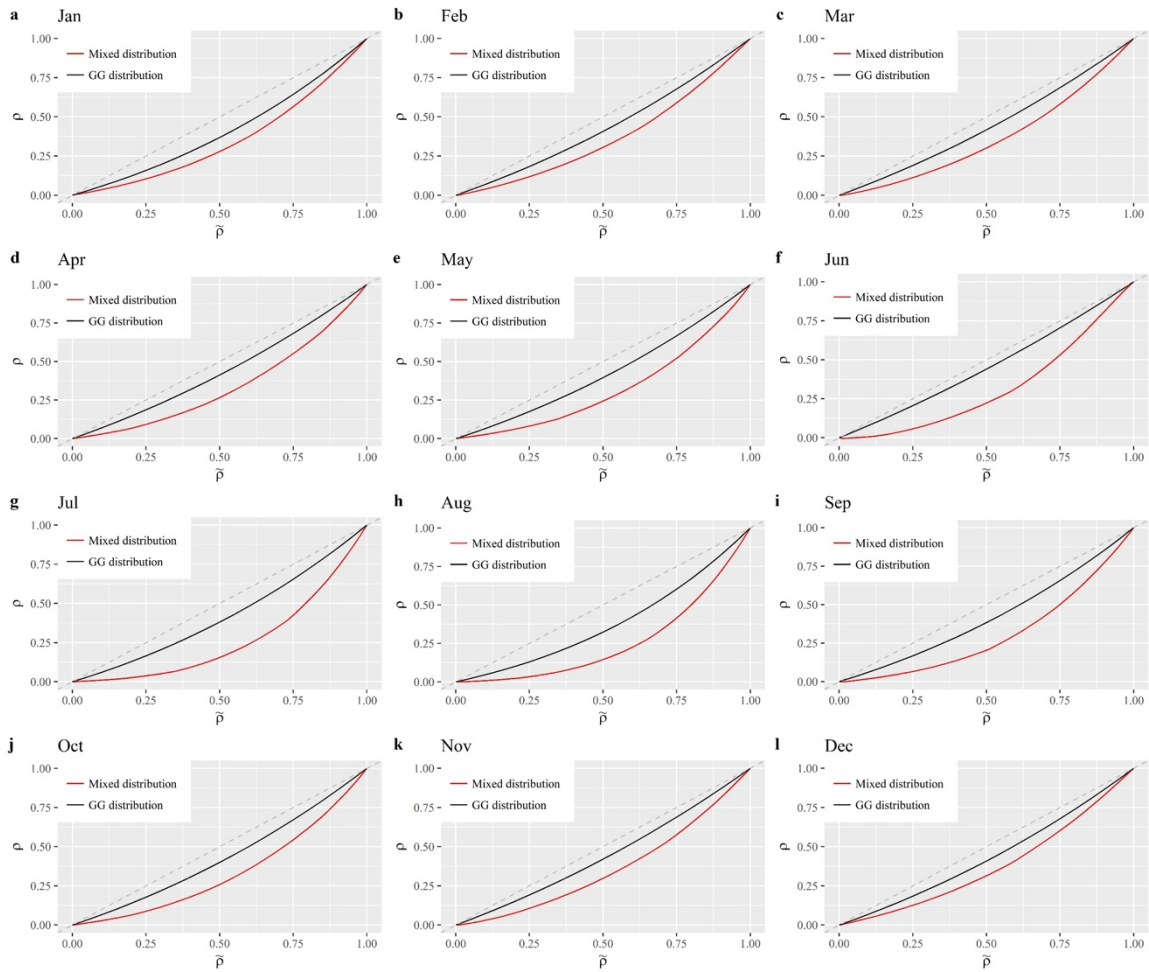
**Figure B.3** | Monthly-based comparison of empirical, simulated and theoretical distribution function of positive daily rainfall at Pavlos station (using the Weibull's plotting position). The title of each plot contains the parameters of the  $GG$  distribution, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.

APPENDIX B



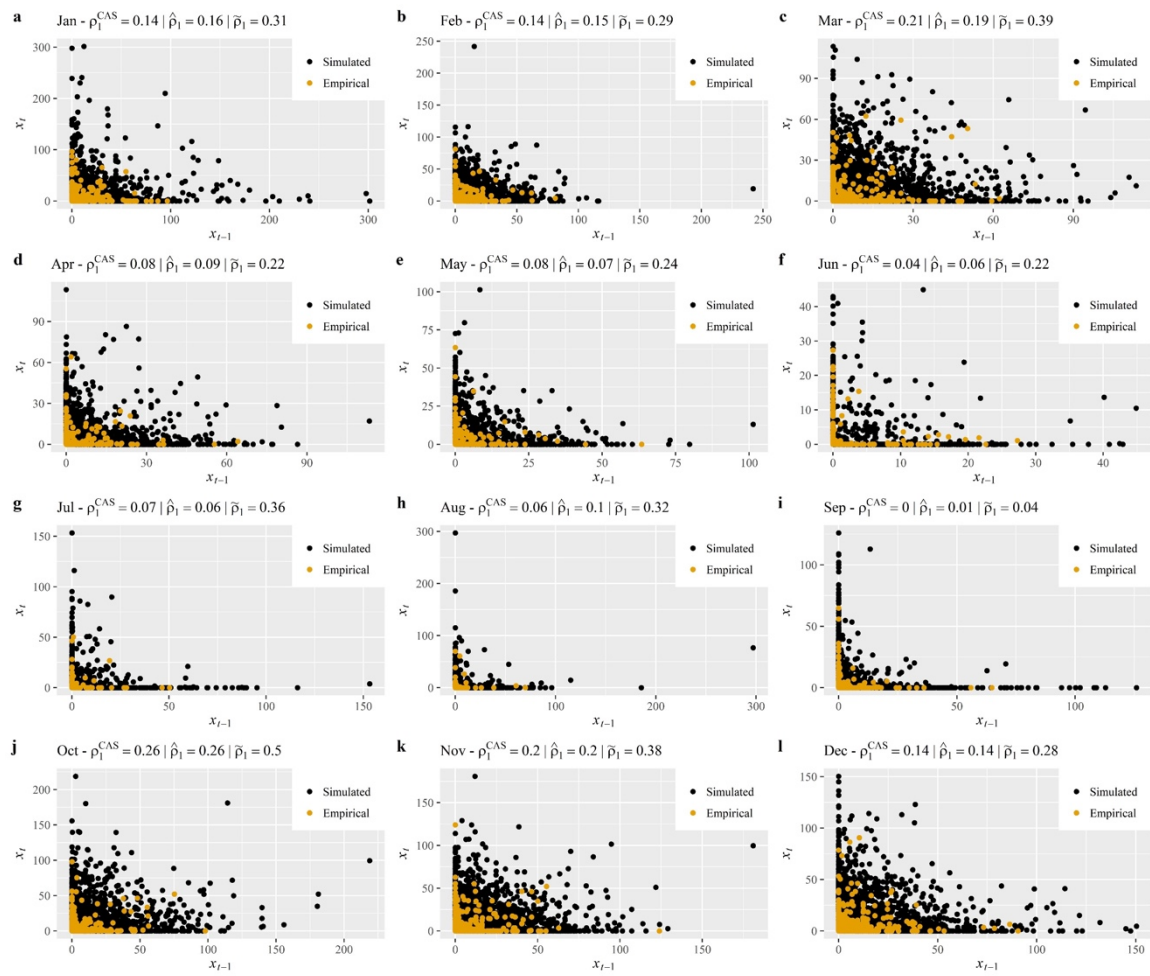
**Figure B.4** | Monthly-based comparison of empirical, simulated and theoretical ACF of daily rainfall at Pavlos station. The title of each plot contains the parameters of the fitted auto-dependence structure (i.e., CAS).





**Figure B.5** | Monthly-based illustration of the relationship between equivalent,  $\tilde{\rho}$  and target  $\rho$  correlation coefficients for the mixed and GG distribution that regard daily rainfall simulation at Pavlos station.

## APPENDIX B



**Figure B.6** | Monthly-based comparison of empirical and simulated dependence pattern for time lag 1. The title of each plot depicts the lag-1, target ( $\rho_1^{CAS}$ ), simulated ( $\hat{\rho}_1$ ), and equivalent ( $\tilde{\rho}_1$ ) autocorrelation coefficients that regard daily rainfall simulation at Pavlos station.

## APPENDIX C

### C.1. THE MULTIVARIATE CONTEMPORANEOUS PAR(1) MODEL

We briefly present the contemporaneous PAR(1) model with Pearson type-III (i.e., 3-parameter Gamma) white noise (referred as PAR-PIII), for multivariate simulation of monthly time series (see [Koutsoyiannis \[1999\]](#)). The model is able to preserve the essential statistics (i.e., mean, variance and skewness coefficient) as well as the lag-1 month-to-month correlations (i.e., autocorrelations) and the lag-0 cross-correlations between *locations*. Particularly, let  $\underline{\mathbf{x}}_{s,t} = [\underline{\mathbf{x}}_{s,t}^1, \dots, \underline{\mathbf{x}}_{s,t}^m]^T$  be a vector of  $m$  stochastically dependent processes at season  $s = 1, \dots, S, 1, \dots, S, \dots$  with period  $t$ . For instance if  $\underline{\mathbf{x}}_{s,t}$  is a cyclostationary monthly process, then  $S = 12$  and  $t$  denotes the year. The process can also be expressed as,  $\underline{\mathbf{x}}_{s,n}$  where  $n \in \mathbb{Z}^>$  is the time index. In this form, the season  $s$  is obtained by,  $s = n \bmod(S)$ , while if  $n \bmod(S) = 0$ , then  $s = S$ . Furthermore, the period  $t$  can be obtained by,  $t = 1 + (n - s)/S$ . For convenience, we use the first formulation, also omitting the period index  $t$ . The models' generating scheme is given by,

$$\underline{\mathbf{x}}_s = \mathbf{A}_s \underline{\mathbf{x}}_{s-1} + \mathbf{B}_s \underline{\mathbf{w}}_s \quad (\text{C.1})$$

where  $\mathbf{A}_s, \mathbf{B}_s$  are seasonally-varying  $m \times m$  parameter matrices and  $\underline{\mathbf{w}}_s = [\underline{w}_s^1, \dots, \underline{w}_s^m]^T$  is a vector of independent random variables generated from Pearson type-III distribution. The matrices  $\mathbf{A}_s$  are calculated as follows:

$$\mathbf{A}_s = \text{diag} \left( \frac{\text{Cov}[\underline{\mathbf{x}}_s^1, \underline{\mathbf{x}}_{s-1}^1]}{\text{Var}[\underline{\mathbf{x}}_{s-1}^1]}, \dots, \frac{\text{Cov}[\underline{\mathbf{x}}_s^m, \underline{\mathbf{x}}_{s-1}^m]}{\text{Var}[\underline{\mathbf{x}}_{s-1}^m]} \right) \quad (\text{C.2})$$

while matrices  $\mathbf{B}_s$  are given by:

$$\mathbf{B}_s \mathbf{B}_s^T = \mathbf{G}_s \quad (\text{C.3})$$

where,

$$\mathbf{G}_s = \text{Cov}[\underline{\mathbf{x}}_s, \underline{\mathbf{x}}_s] - \mathbf{A}_s \text{Cov}[\underline{\mathbf{x}}_{s-1}, \underline{\mathbf{x}}_{s-1}] \mathbf{A}_s^T \quad (\text{C.4})$$

where  $\text{Cov}[\underline{\xi}, \underline{\psi}]$  denotes the covariance of vectors  $\underline{\xi}$  and  $\underline{\psi}$ , i.e.,  $\text{Cov}[\underline{\xi}, \underline{\psi}] = \text{E} \left\{ \left( \underline{\xi} - \text{E}[\underline{\xi}] \right) \left( \underline{\psi}^T - \text{E}[\underline{\psi}^T] \right) \right\}$ . At each season  $s$ , the parameter matrix  $\mathbf{B}_s$  can be estimated either through typical decomposition techniques (e.g., Cholesky or singular value decomposition) or numerically approximated, e.g., through optimization approaches [[Koutsoyiannis, 1999](#); [Higham, 2002](#)].

Regarding the white noise vector  $\underline{\mathbf{w}}_s$ , its statistical structure is associated with the seasonal statistical characteristics of the parent process, through the following equations:

$$E[\underline{\mathbf{w}}_s] = \mathbf{B}_s^{-1} \{E[\underline{\mathbf{x}}_s] - \mathbf{A}_s E[\underline{\mathbf{x}}_{s-1}]\} \tag{C.5}$$

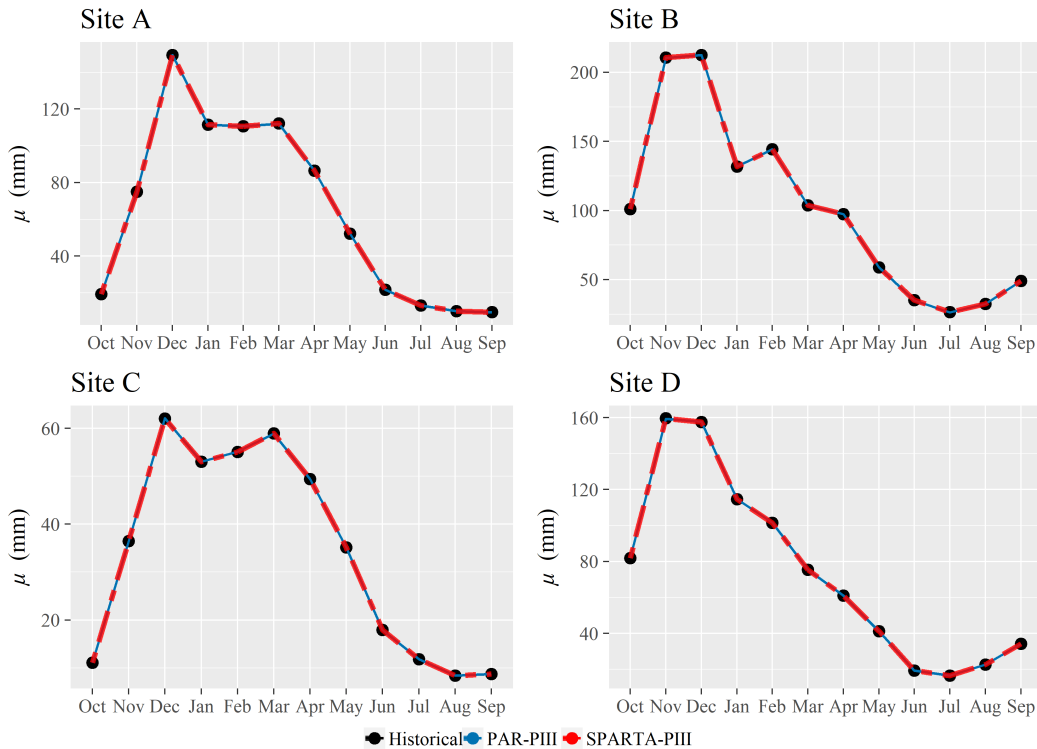
$$\text{Var}[\underline{\mathbf{w}}_s] = [1, \dots, 1]^T \tag{C.6}$$

$$\mu_3[\underline{\mathbf{w}}_s] = (\mathbf{B}_s^{(3)})^{-1} \{\mu_3[\underline{\mathbf{x}}_s] - \mathbf{A}_s^{(3)} \mu_3[\underline{\mathbf{x}}_{s-1}]\} \tag{C.7}$$

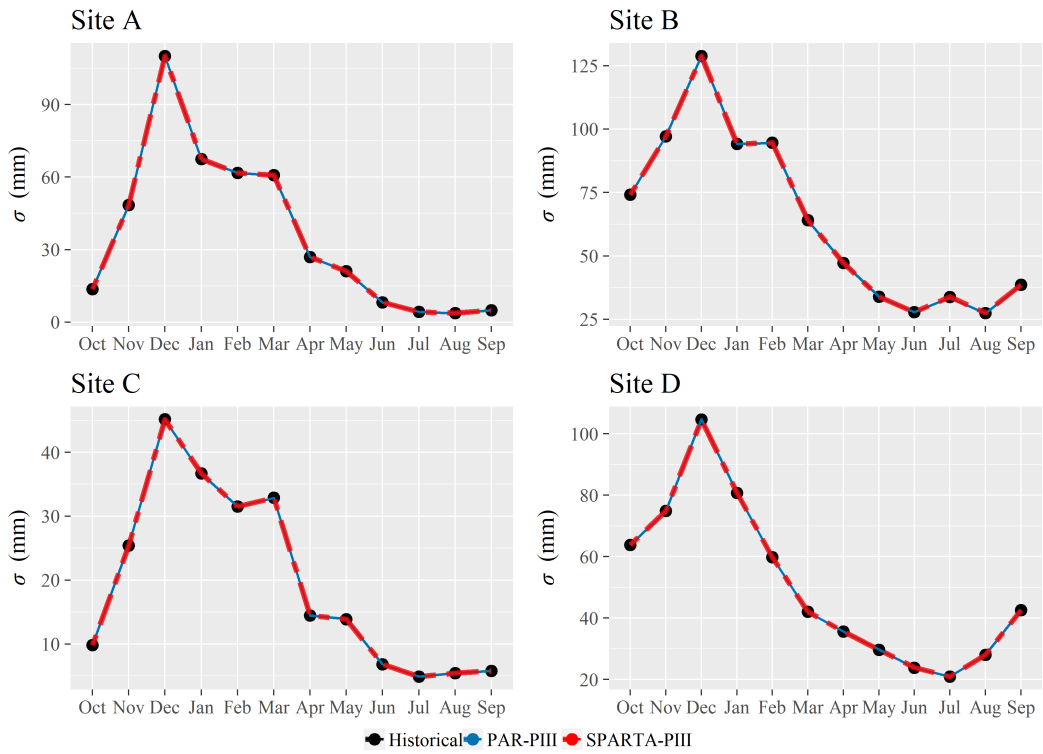
where  $\mathbf{B}_s^{(k)}$  is a matrix whose elements are raised to power  $k$  while  $\mu_3[\underline{\mathbf{w}}_s]$  and  $\mu_3[\underline{\mathbf{x}}_s]$  are vectors that denote the third central moments of  $\underline{\mathbf{w}}_s$  and  $\underline{\mathbf{x}}_s$  respectively. The white noise is produced by a suitable random number generator, in particular the Pearson type-III distribution, which can explicitly preserve  $E[\underline{\mathbf{w}}_s]$ ,  $\text{Var}[\underline{\mathbf{w}}_s]$  and  $\mu_3[\underline{\mathbf{w}}_s]$ .

### C.2. SUPPLEMENTARY MATERIAL OF CHAPTER 6

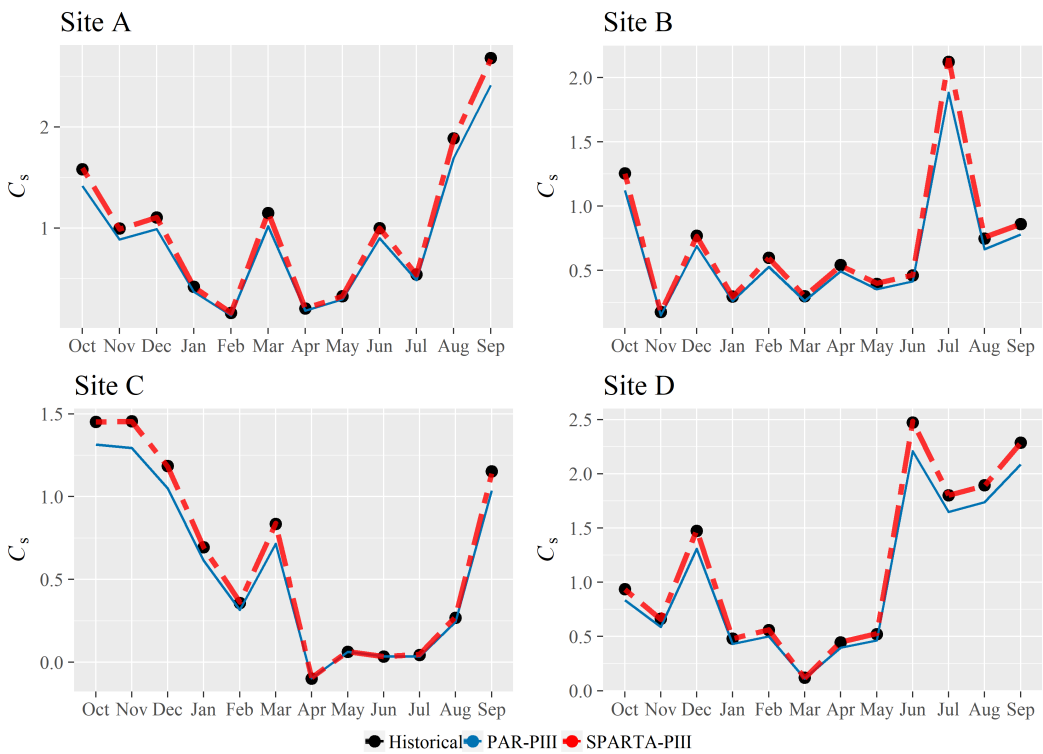
The following figures (**Figure C.1 – Figure C.5**) and tables (**Table C-1 – Table C-2**) illustrate the performance of SPARTA for the case study of section 6.5.3 (multivariate time series simulation) for a long simulation period of 500 000 years. It is noted that in this case, the simulated negative values have not been truncated to zero since we want to validate the theoretical basis of the proposed scheme. The following highlight the solid theoretical background of SPARTA as well as its ability to *exactly* reproduce the desired marginal distributions and the statistics of interest.



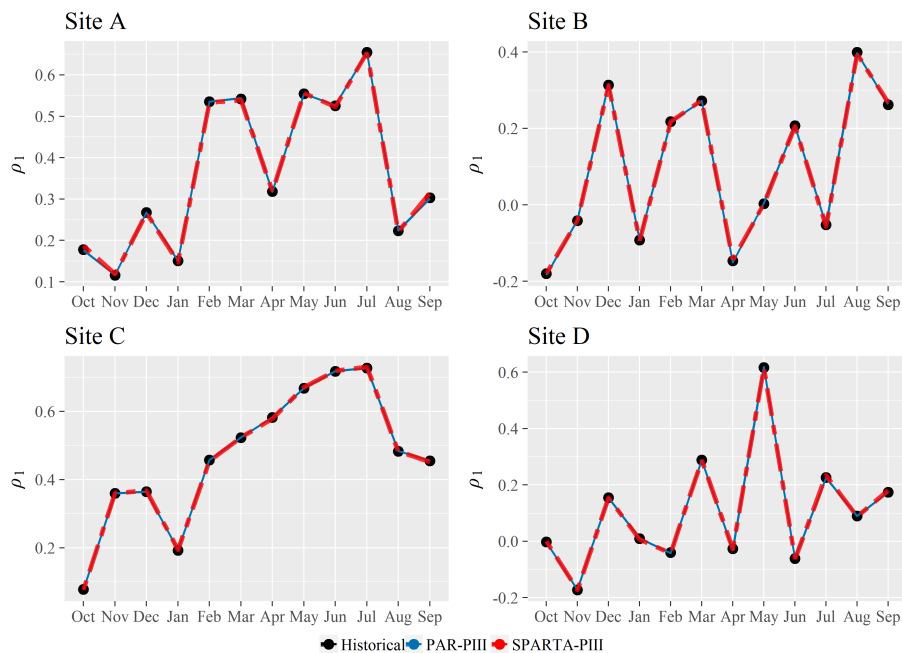
**Figure C.1** | Comparison of monthly mean values,  $\mu$ , of historical and synthetic data (simulation length: 500 000 years).



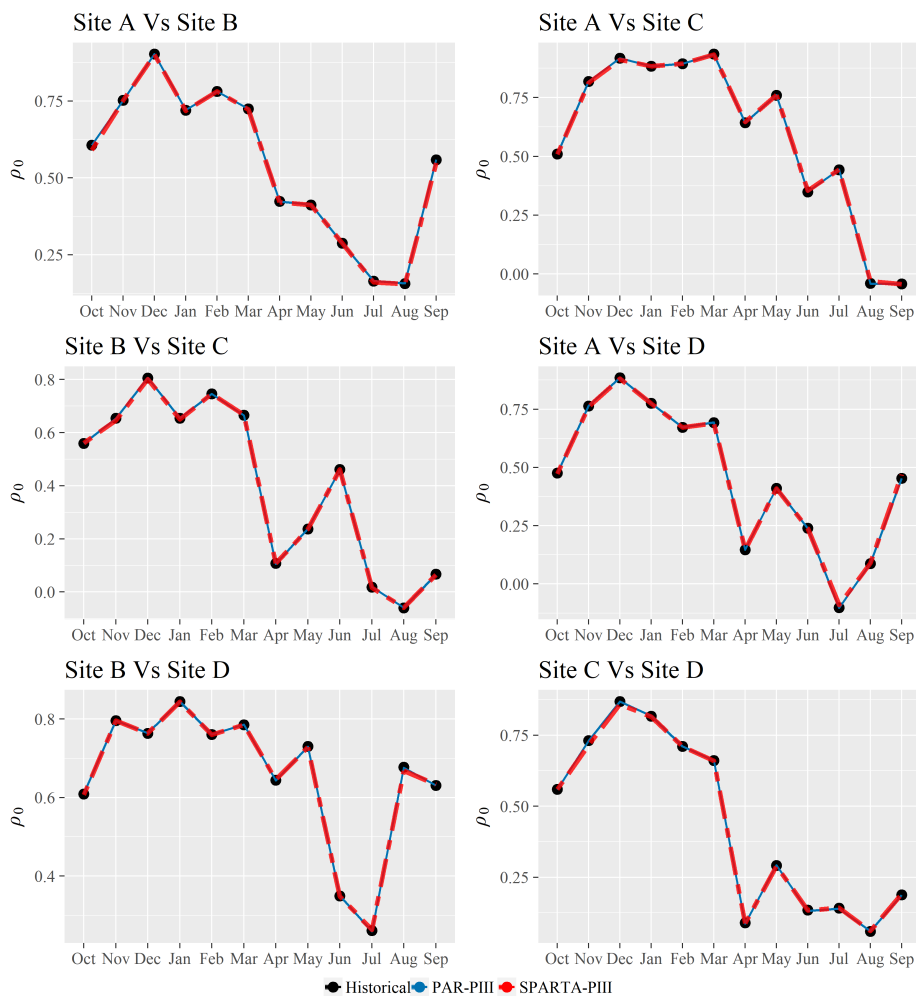
**Figure C.2** | Comparison of monthly standard deviation values,  $\sigma$ , of historical and synthetic data (simulation length: 500 000 years).



**Figure C.3** | Comparison of monthly skewness values,  $C_s$ , of historical and synthetic data (simulation length: 500 000 years).



**Figure C.4** | Comparison of month-to-month lag-1 correlations,  $\rho_1$ , of historical and synthetic data (simulation length: 500 000 years).



**Figure C.5** | Comparison of monthly lag-0 cross-correlations,  $\rho_0$ , between sites of historical and synthetic data (simulation length: 500 000 years).

**Table C-1** | Parameters of PIII for historical and simulated data (from PAR-PIII and SPARTA-PIII); identified with the method of moments.

| Month/ Parameter       | Oct   | Nov     | Dec    | Jan    | Feb    | Mar    | Apr    | May    | June   | July   | Aug   | Sep   |
|------------------------|-------|---------|--------|--------|--------|--------|--------|--------|--------|--------|-------|-------|
| <b>Site A</b>          |       |         |        |        |        |        |        |        |        |        |       |       |
| <i>a</i> (Hist.)       | 1.6   | 4.0     | 3.3    | 22.7   | 155.0  | 3.0    | 95.1   | 37.5   | 4.0    | 13.7   | 1.1   | 0.6   |
| <i>a</i> (SPARTA Sim.) | 1.6   | 4.1     | 3.3    | 23.0   | 150.1  | 3.0    | 94.8   | 37.1   | 4.0    | 13.8   | 1.1   | 0.6   |
| <i>a</i> (PAR Sim.)    | 2.0   | 5.1     | 4.1    | 29.1   | 208.6  | 3.9    | 119.0  | 45.2   | 5.0    | 17.0   | 1.4   | 0.7   |
| <i>b</i> (Hist.)       | 10.8  | 24.1    | 61.0   | 14.1   | 5.0    | 34.9   | 2.8    | 3.4    | 4.1    | 1.1    | 3.5   | 6.6   |
| <i>b</i> (SPARTA Sim.) | 10.9  | 24.0    | 60.8   | 14.1   | 5.0    | 35.2   | 2.8    | 3.5    | 4.1    | 1.1    | 3.5   | 6.6   |
| <i>b</i> (PAR Sim.)    | 9.7   | 21.4    | 54.6   | 12.5   | 4.3    | 30.9   | 2.5    | 3.1    | 3.7    | 1.0    | 3.1   | 5.9   |
| <i>c</i> (Hist.)       | 2.0   | -22.4   | -49.6  | -210.2 | -658.4 | 6.2    | -176.2 | -77.1  | 5.2    | -2.4   | 6.2   | 5.8   |
| <i>c</i> (SPARTA Sim.) | 2.2   | -23.0   | -50.0  | -212.1 | -646.2 | 6.8    | -175.9 | -76.4  | 5.2    | -2.5   | 6.2   | 5.8   |
| <i>c</i> (PAR Sim.)    | 0.1   | -34.3   | -72.7  | -252.3 | -781.6 | -7.4   | -207.5 | -89.9  | 3.4    | -4.1   | 5.7   | 5.4   |
| <b>Site B</b>          |       |         |        |        |        |        |        |        |        |        |       |       |
| <i>a</i> (Hist.)       | 2.5   | 128.6   | 6.7    | 45.5   | 11.1   | 44.1   | 13.5   | 25.5   | 18.7   | 0.9    | 7.2   | 5.4   |
| <i>a</i> (SPARTA Sim.) | 2.5   | 134.3   | 6.8    | 45.6   | 11.3   | 44.7   | 13.8   | 25.0   | 18.5   | 0.9    | 7.0   | 5.4   |
| <i>a</i> (PAR Sim.)    | 3.2   | 174.0   | 8.4    | 57.3   | 14.4   | 57.8   | 16.5   | 32.1   | 23.2   | 1.1    | 9.0   | 6.5   |
| <i>b</i> (Hist.)       | 46.6  | 8.6     | 49.7   | 14.0   | 28.4   | 9.6    | 12.8   | 6.7    | 6.4    | 35.9   | 10.3  | 16.6  |
| <i>b</i> (SPARTA Sim.) | 46.5  | 8.4     | 49.4   | 13.9   | 28.2   | 9.6    | 12.7   | 6.8    | 6.5    | 36.5   | 10.4  | 16.6  |
| <i>b</i> (PAR Sim.)    | 41.6  | 7.4     | 44.4   | 12.4   | 24.9   | 8.4    | 11.7   | 6.0    | 5.8    | 31.6   | 9.1   | 15.1  |
| <i>c</i> (Hist.)       | -17.2 | -891.1  | -121.9 | -503.4 | -171.2 | -321.7 | -76.4  | -112.3 | -85.5  | -5.4   | -41.1 | -40.6 |
| <i>c</i> (SPARTA Sim.) | -17.3 | -914.9  | -123.4 | -504.4 | -173.2 | -324.2 | -78.1  | -110.9 | -84.8  | -5.0   | -40.2 | -40.6 |
| <i>c</i> (PAR Sim.)    | -30.9 | -1071.8 | -161.3 | -581.6 | -214.2 | -382.4 | -94.8  | -133.3 | -99.2  | -9.3   | -50.0 | -49.8 |
| <b>Site C</b>          |       |         |        |        |        |        |        |        |        |        |       |       |
| <i>a</i> (Hist.)       | 1.9   | 1.9     | 2.8    | 8.3    | 31.3   | 5.7    | 404.0  | 1018.5 | 3293.8 | 2140.8 | 55.2  | 3.0   |
| <i>a</i> (SPARTA Sim.) | 1.9   | 1.9     | 2.9    | 8.3    | 30.4   | 5.7    | 459.8  | 942.6  | 3239.7 | 1708.3 | 53.4  | 3.1   |
| <i>a</i> (PAR Sim.)    | 2.3   | 2.4     | 3.6    | 10.6   | 40.3   | 7.8    | 427.9  | 1024.6 | 3179.6 | 3377.9 | 67.1  | 3.7   |
| <i>b</i> (Hist.)       | 7.2   | 18.5    | 26.8   | 12.7   | 5.6    | 13.7   | -0.7   | 0.4    | 0.1    | 0.1    | 0.7   | 3.3   |
| <i>b</i> (SPARTA Sim.) | 7.2   | 18.5    | 26.7   | 12.8   | 5.7    | 13.8   | -0.7   | 0.5    | 0.1    | 0.1    | 0.7   | 3.3   |
| <i>b</i> (PAR Sim.)    | 6.5   | 16.4    | 23.7   | 11.3   | 5.0    | 11.8   | -0.7   | 0.4    | 0.1    | 0.1    | 0.7   | 3.0   |
| <i>c</i> (Hist.)       | -2.5  | 1.5     | -14.1  | -52.5  | -121.1 | -19.8  | 340.2  | -408.4 | -372.9 | -214.3 | -32.2 | -1.3  |
| <i>c</i> (SPARTA Sim.) | -2.5  | 1.5     | -14.3  | -52.4  | -118.6 | -19.5  | 359.6  | -391.6 | -369.6 | -190.2 | -31.6 | -1.4  |
| <i>c</i> (PAR Sim.)    | -3.9  | -2.8    | -23.8  | -66.1  | -144.8 | -33.0  | 349.0  | -410.2 | -366.1 | -272.2 | -36.3 | -2.5  |
| <b>Site D</b>          |       |         |        |        |        |        |        |        |        |        |       |       |
| <i>a</i> (Hist.)       | 4.6   | 9.1     | 1.8    | 17.3   | 12.9   | 276.7  | 20.0   | 14.7   | 0.7    | 1.2    | 1.1   | 0.8   |
| <i>a</i> (SPARTA Sim.) | 4.6   | 9.2     | 1.8    | 17.2   | 12.7   | 299.1  | 19.9   | 14.5   | 0.6    | 1.2    | 1.1   | 0.8   |
| <i>a</i> (PAR Sim.)    | 5.7   | 11.5    | 2.3    | 21.6   | 15.9   | 325.3  | 25.6   | 18.5   | 0.8    | 1.5    | 1.3   | 0.9   |
| <i>b</i> (Hist.)       | 29.9  | 24.8    | 77.2   | 19.4   | 16.7   | 2.5    | 7.9    | 7.8    | 29.5   | 18.9   | 26.6  | 48.8  |
| <i>b</i> (SPARTA Sim.) | 29.8  | 24.6    | 77.6   | 19.5   | 16.8   | 2.4    | 8.0    | 7.8    | 29.7   | 18.9   | 26.5  | 48.7  |
| <i>b</i> (PAR Sim.)    | 26.7  | 22.0    | 68.5   | 17.4   | 15.0   | 2.3    | 7.0    | 6.9    | 26.3   | 17.3   | 24.4  | 44.5  |
| <i>c</i> (Hist.)       | -54.7 | -66.3   | 15.6   | -221.0 | -113.0 | -624.3 | -98.4  | -72.8  | 0.0    | -6.8   | -7.0  | -3.1  |
| <i>c</i> (SPARTA Sim.) | -55.0 | -68.0   | 16.4   | -220.6 | -111.6 | -652.0 | -97.7  | -71.9  | 0.1    | -6.8   | -7.1  | -3.1  |
| <i>c</i> (PAR Sim.)    | -71.1 | -94.6   | -2.4   | -261.1 | -137.1 | -681.8 | -119.4 | -86.9  | -2.3   | -9.0   | -9.7  | -6.8  |

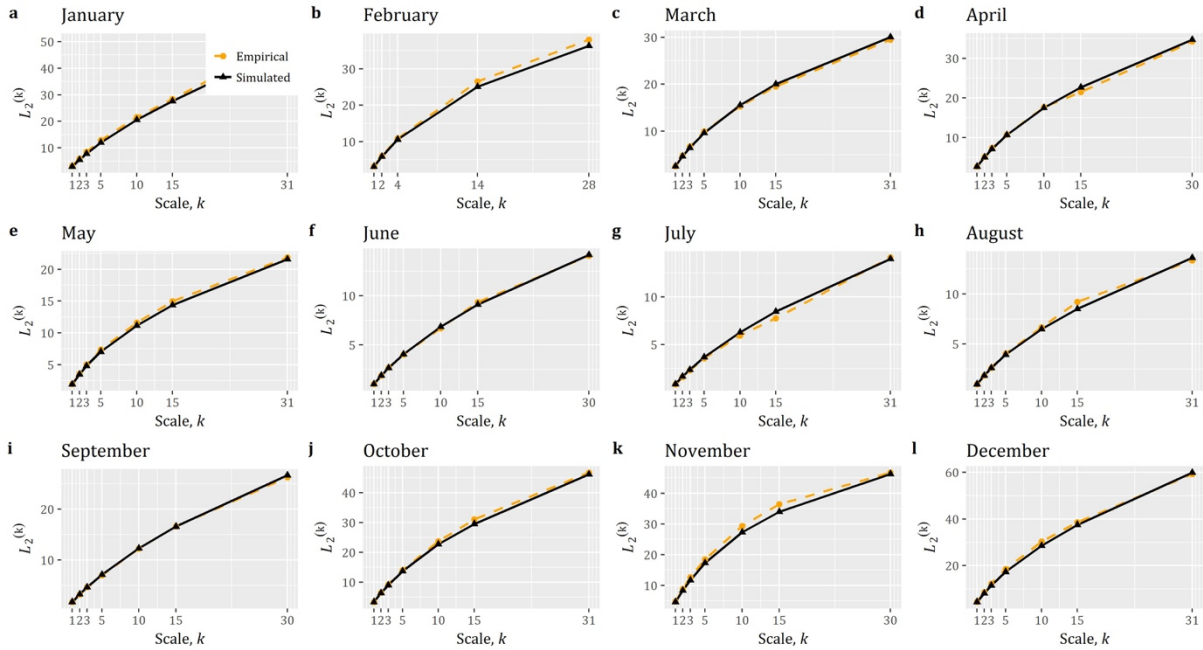
**Table C-2** | Root mean square error between the theoretical values, (i.e., the historical) and the distribution parameters of simulated data of PAR-PIII and SPARTA-PIII models (see **Table C-1**).

| Site/ Parameter        | Site A | Site B | Cite C | Cite D |
|------------------------|--------|--------|--------|--------|
| <i>a</i> (SPARTA Sim.) | 1.42   | 1.66   | 128.73 | 6.46   |
| <i>a</i> (PAR Sim.)    | 17.22  | 14.37  | 358.73 | 14.2   |
| <i>b</i> (SPARTA Sim.) | 0.11   | 0.22   | 0.06   | 0.16   |
| <i>b</i> (PAR Sim.)    | 2.40   | 2.81   | 1.30   | 3.39   |
| <i>c</i> (SPARTA Sim.) | 3.58   | 6.99   | 10.23  | 8.03   |
| <i>c</i> (PAR Sim.)    | 39.75  | 62.70  | 19.43  | 25.09  |

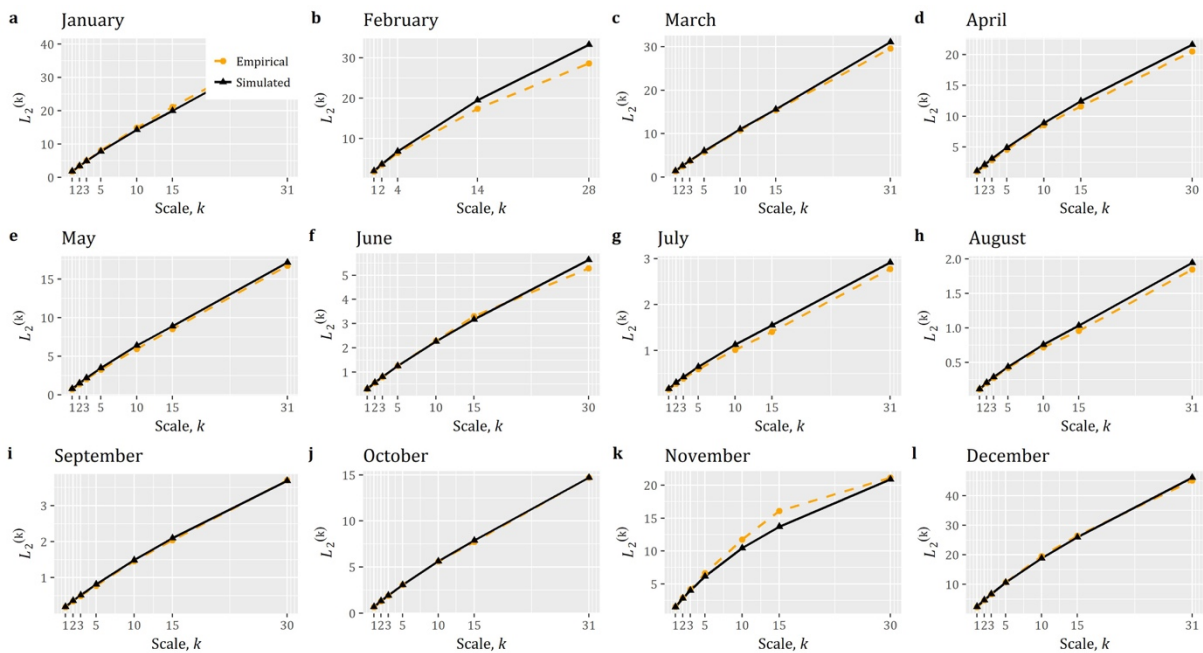


## APPENDIX D

### D.1 SUPPLEMENTARY MATERIAL OF SECTION 7.4

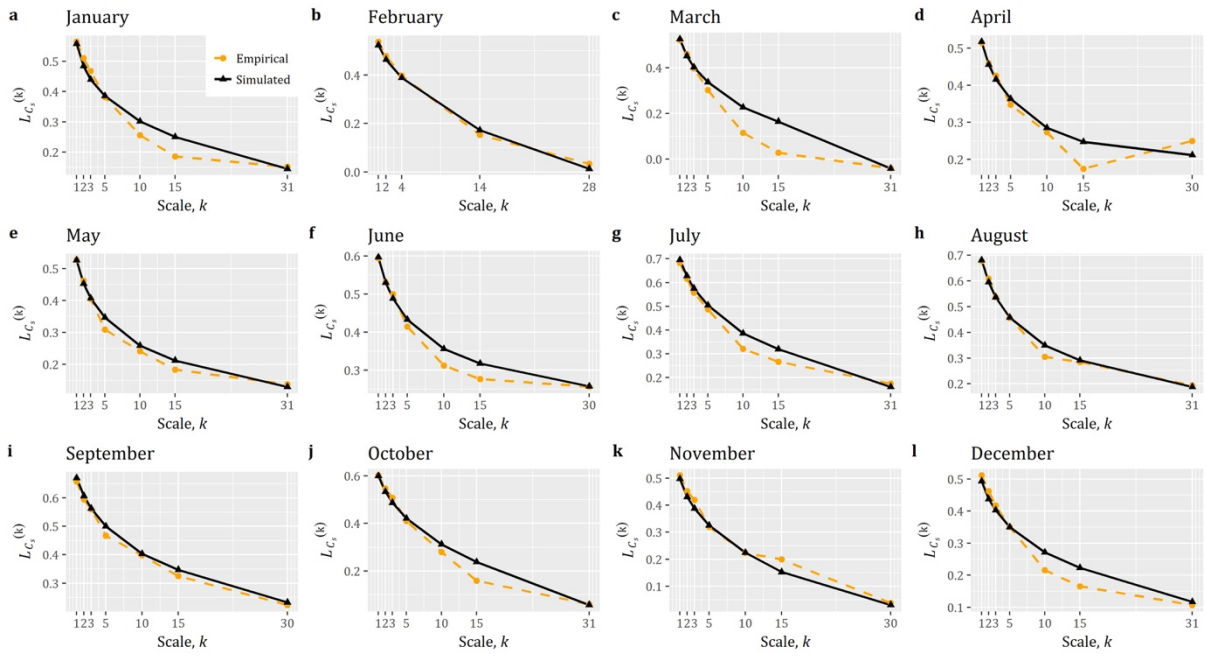


**Figure D.1** | Rainfall - Monthly-based summary of L-scale ( $L_2$ ) as a function of aggregation scale  $k$ .

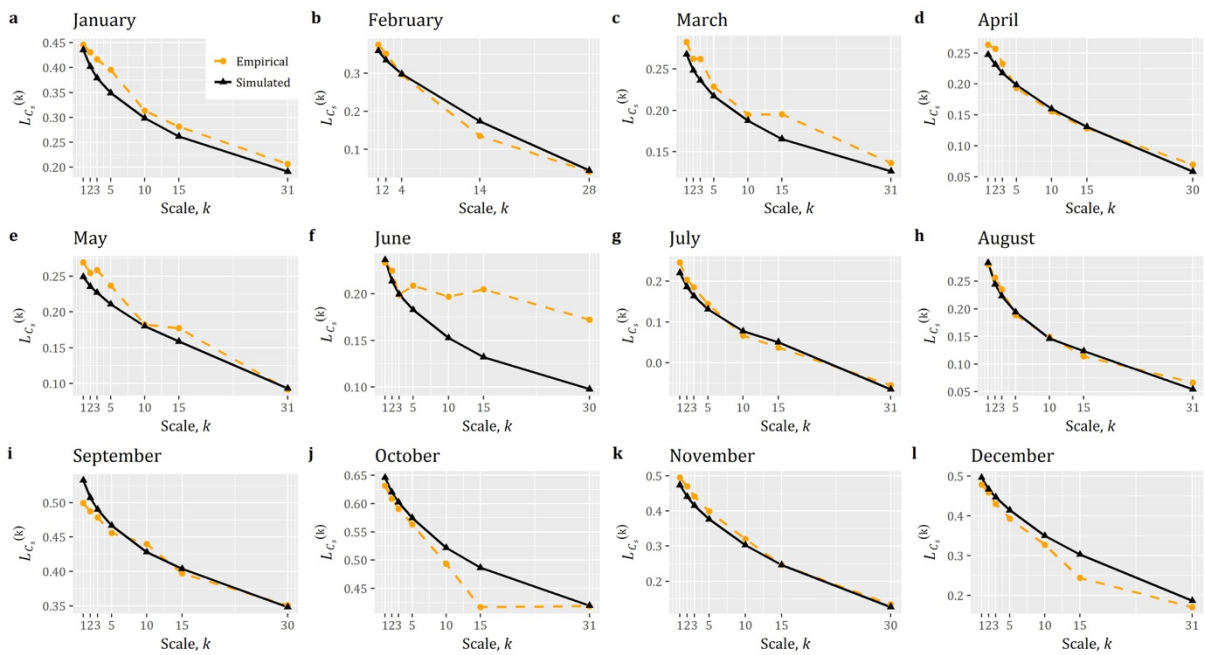


**Figure D.2** | Runoff - Monthly-based summary of L-scale ( $L_2$ ) as a function of aggregation scale  $k$ .

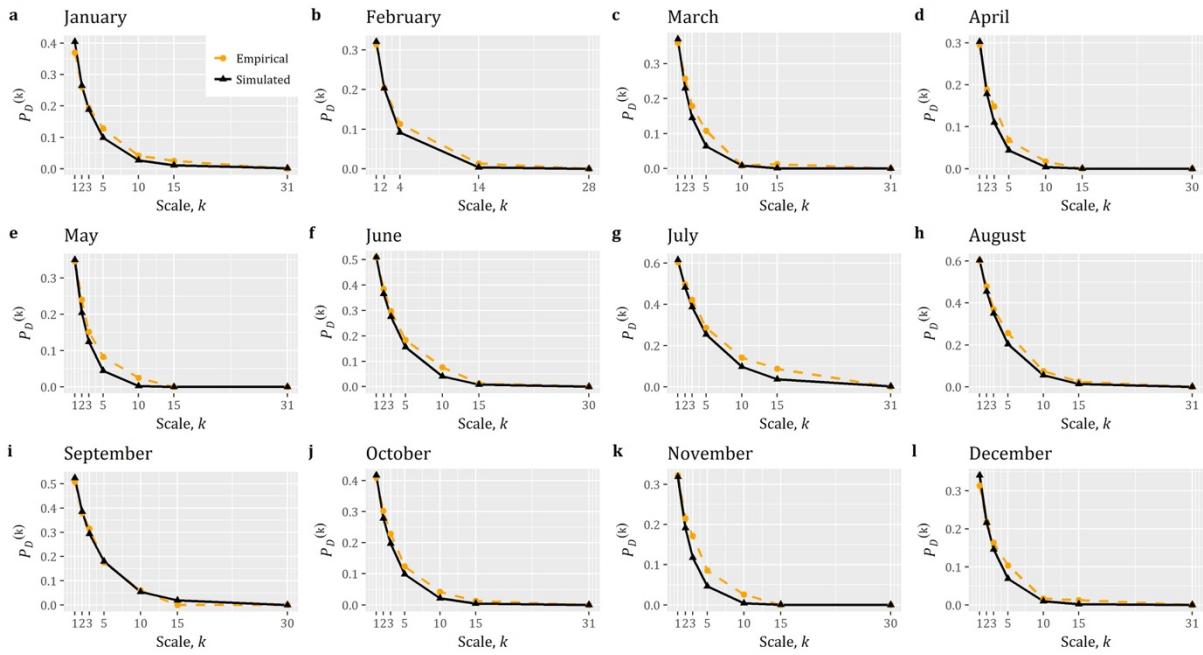
APPENDIX D



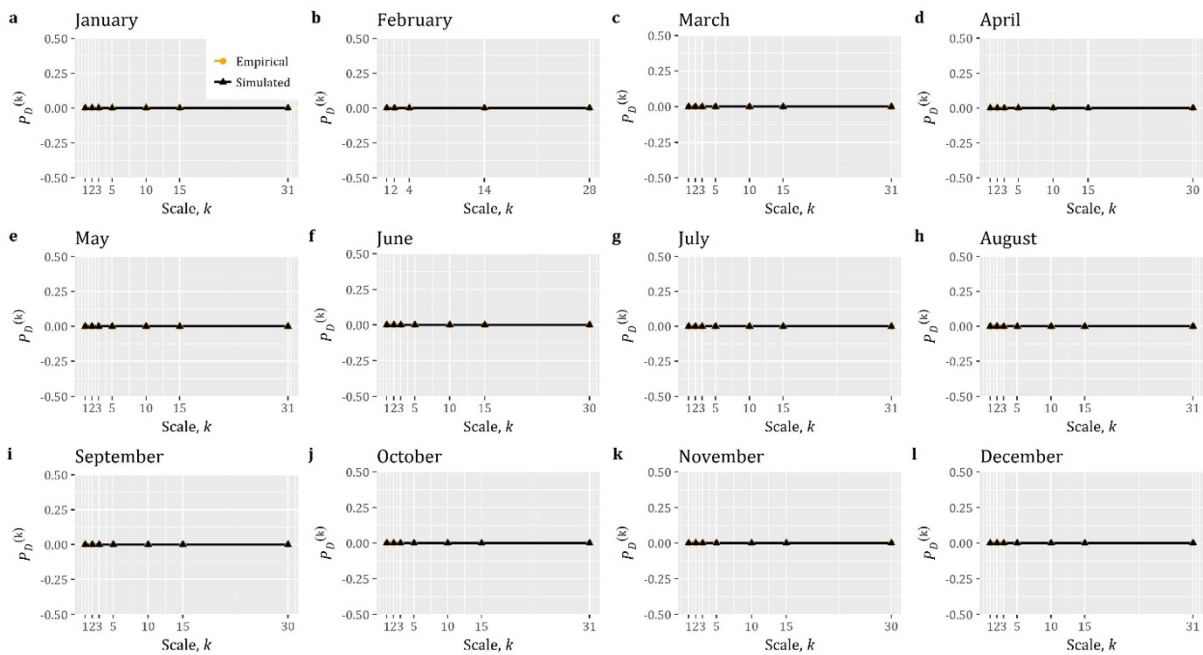
**Figure D.3 |** Rainfall - Monthly-based summary of L-skewness ( $L_{Cs}$ ) as a function of aggregation scale  $k$ .



**Figure D.4 |** Runoff - Monthly-based summary of L-skewness ( $L_{Cs}$ ) as a function of aggregation scale  $k$ .



**Figure D.5** | Rainfall - Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$ .



**Figure D.6** | Runoff - Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$ .

## D.2 MULTI-TEMPORAL SIMULATION OF MULTIVARIATE DAILY RAINFALL PROCESSES

To further explore the capabilities of the NDA-based three-level configuration scheme of Chapter 7, we employ it for the synthesis of long daily rainfall time series (2 000 years) at four locations. More specifically, the historical data<sup>13</sup> concern four rain gauges located at Boeotikos Kephisos river basin, Eastern Greece. The historical data, that span from 1/1/1964 to 31/12/2006, were obtained from the rainfall stations of Pavlos, Atalanti, Leivadia and Tithorea, which hereafter are referred to as site A, B, C and D respectively. See also, [Efstratiadis et al. \[2014a\]](#) for further details regarding the dataset. In this case, in order to account for the intermittent character of rainfall, at monthly and daily time scale, we employ the mixed, zero-inflated, distribution model discussed in section 4.4. Its CDF reads,

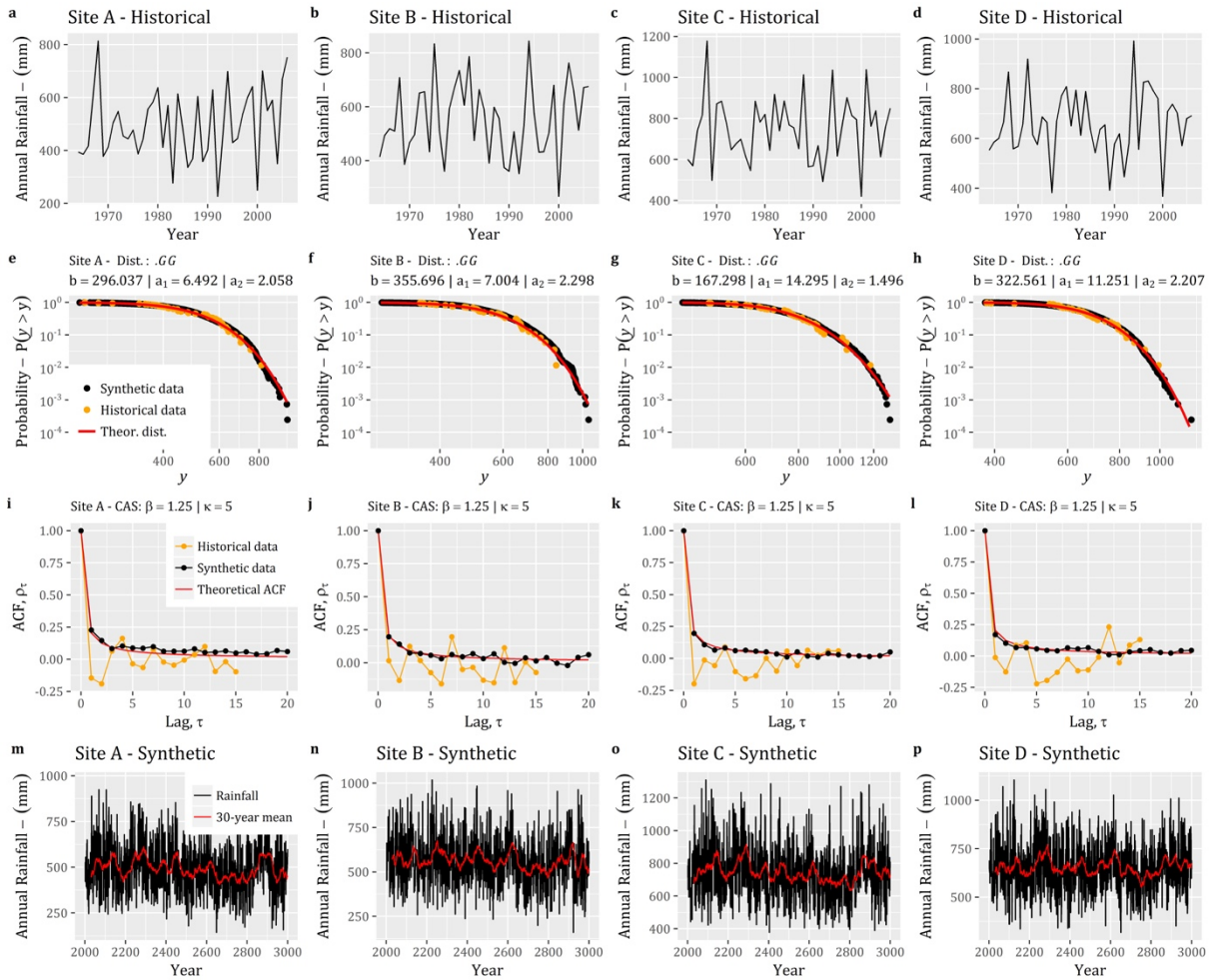
$$F_{\underline{x}}(x) = \begin{cases} p_D, & x \leq 0 \\ p_D + (1 - p_D)G_{\underline{x}}(x), & x > 0 \end{cases} \quad (\text{D.1})$$

where,  $p_D$  denotes the probability of a dry interval (abbreviated as probability dry), i.e.,  $p_D := P(\underline{x} \leq x_D)$  and  $G_{\underline{x}}$  stands for the distribution of amounts greater than the threshold  $x_D$ , i.e.,  $G_{\underline{x}} := F_{\underline{x}|\underline{x} > x_D} = P(\underline{x} \leq x | \underline{x} > x_D)$ . Herein it was assumed that  $x_D := 0$ , while  $G_{\underline{x}}$  was obtained by fitting (using the L-moments method) both the Generalized Gamma ( $\mathcal{GG}$ ; Eq. (5.44)) and the Burr type-XII ( $\mathcal{BrXII}$ ; Eq. (5.41)) distributions and selecting the one that better describes that data at hand.

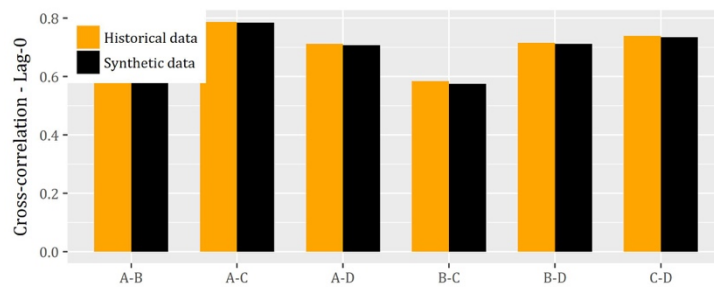
Similar to the case study of section 7.4, we shall start the assessment of the model and presentation of the results from the annual time scale and subsequently move to the monthly and daily ones. **Figure D.7** presents a summary of the simulation results at the annual time scale and verifies that the model was capable of preserving the target distribution functions and autocorrelation structure (i.e., CAS). It is noted that for demonstration purposes the parameters of CAS at the annual scale were manually set to  $\beta = 0$  and  $\kappa = 1.5$  for all processes. **Figure D.8** depicts the lag-0 cross-correlations among the four sites, which are all very well preserved by the model.

---

<sup>13</sup> <http://main.hydroscope.gr/>



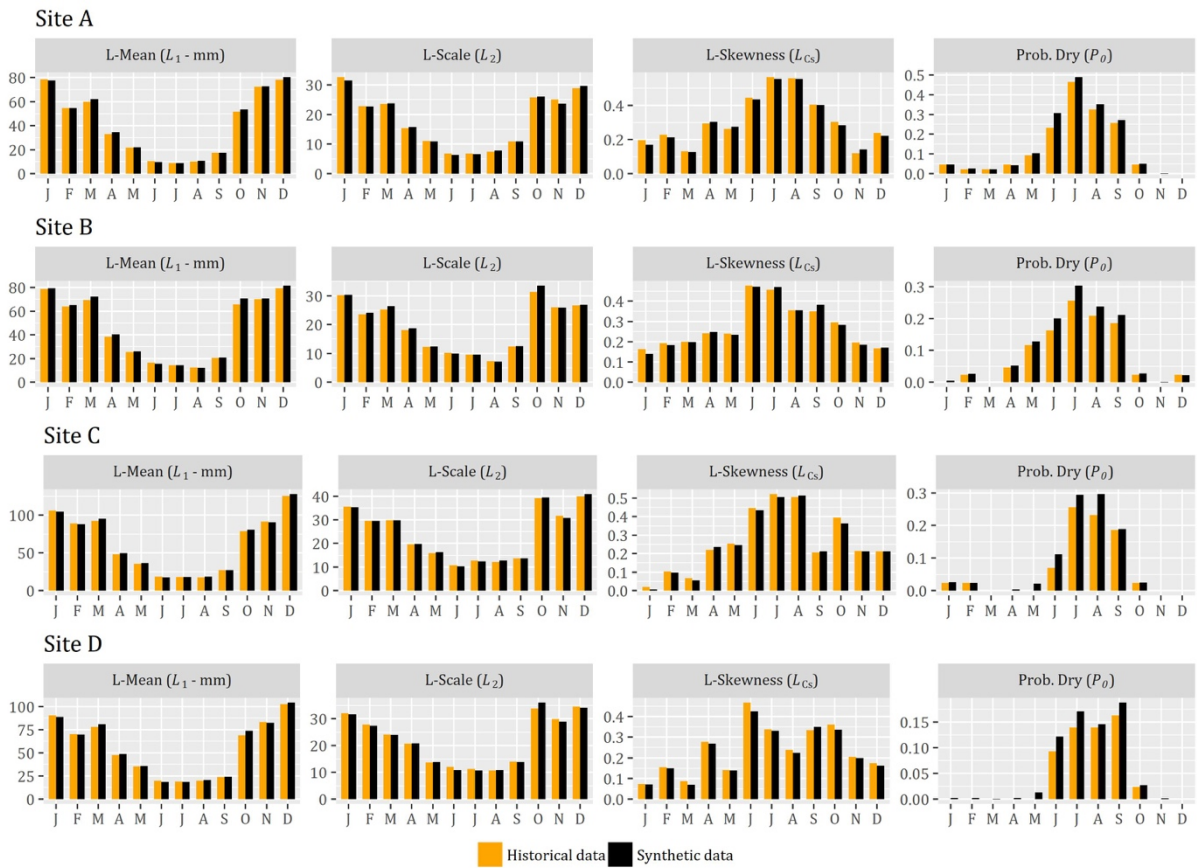
**Figure D.7** | (a-d) Historical annual time series for sites A-D. (e-h) Empirical, simulated and theoretical distribution functions for sites A-D (using the Weibull’s plotting position) (i-l) Empirical, simulated and theoretical ACFs for sites A-D. (m-p) Synthetic annual time series (randomly selected window of 1 000 years).



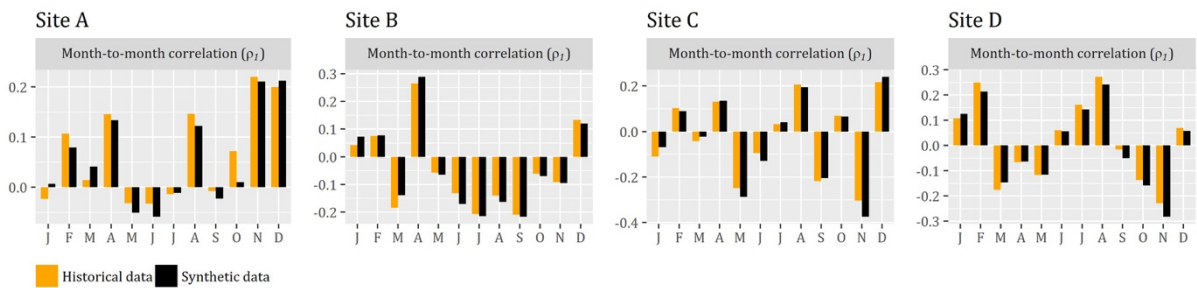
**Figure D.8** | Comparison of historical and simulated lag-0 cross-correlations at the annual time scale.

Moving to the monthly time scale, **Figure D.9** provides a brief summary of the simulation results and highlights the ability of the three-level configuration to resemble the first three L-moments (i.e., L-mean, L-scale and L-Skewness), as well as reproduce the moderate intermittent behavior of the rainfall data, mostly observed during the summer months. Furthermore, **Figure D.10** and **Figure D.11** compare the historical and simulated lag-1 month-to-month correlations and the lag-0 cross-correlations of the monthly time scale respectively. Inspection of these graphs reveals that model closely resembles the target correlations in all cases.

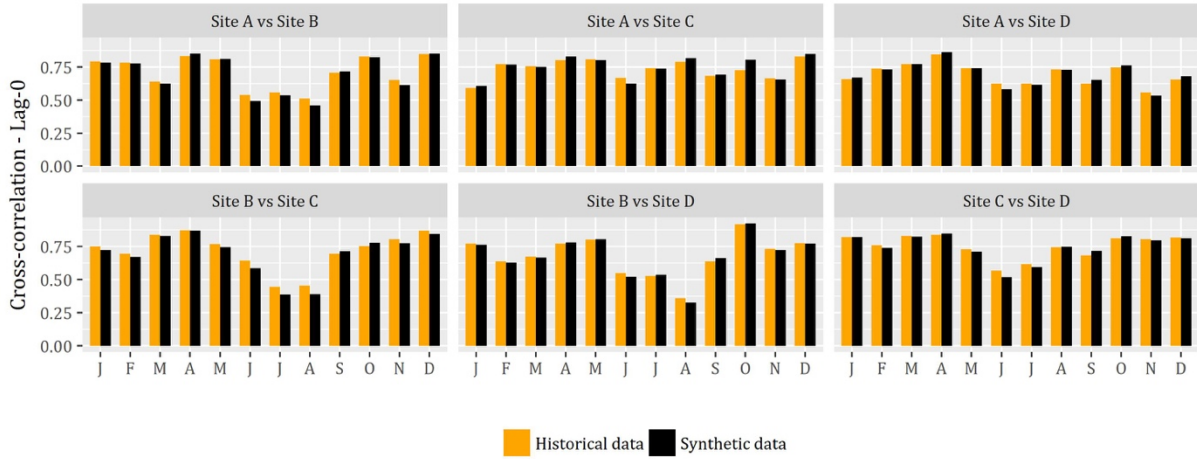
APPENDIX D



**Figure D.9** | Monthly-based comparison of monthly empirical and simulated L-Mean, L-Scale and L-Skewness, as well as probability dry.

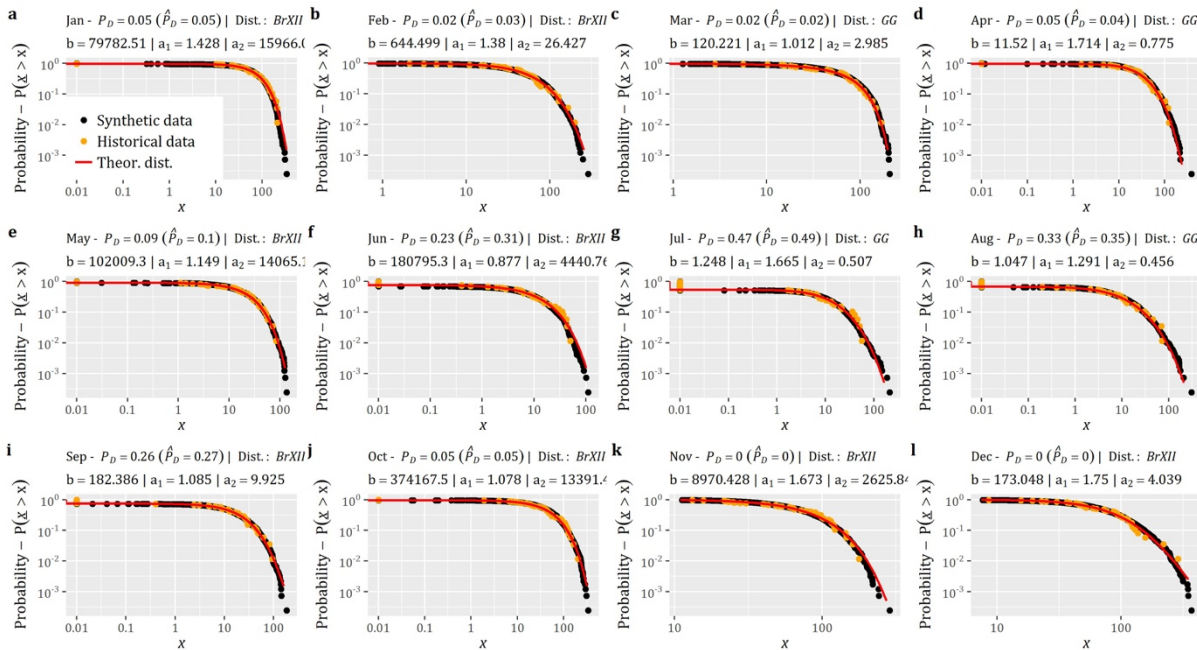


**Figure D.10** | Comparison of historical and simulated lag-1 month-to-month correlations for sites A-D.



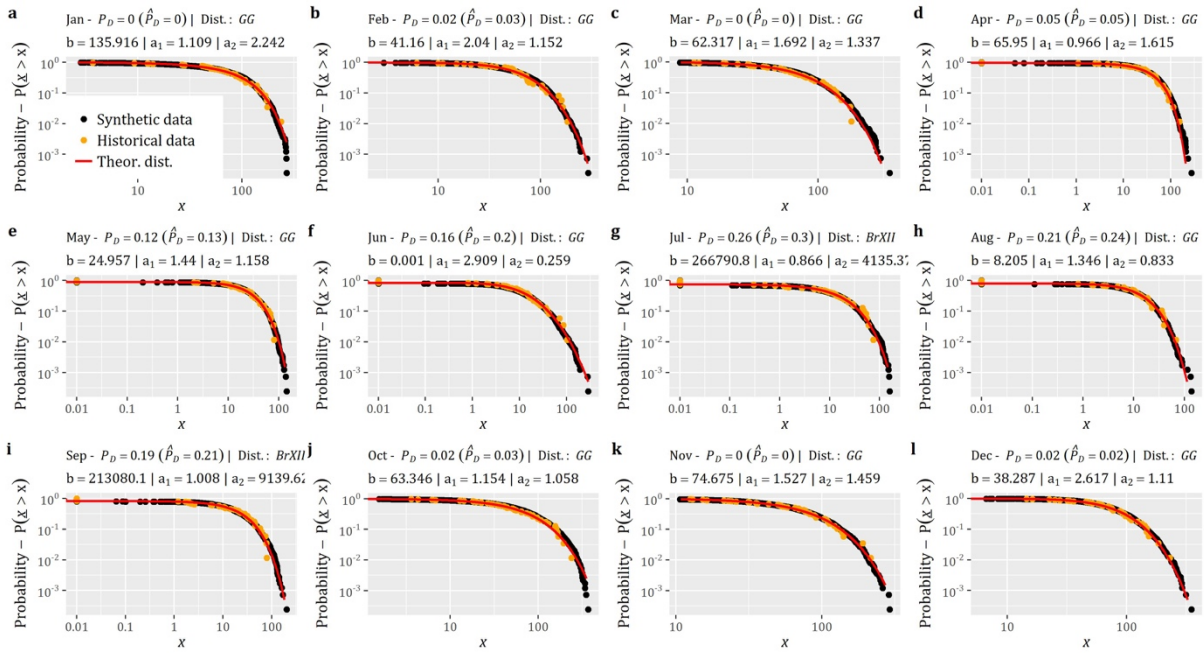
**Figure D.11** | Comparison of monthly historical and simulated lag-0 cross-correlations for sites A-D.

**Figure D.12** to **Figure D.15** provide a comparison among the monthly empirical, simulated and theoretical distribution functions (as well as their parameters) of all months for sites A-D. These figures highlight the ability of the model to preserve the target distribution functions (*GG* or *BrXII*) of the monthly time scale with notable accuracy.

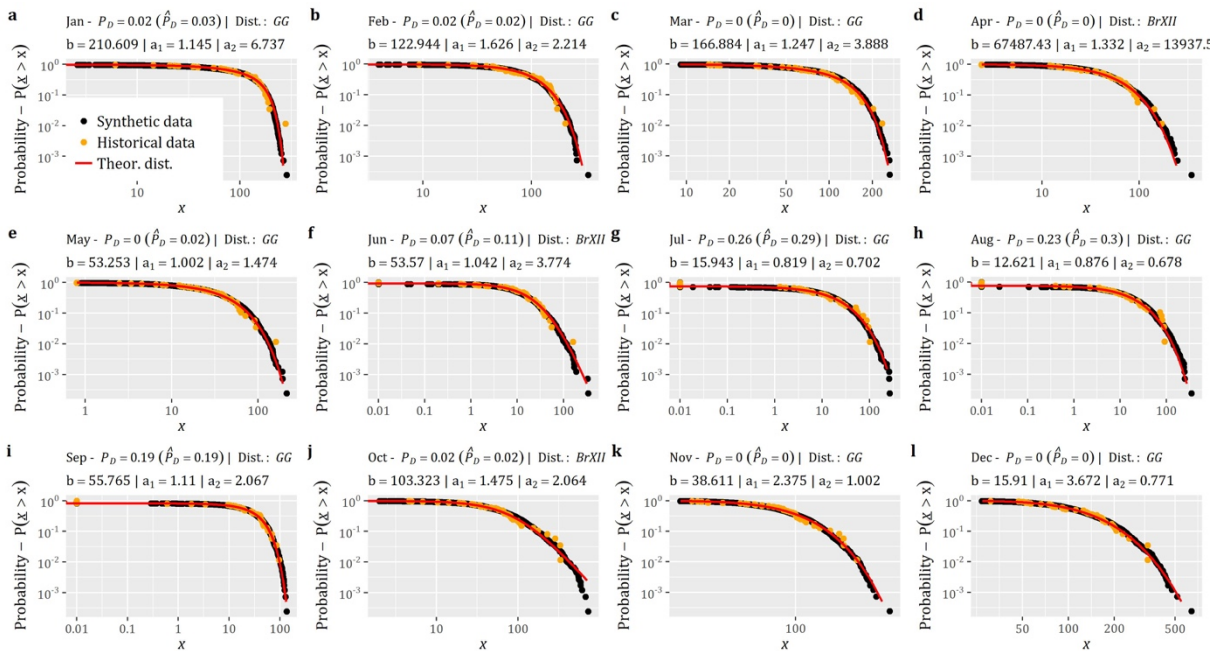


**Figure D.12** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site A (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.

APPENDIX D

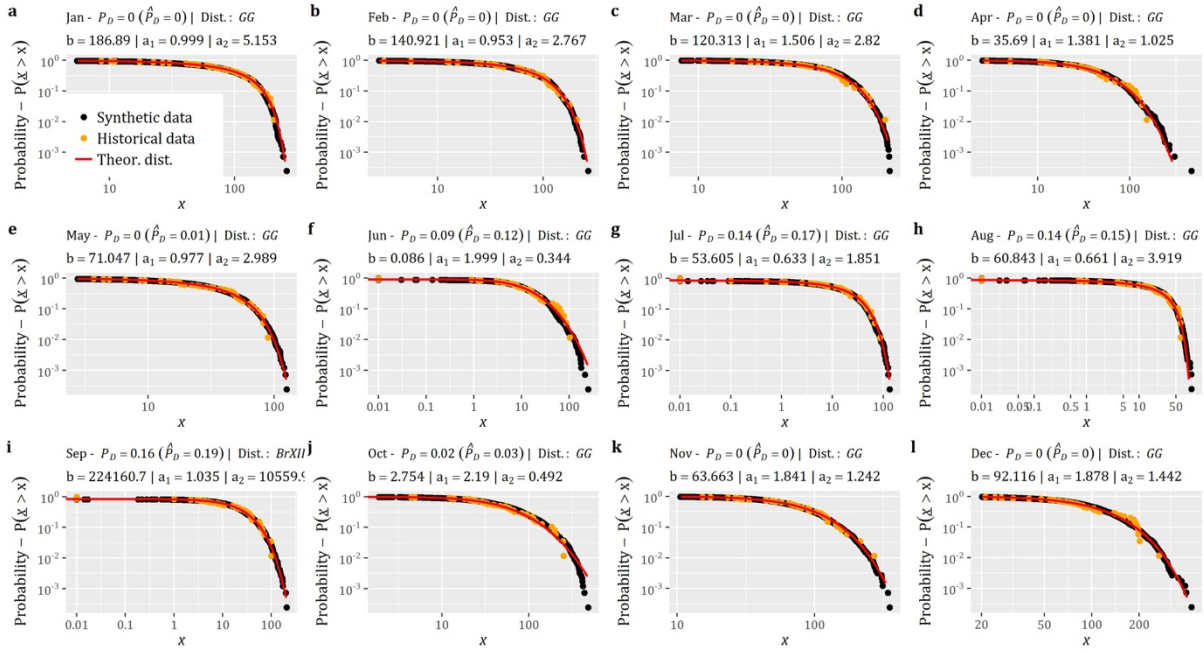


**Figure D.13** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site B (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.



**Figure D.14** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site C (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.



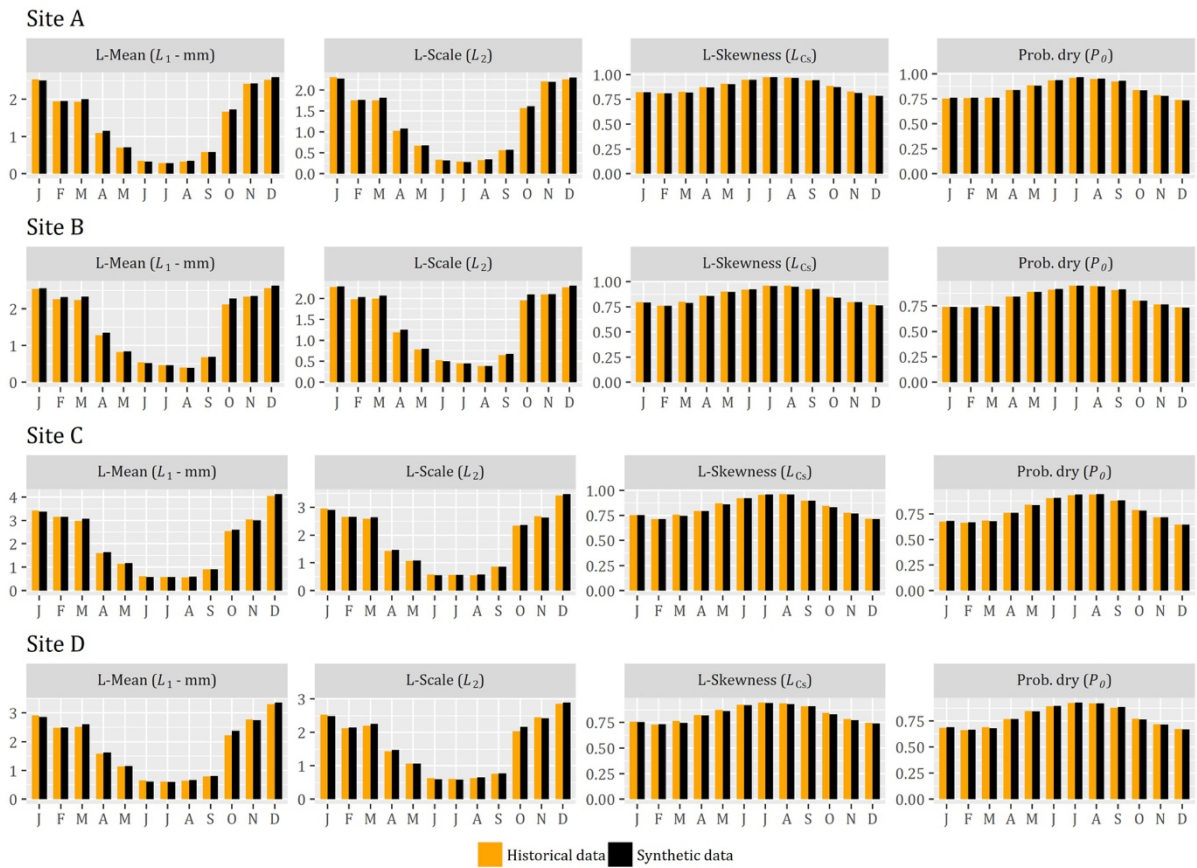


**Figure D.15** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at monthly time scale for site D (using the Weibull’s plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.

Regarding the daily time scale, **Figure D.16** provides a quick summary of the simulation results in terms of reproducing some key summary daily statistics (L-Mean, L-Scale, L-Skewness and probability dry), while **Figure D.17** presents a comparison among the daily historical and simulated lag-0 cross-correlation coefficients for sites A-D. As shown, the model reproduced the first three L-moments, as well as the historical probability dry with high accuracy, while it managed to satisfactory reassemble the lag-0 cross-correlation coefficients of all sites and for all months. The slight difference between the historical and simulated lag-0 cross-correlations can be attributed to the introduction of bias through the application of the proportional adjusting procedure.

To further assess the ability of the scheme to reproduce the target marginal distributions at the daily time scale, **Figure D.18** to **Figure D.21** present a monthly-based comparison among the empirical, simulated and theoretical distribution functions for sites A-D, that highlight the potential of the model to reproduce not only the target summary statistics (i.e., L-moments and correlations), but the entire target distribution functions. Furthermore, regarding the reproduction of the auto-dependence structure at the daily time scale, **Figure D.22** to **Figure D.25** validate the capabilities of the scheme to resemble the target daily autocorrelation functions (i.e., the fitted CAS to the historical data) for all sites and months. In addition, to explore the performance of the model in the intermediate temporal scales (i.e., those between the monthly and daily time scale) we performed a similar analysis as the one presented in the previous case study. **Figure D.26** to **Figure D.29**, **Figure D.30** to **Figure D.33** and **Figure D.34** to **Figure D.37** depict a monthly-based summary of the L-Scale ( $L_2^{(k)}$ ), L-Skewness ( $L_{CS}^{(k)}$ ) and probability dry ( $P_D^{(k)}$ ) as a function of time scale  $k$  respectively.

APPENDIX D

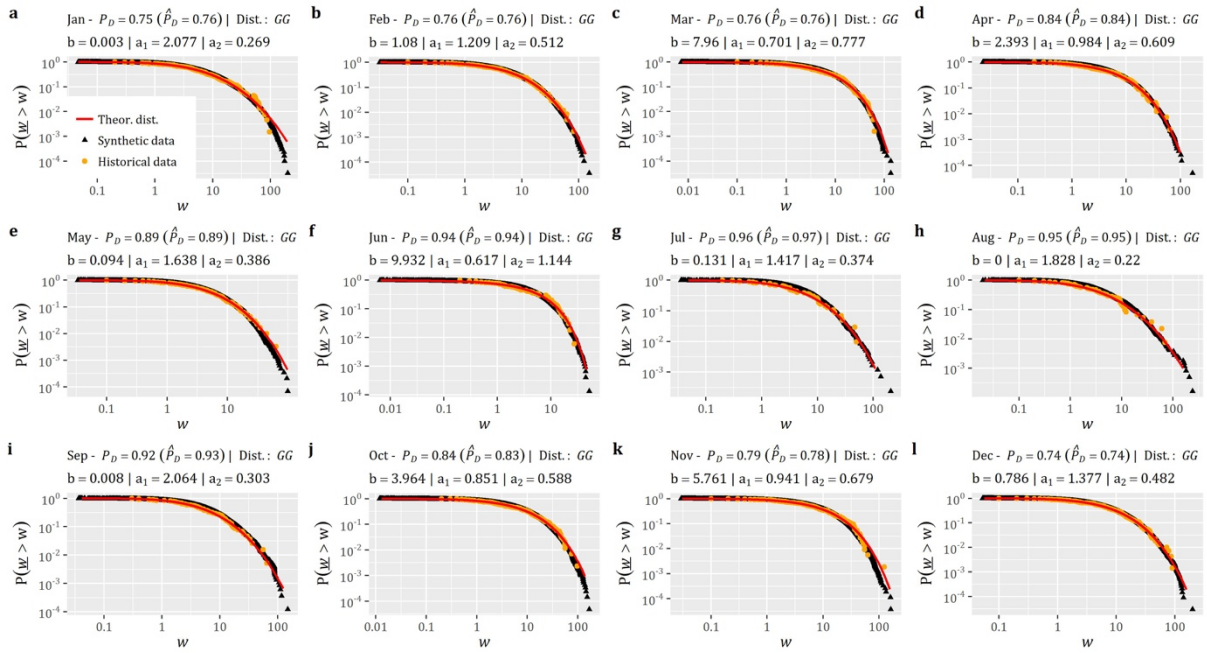


**Figure D.16** | Monthly-based comparison of daily empirical and simulated L-Mean, L-Scale and L-Skewness, as well as probability dry.

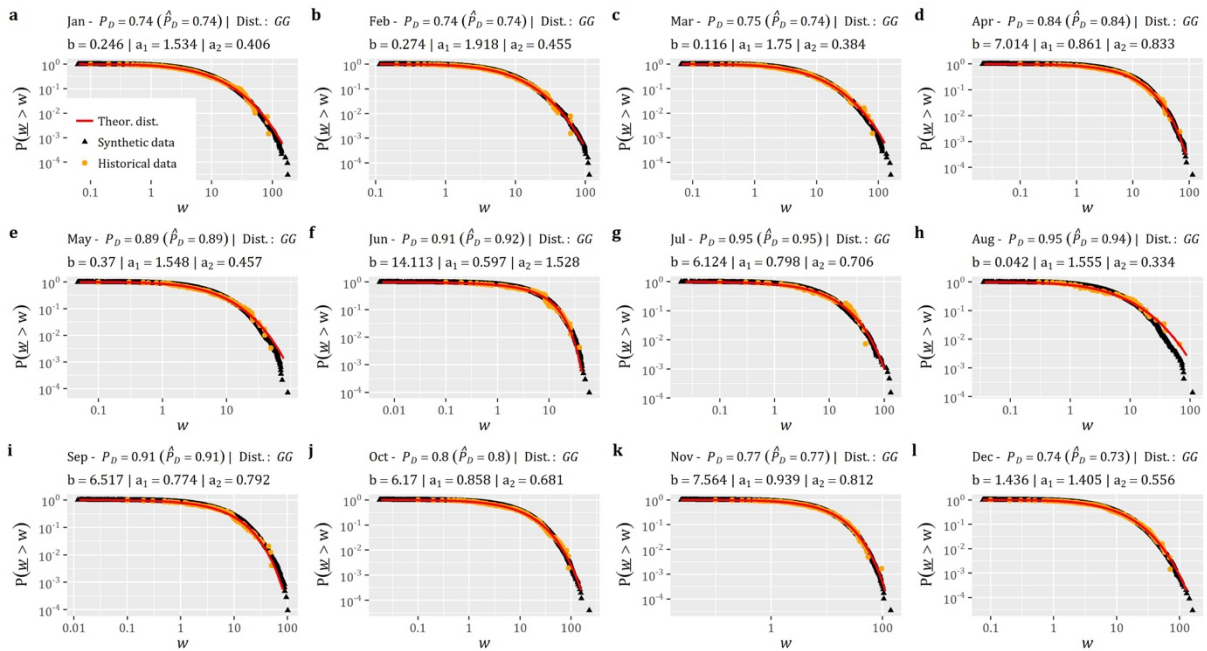


**Figure D.17** | Comparison of daily historical and simulated lag-0 cross-correlations for sites A-D.

## D.2 MULTI-TEMPORAL SIMULATION OF MULTIVARIATE DAILY RAINFALL PROCESSES

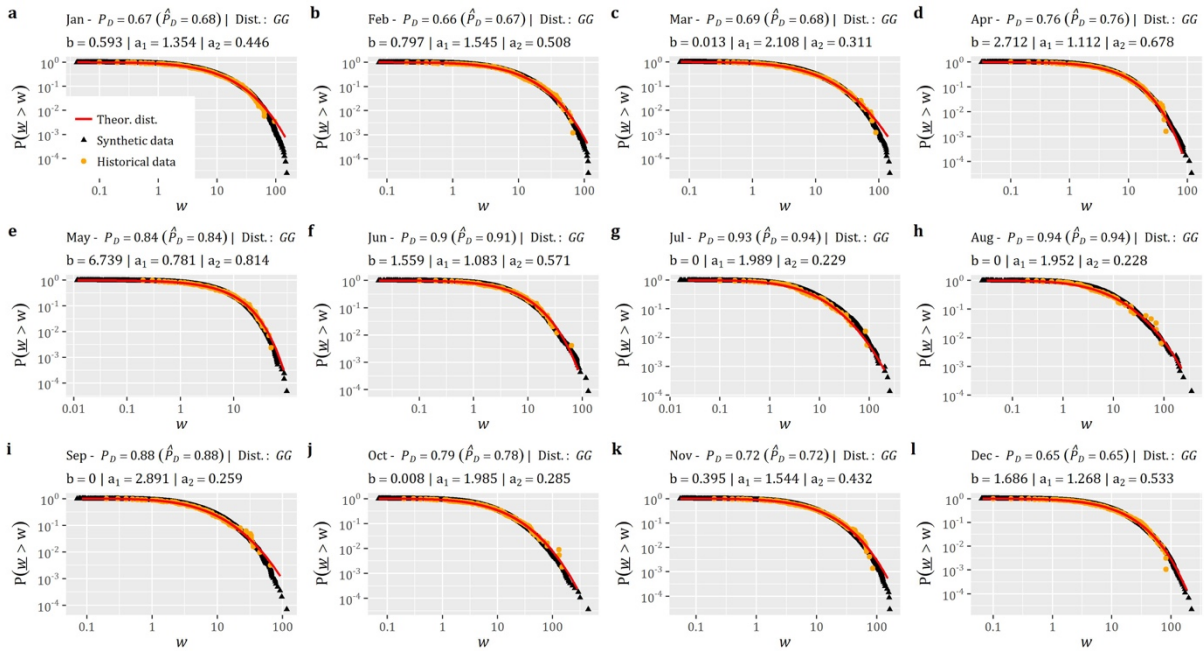


**Figure D.18** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site A (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.

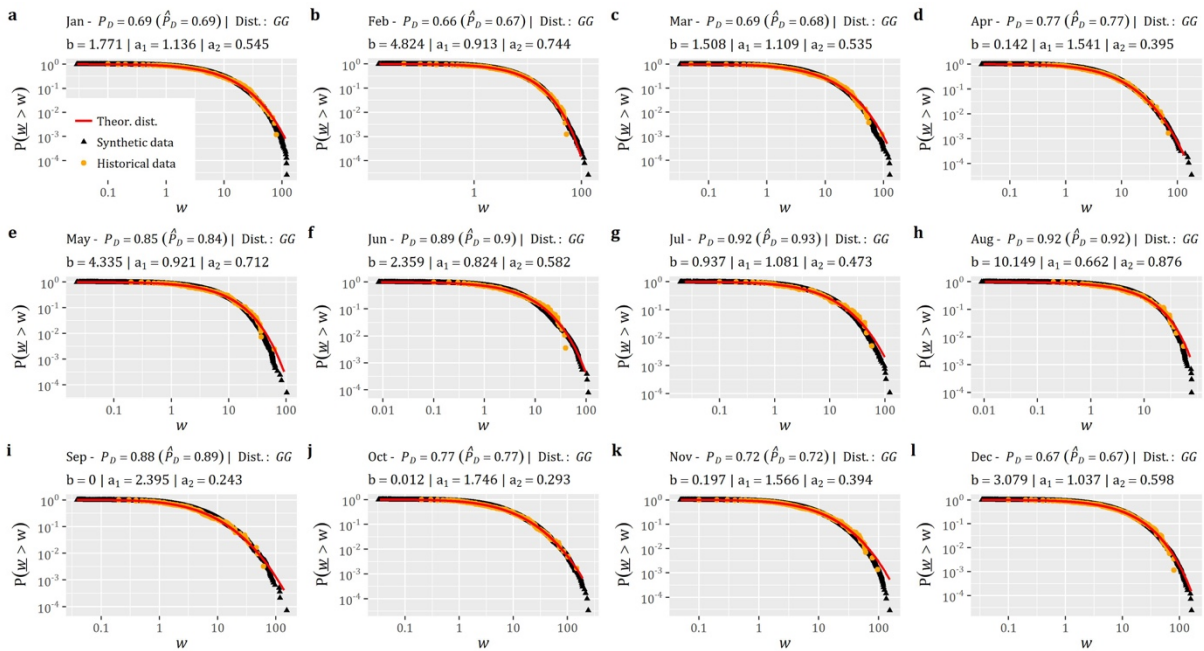


**Figure D.19** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site B (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.

APPENDIX D

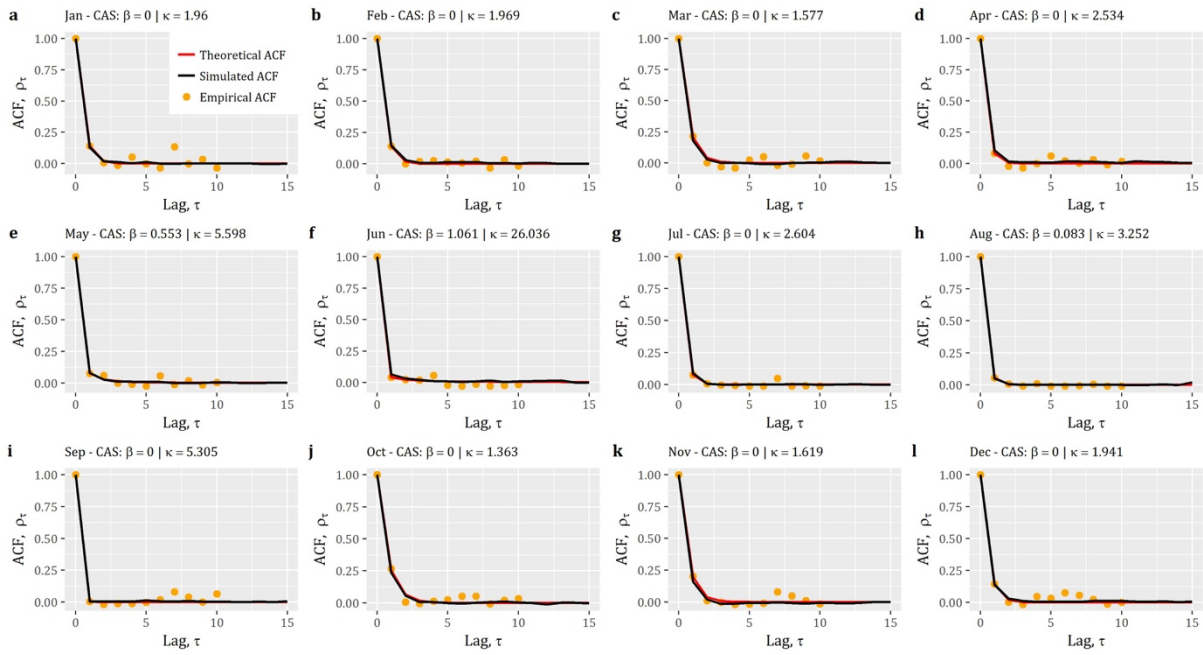


**Figure D.20** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site C (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.

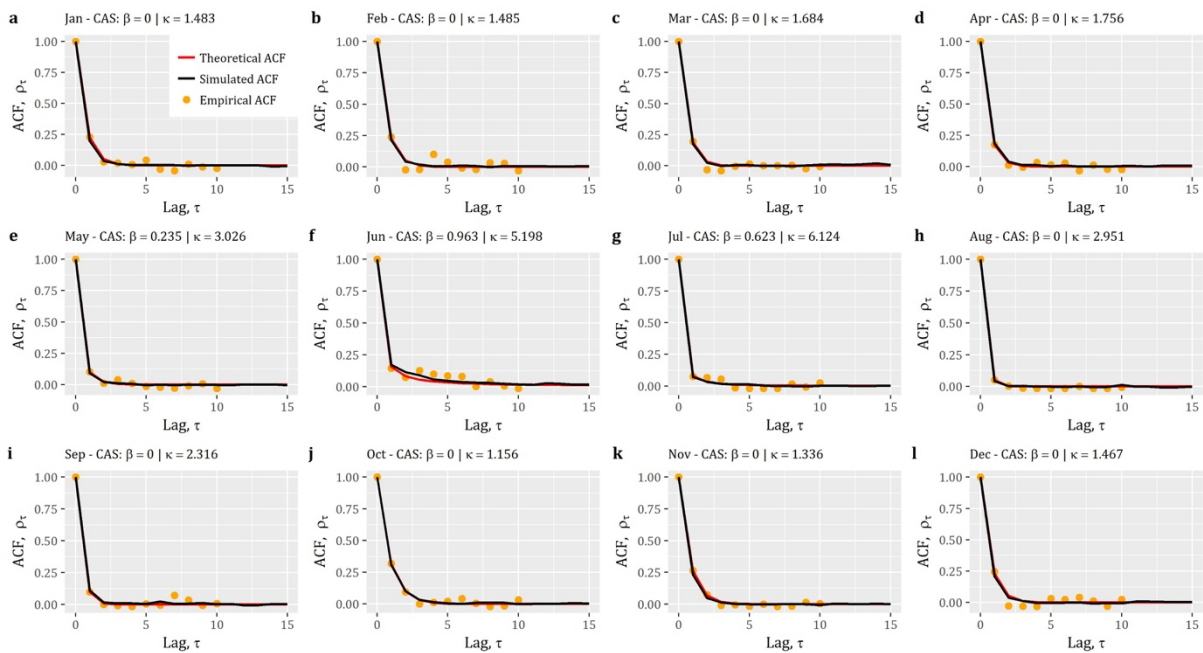


**Figure D.21** | Monthly-based comparison of empirical, simulated and theoretical distribution functions at daily time scale for site D (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry.

## D.2 MULTI-TEMPORAL SIMULATION OF MULTIVARIATE DAILY RAINFALL PROCESSES

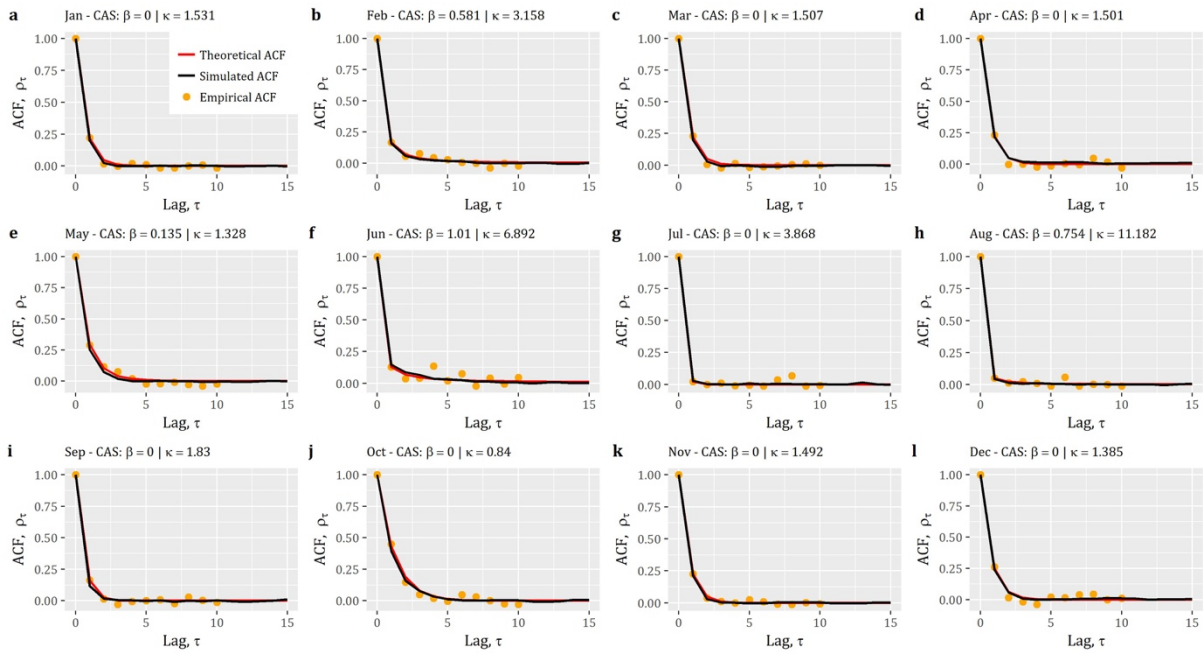


**Figure D.22** | Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site A; the parameters of CAS are given on the title of each subplot.

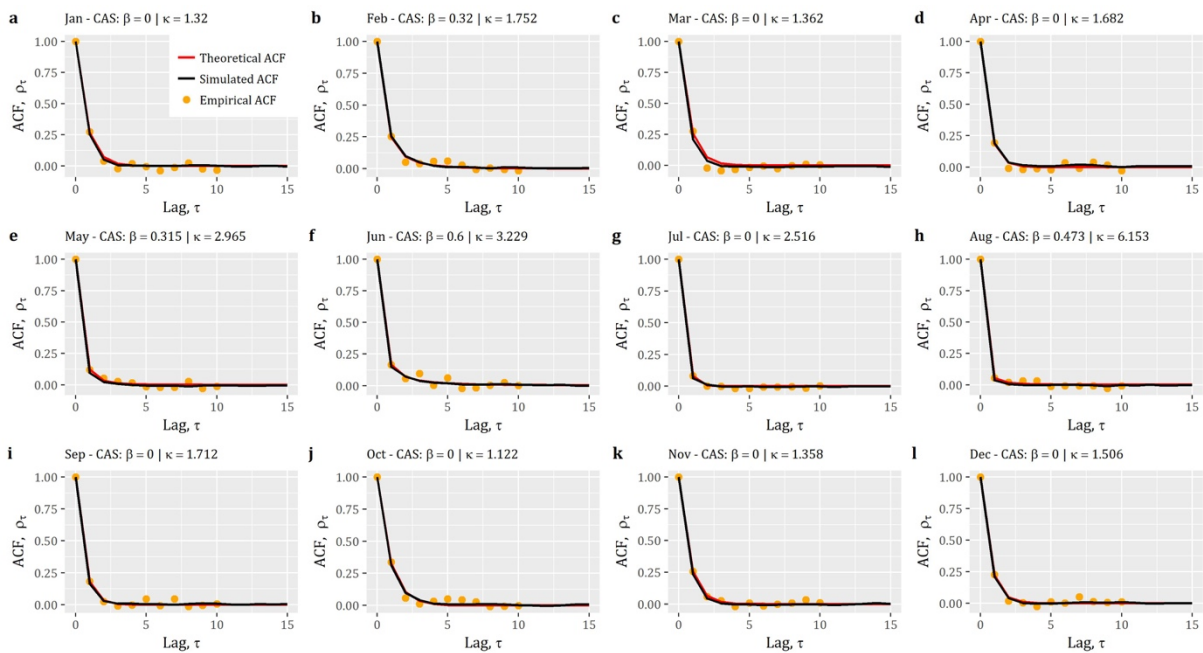


**Figure D.23** | Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site B; the parameters of CAS are given on the title of each subplot.

APPENDIX D

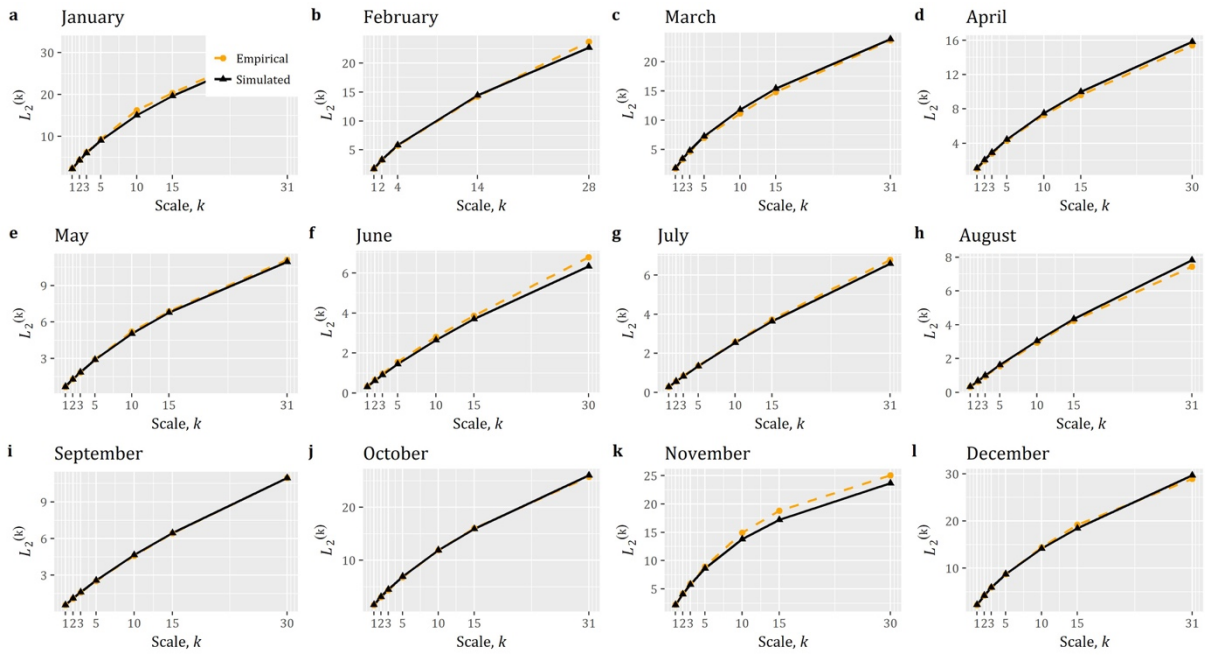


**Figure D.24** | Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site C; the parameters of CAS are given on the title of each subplot.

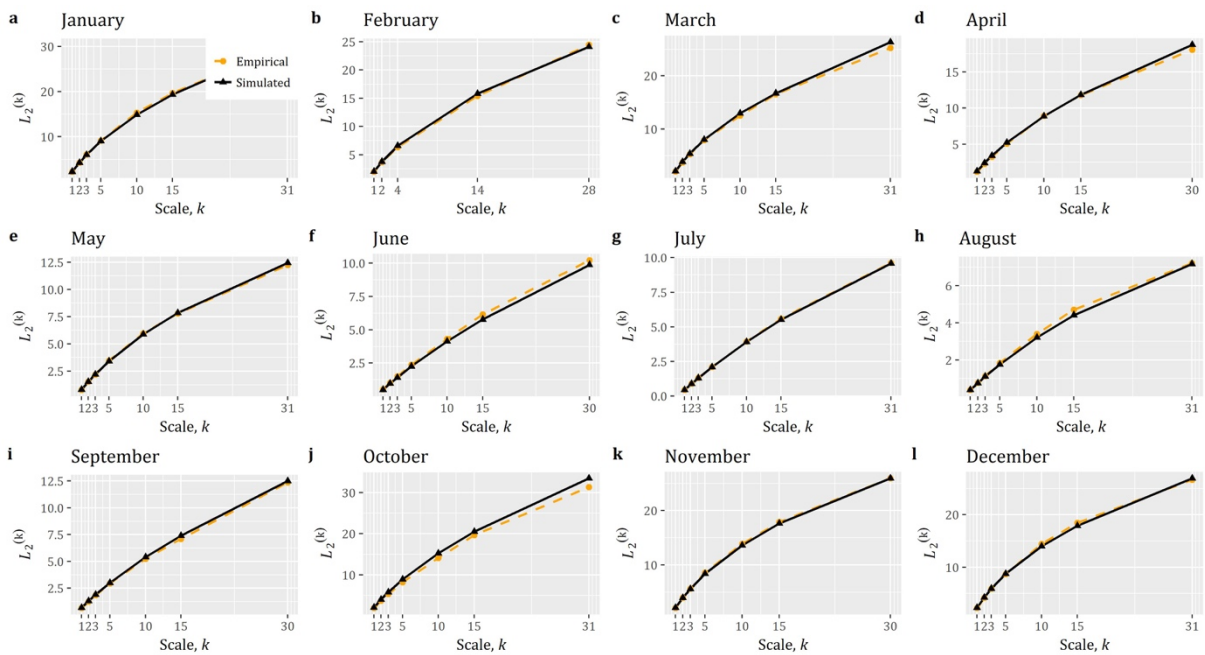


**Figure D.25** | Monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF) at daily time scale for site D; the parameters of CAS are given on the title of each subplot.

## D.2 MULTI-TEMPORAL SIMULATION OF MULTIVARIATE DAILY RAINFALL PROCESSES

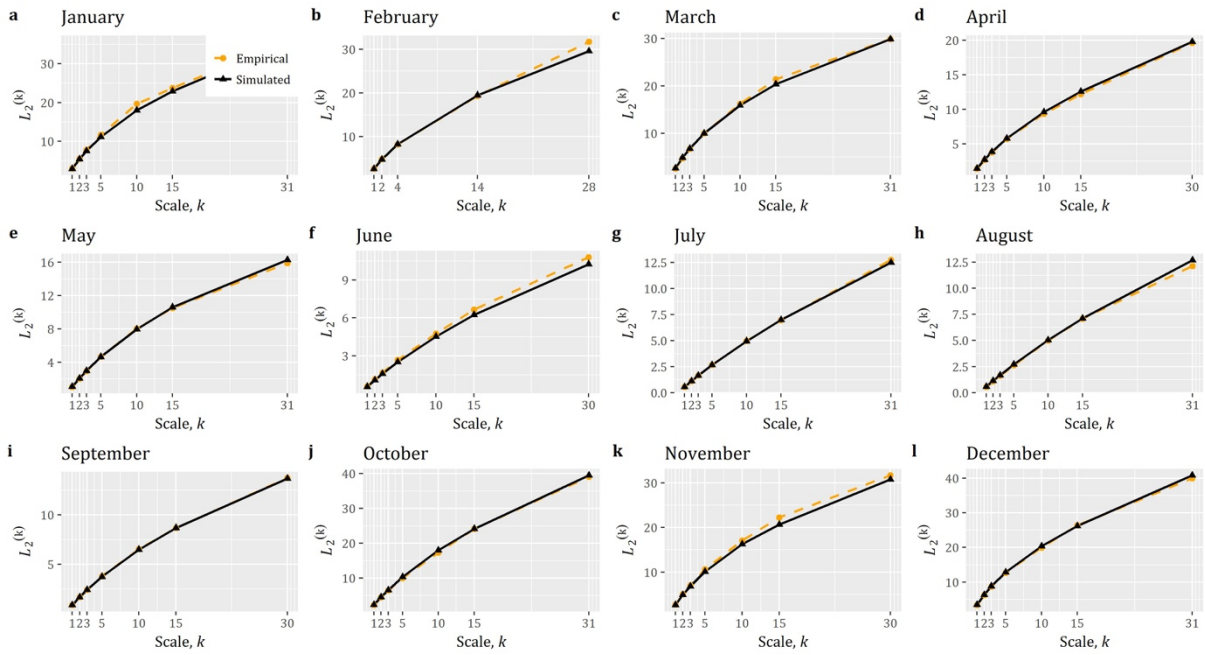


**Figure D.26** | Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale  $k$  for site A.

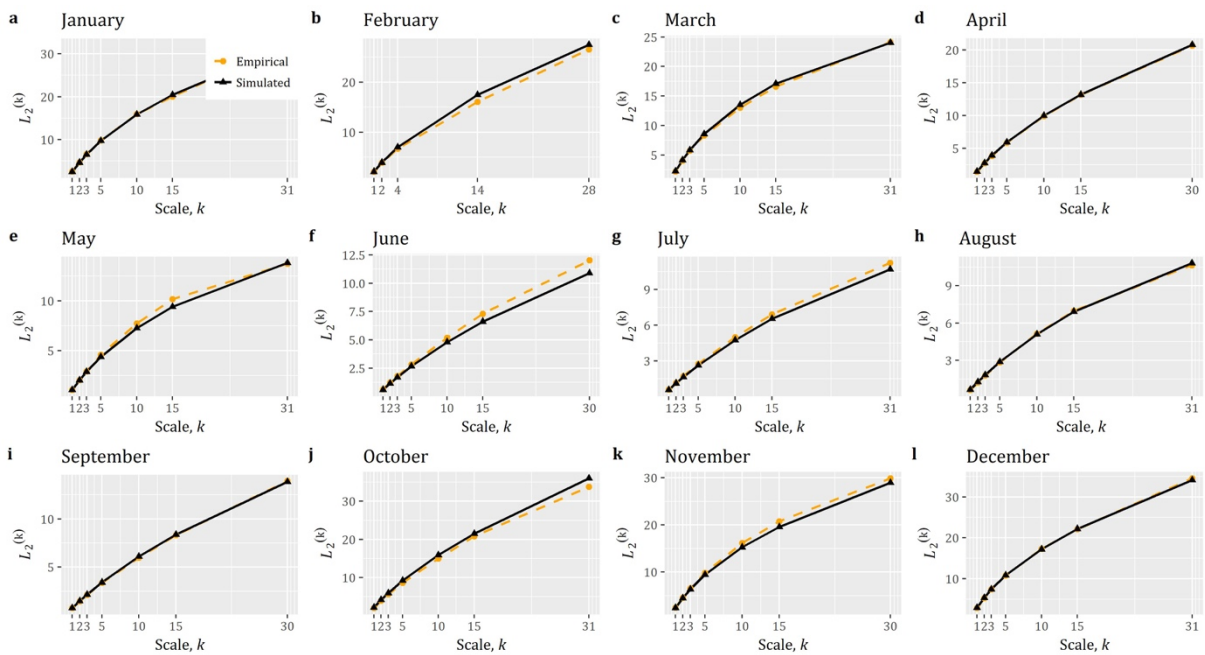


**Figure D.27** | Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale  $k$  for site B.

APPENDIX D

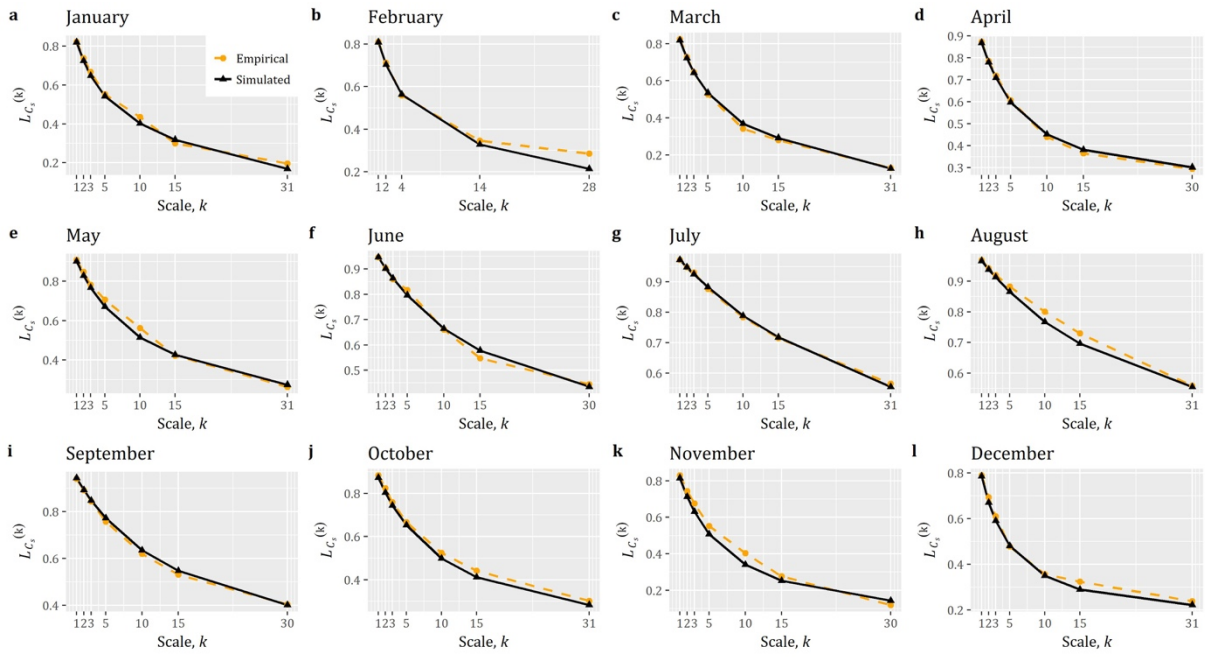


**Figure D.28** | Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale  $k$  for site C.

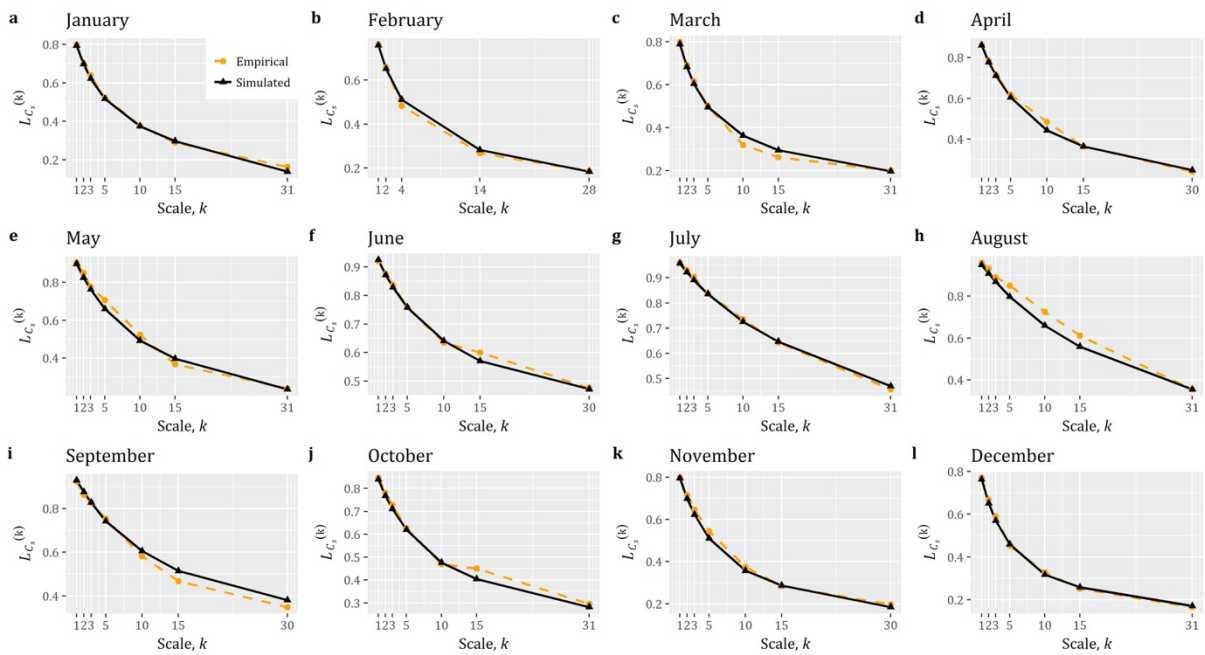


**Figure D.29** | Monthly-based summary of L-Scale ( $L_2$ ) as a function of aggregation scale  $k$  for site D.



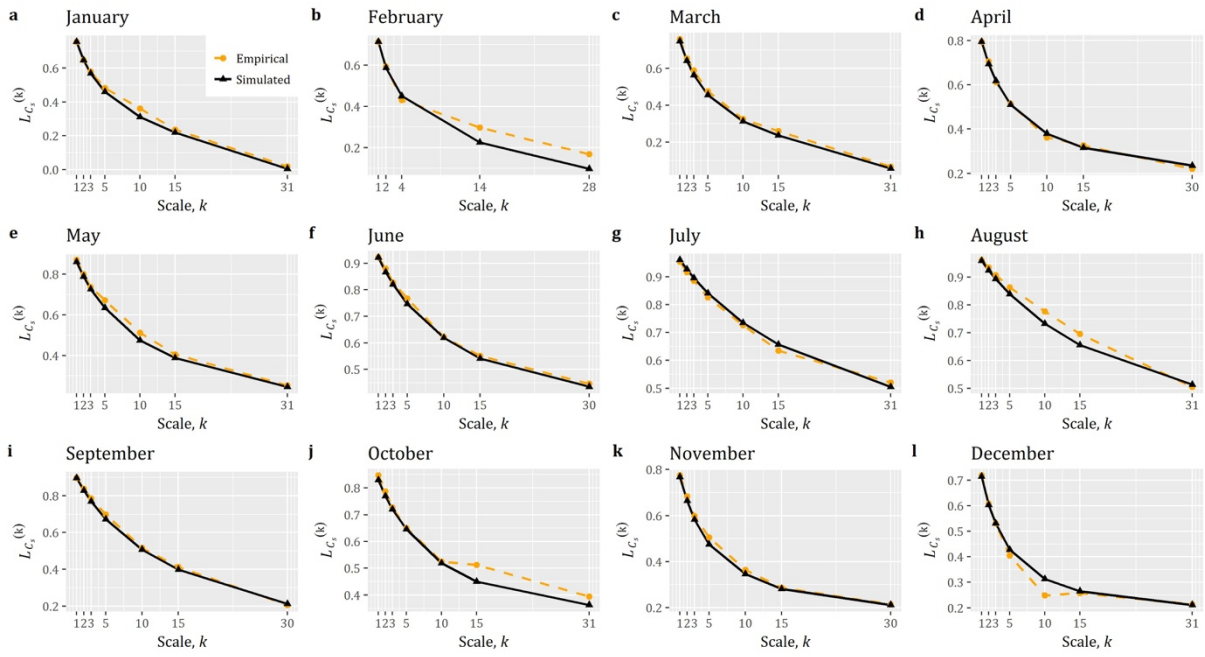


**Figure D.30** | Monthly-based summary of L-Skewness ( $L_{C_s}$ ) as a function of aggregation scale  $k$  for site A.

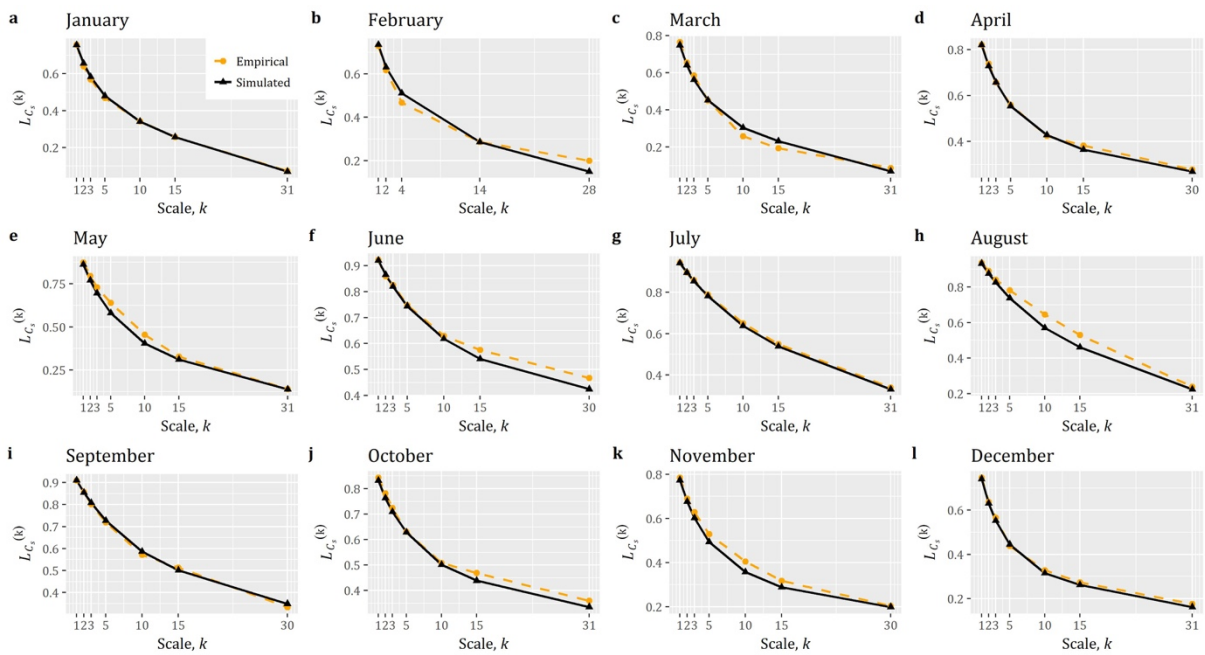


**Figure D.31** | Monthly-based summary of L-Skewness ( $L_{C_s}$ ) as a function of aggregation scale  $k$  for site B.

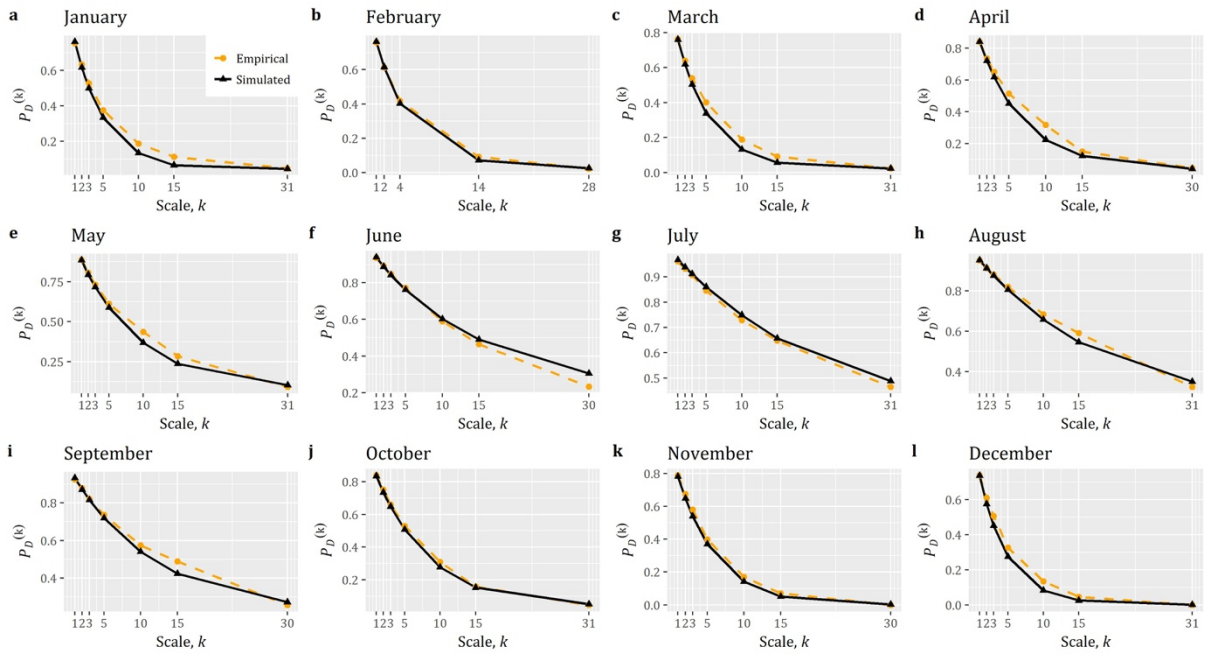
APPENDIX D



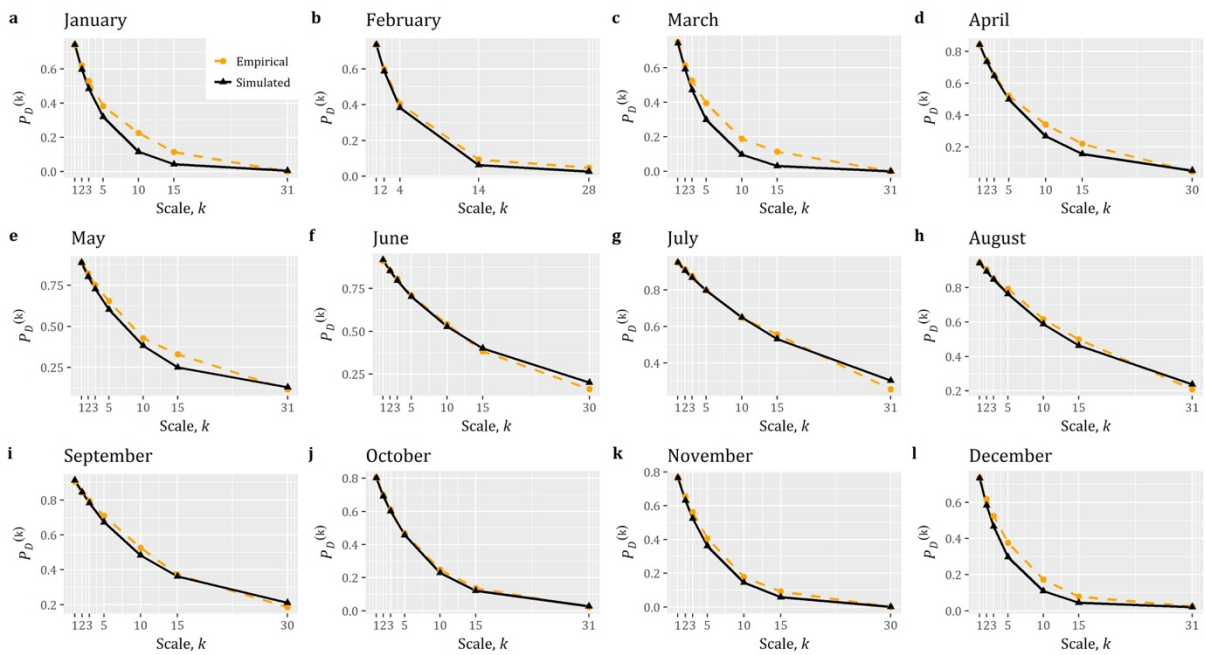
**Figure D.32** | Monthly-based summary of L-Skewness ( $L_{C_s}$ ) as a function of aggregation scale  $k$  for site C.



**Figure D.33** | Monthly-based summary of L-Skewness ( $L_{C_s}$ ) as a function of aggregation scale  $k$  for site D.

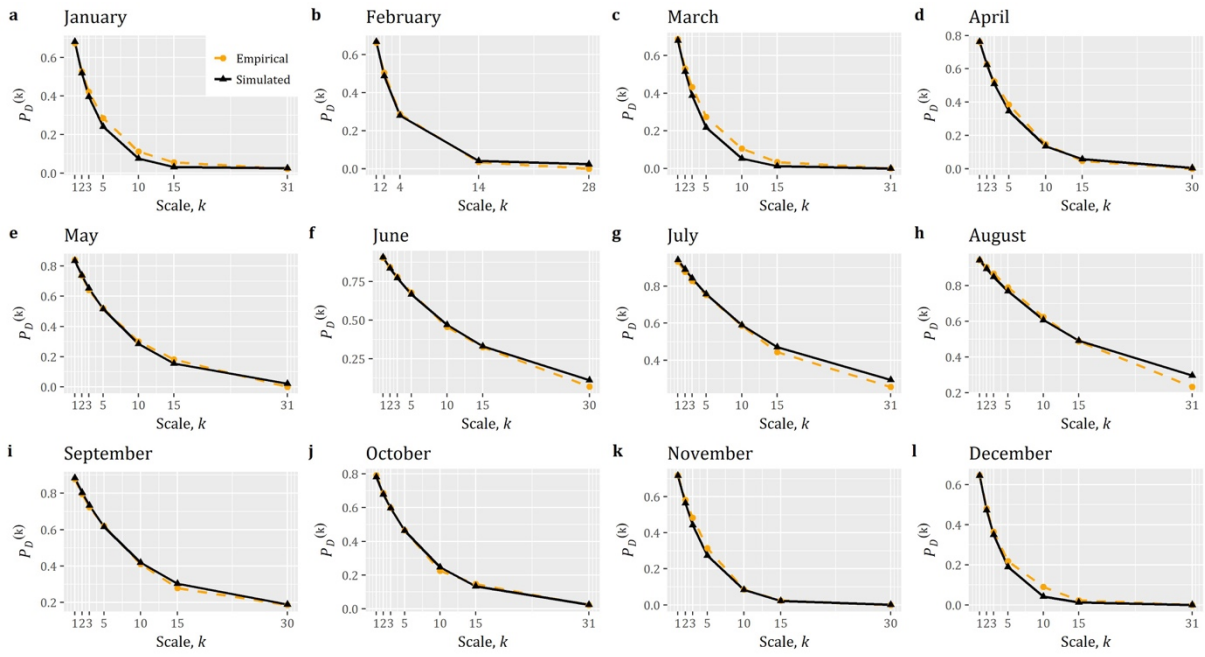


**Figure D.34** | Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$  for site A.

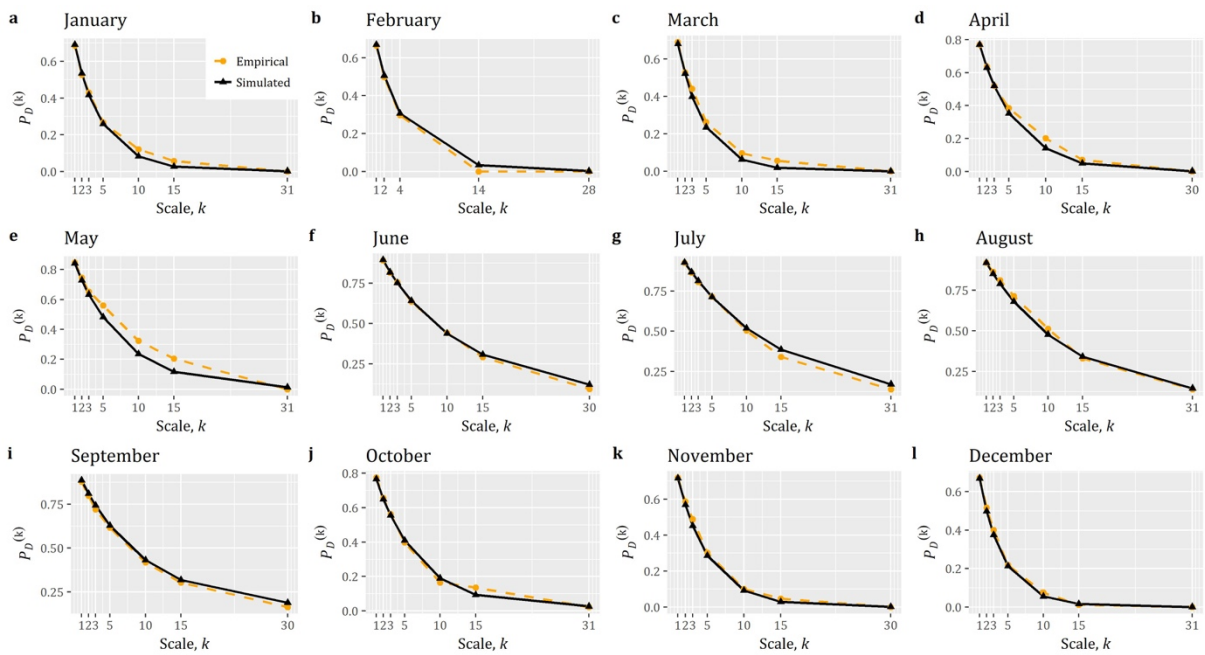


**Figure D.35** | Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$  for site B.

APPENDIX D

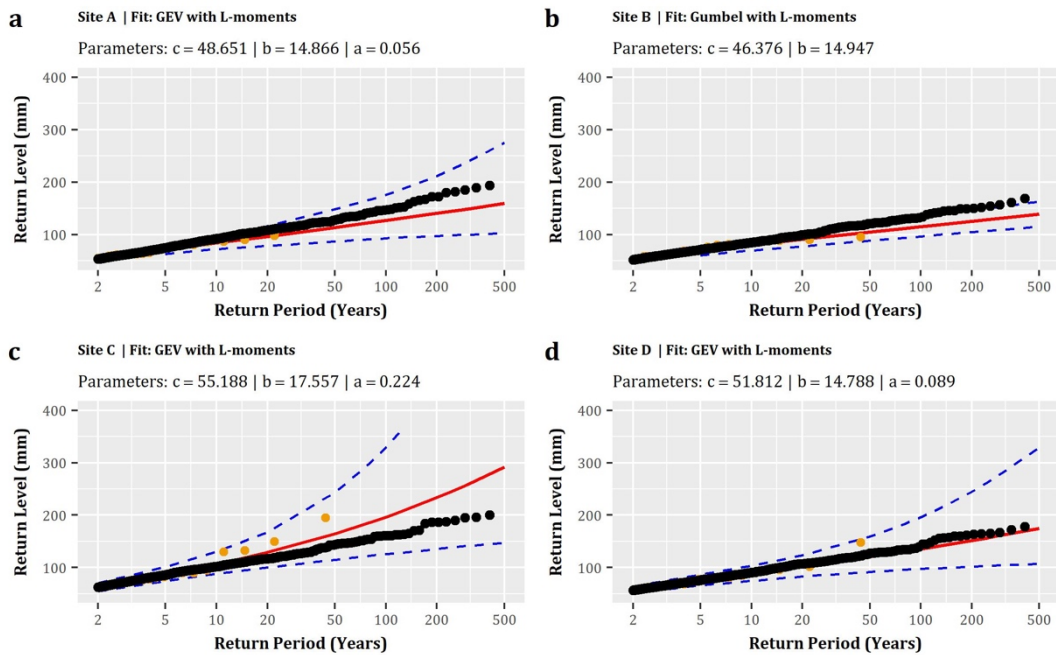


**Figure D.36** | Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$  for site C.



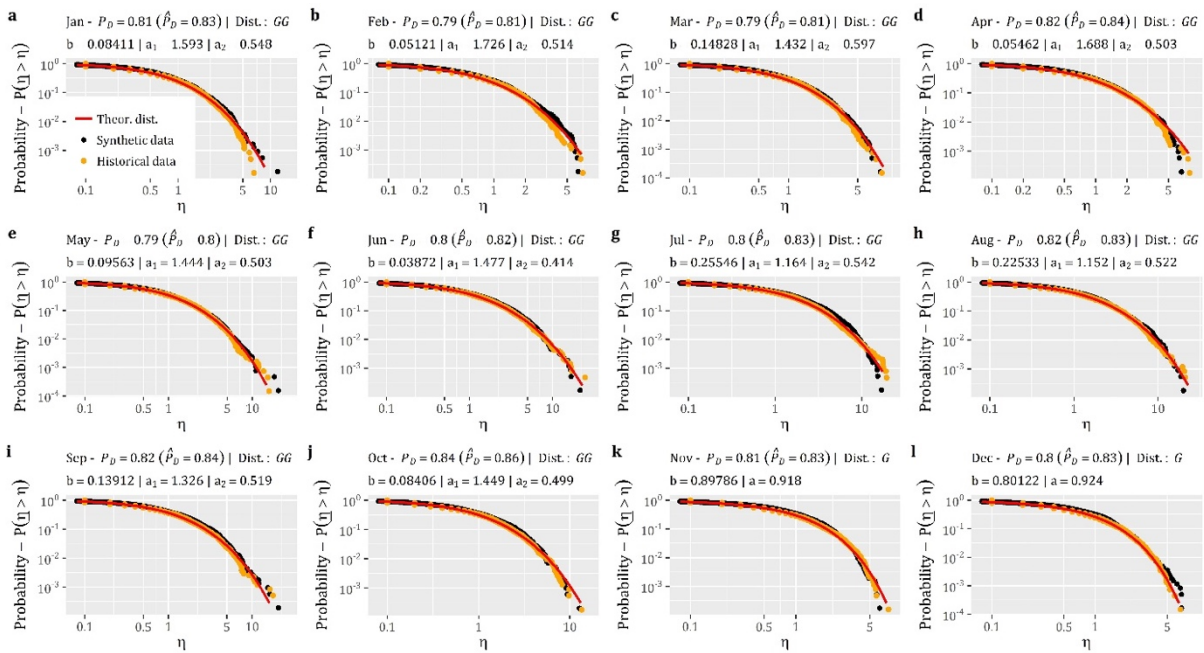
**Figure D.37** | Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$  for site D.

Finally, in order to investigate the behavior of the three-level configuration regarding the simulation of daily extreme events, the series of historical and simulated annual rainfall maxima has been extracted and depicted as a function of the return period in **Figure D.38**. The plot also depicts the parameters of the fitted (using the L-moments method) to the historical data  $\mathcal{GEV}$  distribution (i.e., Eq. (7.6)). It is noted that Site B (i.e., the annual rainfall maxima obtained from Atalanti gauge) exhibited negative shape parameter ( $a = -0.050$ ), which is not consistent from hydrological point of view, since it implies that the distribution is bounded from above. Thereby, for this site we fitted the Gumbel distribution ( $a = 0$ ; in Eq. (7.6)). Regarding site B, visual inspection of **Figure D.38**, reveals that the scheme generates annual daily maxima with arguably heavier tails; a fact also confirmed by the shape parameter of the  $\mathcal{GEV}$  distribution fitted to the simulated data. In this case, the identified  $\mathcal{GEV}$  parameters are for Site B- $\mathcal{GEV}(c = 46.941, b = 16.091, a = 0.068)$ . As far it concerns the behavior of the extremes of sites A, C and D, the historical, simulated data are in better agreement, while the theoretical  $\mathcal{GEV}$  distribution (extracted from the historical data) has in both cases a positive value. Further to this, it is noted that the simulated annual daily maxima of sites A, C and D lie within the 95% confidence intervals (estimated using the parametric bootstrap method). The parameters of the fitted  $\mathcal{GEV}$  distribution to the simulated annual rainfall maxima are: Site A- $\mathcal{GEV}(c = 47.796, b = 18.549, a = 0.073)$ , Site C- $\mathcal{GEV}(c = 56.829, b = 19.479, a = 0.143)$  and site D- $\mathcal{GEV}(c = 51.382, b = 16.816, a = 0.116)$ , which are relatively close (considering the associated large uncertainty) to those obtained from the historical maxima (see the titles of **Figure D.38a-d**).

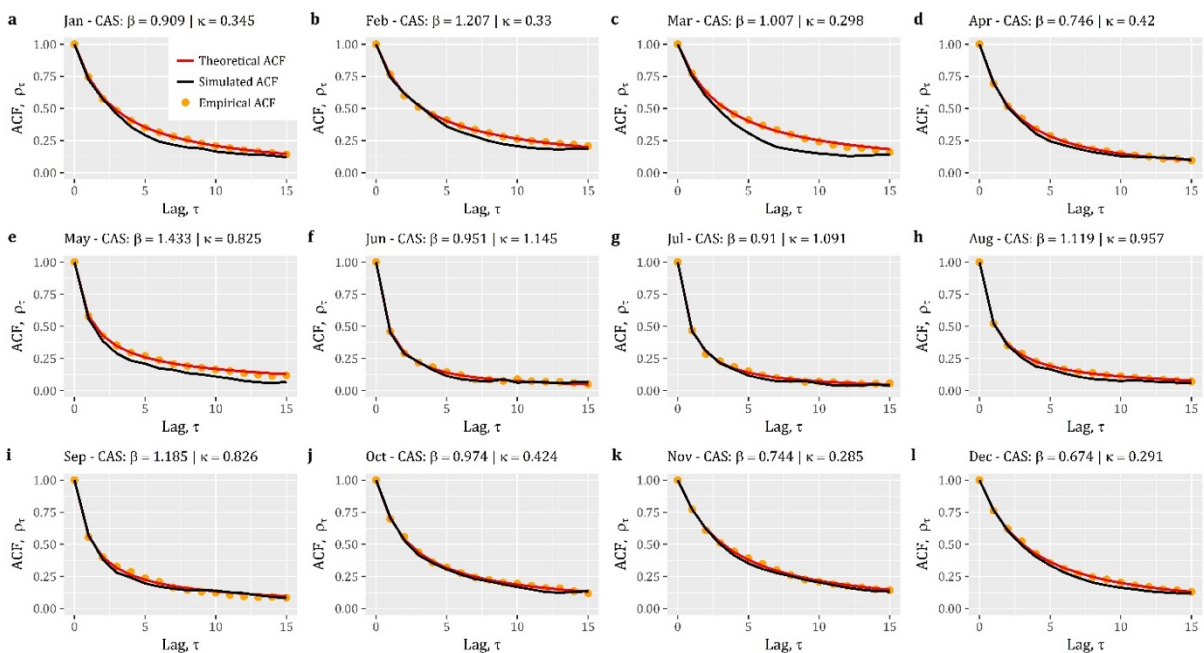


**Figure D.38** | Empirical (•) and simulated (●) daily annual rainfall maxima of sites A-D, as a function of the return period. The solid red line (—) depicts the fitted to historical data Generalized Extreme Value ( $\mathcal{GEV}$ ) distribution (parameters: location ( $c$ ), scale ( $b$ ) and shape ( $a$ )). The dashed blue line (---) represents the 95% confidence intervals (estimated using the parametric bootstrap method).

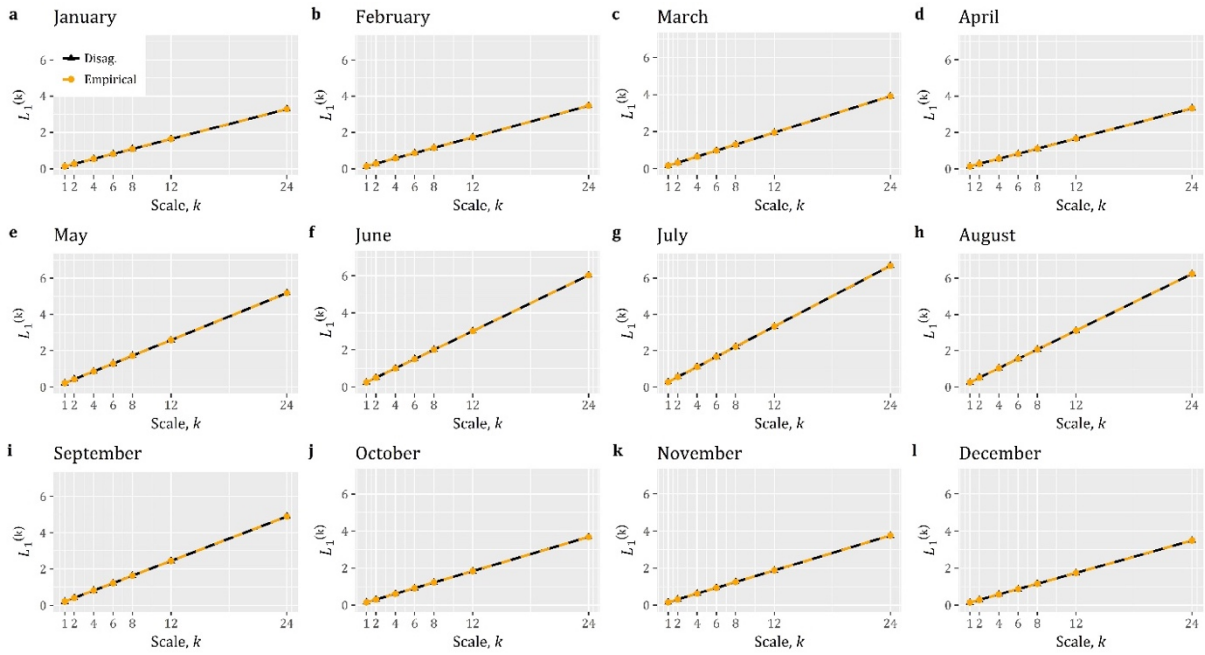
D.3 SUPPLEMENTARY MATERIAL OF SECTION 7.5



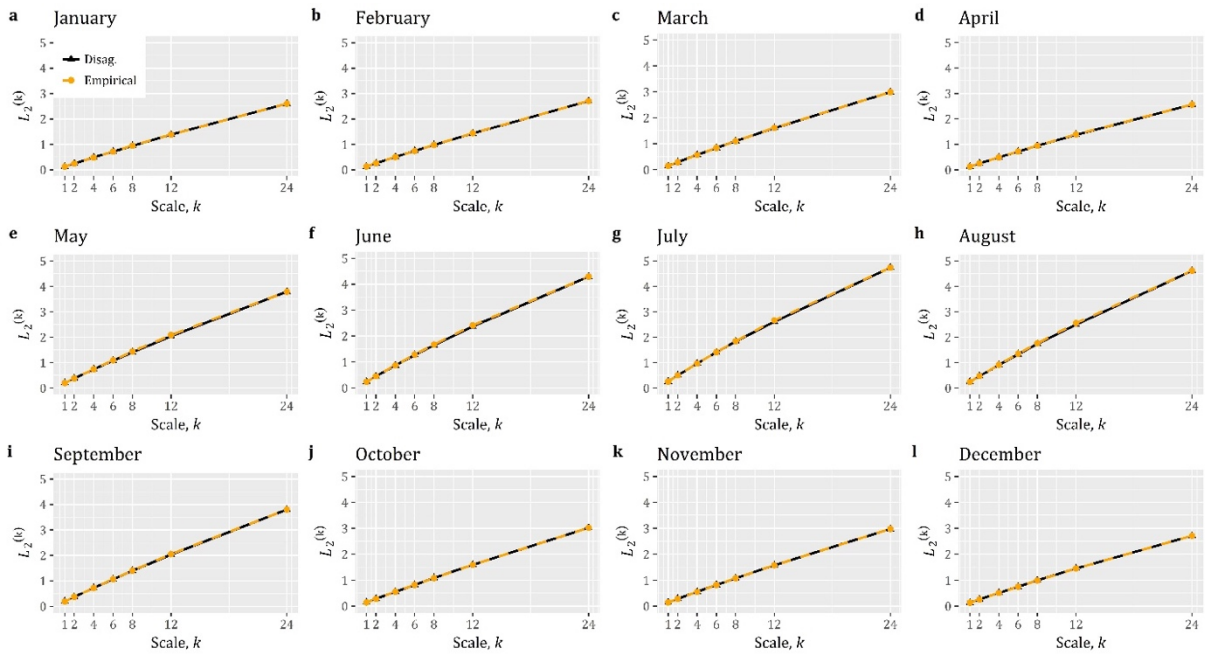
**Figure D.39** | Disaggregated hourly rainfall (non-zero) - monthly-based comparison of empirical, simulated and theoretical distribution functions (using the Weibull's plotting position). The title of each subplot provides the selected distribution and its parameters, as well as the historical ( $p_D$ ) and simulated ( $\hat{p}_D$ ) values of probability dry



**Figure D.40** | Disaggregated hourly rainfall - monthly-based comparison of empirical, simulated and theoretical autocorrelation function (ACF); the parameters of CAS are given on the title of each subplot.

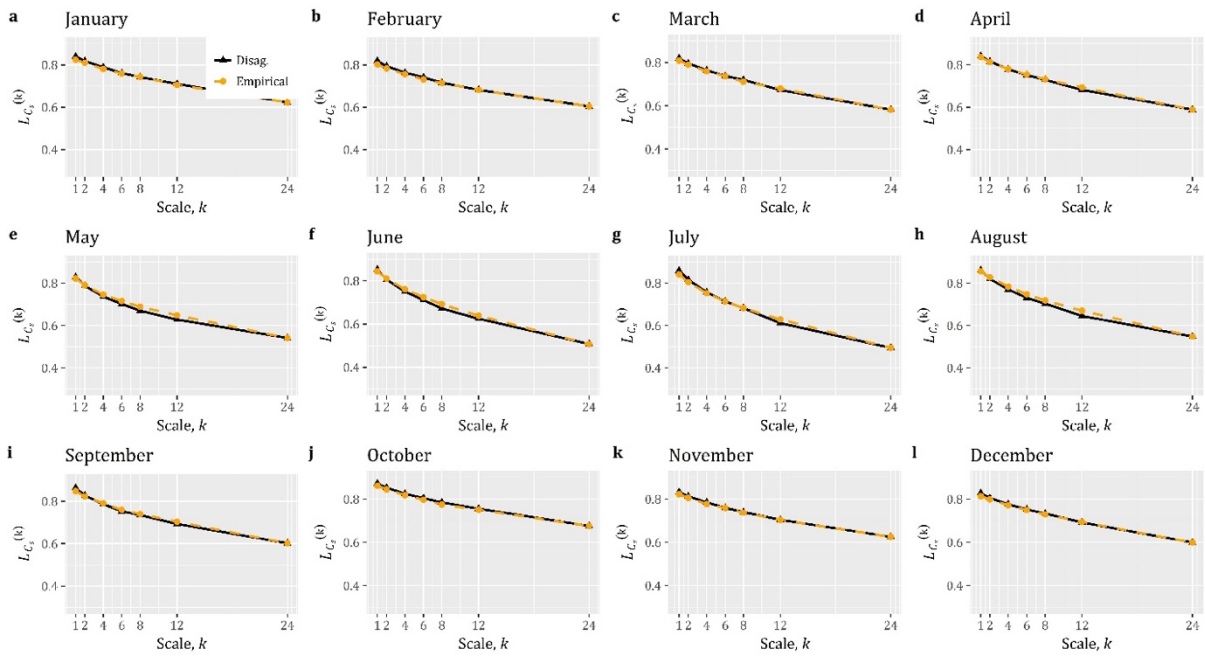


**Figure D.41** | Disaggregated hourly rainfall - Monthly-based summary of L-mean ( $L_1$ ) as a function of aggregation scale  $k$ .

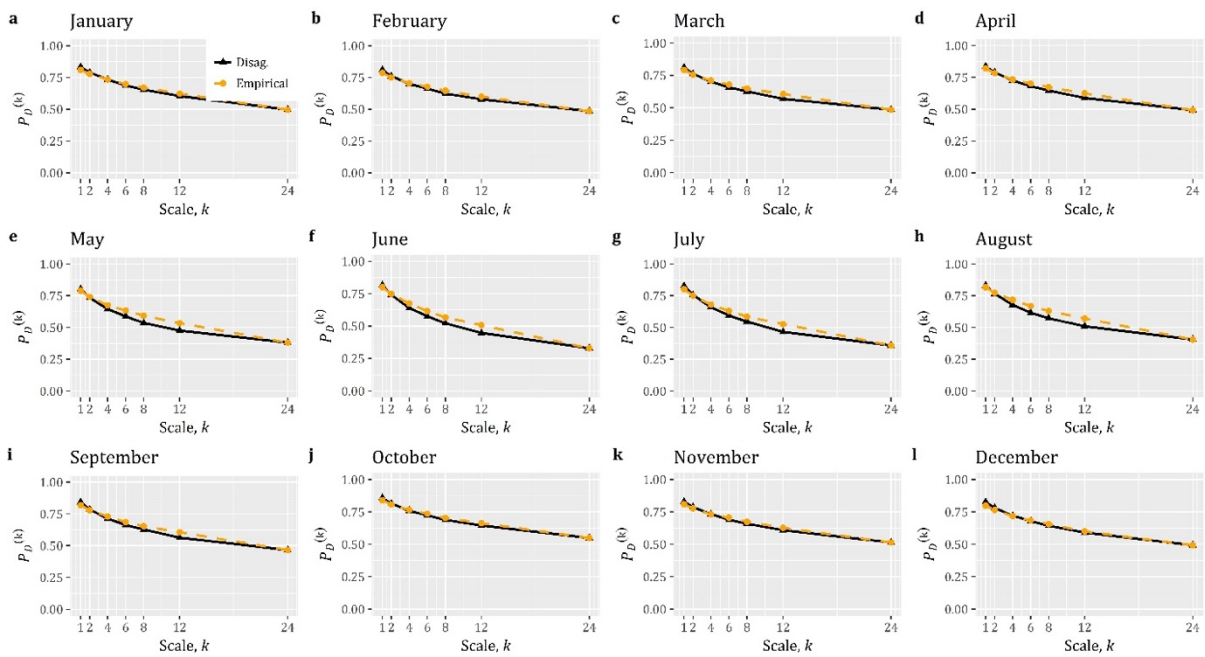


**Figure D.42** | Disaggregated hourly rainfall - Monthly-based summary of L-scale ( $L_2$ ) as a function of aggregation scale  $k$ .

APPENDIX D

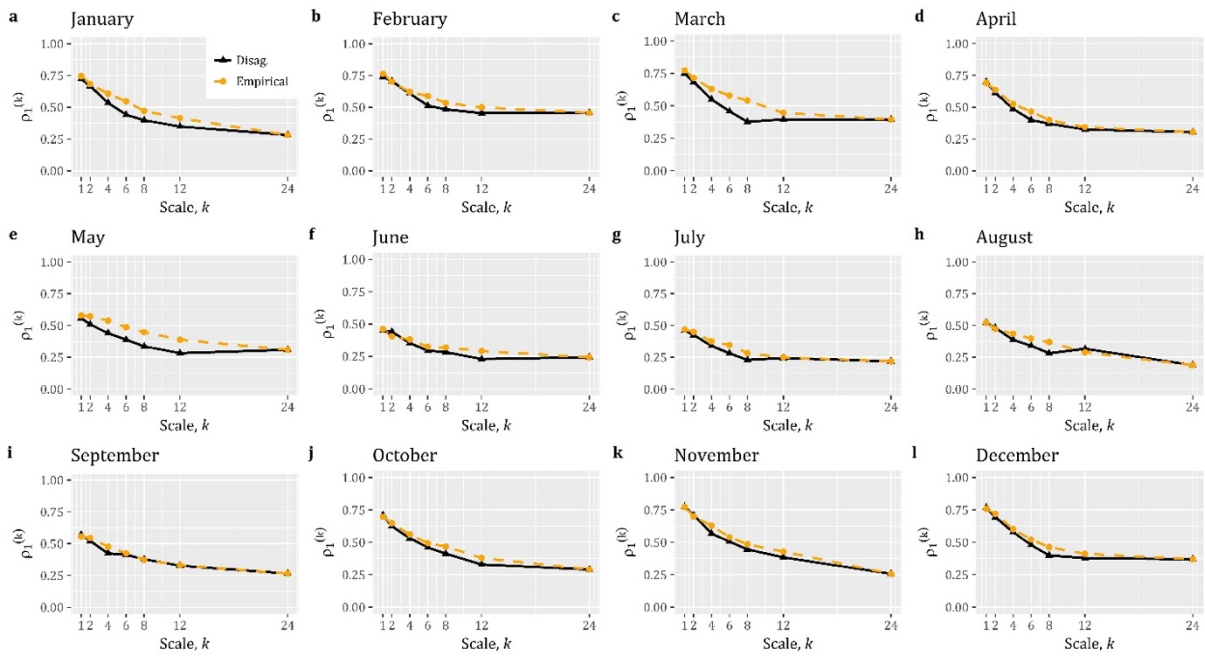


**Figure D.43** | Disaggregated hourly rainfall - Monthly-based summary of L-skewness ( $L_{Cs}$ ) as a function of aggregation scale  $k$ .



**Figure D.44** | Disaggregated hourly rainfall - Monthly-based summary of prob. dry ( $P_D$ ) as a function of aggregation scale  $k$ .





**Figure D.45** | Disaggregated hourly rainfall - Monthly-based summary of lag-1 autocorrelation coefficient ( $\rho_1$ ) as a function of aggregation scale  $k$ .

# LIST OF PUBLICATIONS

---

## 1 Publications in scientific journals

### 1.1 Phd related

- [1] **Tsoukalas I**, Papalexiou, S.M., Efstratiadis A., and Makropoulos, C., (2018). *A Cautionary Note on the Reproduction of Dependencies through Linear Stochastic Models with Non-Gaussian White Noise*, *Water*, 10 (6), 771.
- [2] **Tsoukalas I**, Makropoulos C., and Koutsoyiannis D., (2018). *Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions*, *Water Resources Research*. DOI:10.1029/2017WR022462
- [3] **Tsoukalas, I.**, Efstratiadis, A., and Makropoulos, C. (2018). *Stochastic periodic autoregressive to anything (SPARTA): Modeling and simulation of cyclostationary processes with arbitrary marginal distributions*. *Water Resources Research*, 54(1), 161-185.
- [4] **Tsoukalas, I.**, Efstratiadis, A., and Makropoulos, C. (2018). *Building a puzzle to solve a riddle: a new approach to multi-temporal stochastic simulation* (in review).
- [5] **Tsoukalas, I.**, Kossieris, P., Efstratiadis, A., and Makropoulos, C., (2016). *Surrogate-enhanced evolutionary annealing simplex algorithm for effective and efficient optimization of water resources problems on a budget*. *Environmental Modelling & Software*, 77 122-142. DOI:10.1016/j.envsoft.2015.12.008.
- [6] **Tsoukalas I.** and Makropoulos C. (2015). *Multiobjective optimization on a budget: Exploring surrogate modelling for robust multi-reservoir rules generation under hydrological uncertainty*, *Environmental Modelling and software*, 69 396-413, DOI:10.1016/j.envsoft.2014.09.023.
- [7] **Tsoukalas I.**, and C. Makropoulos., (2015). *A Surrogate Based Optimization Approach for the Development of Uncertainty-Aware Reservoir Operational Rules: The Case of Nestos Hydrosystem*. *Water Resources Management* 10/2015; 29(13), 4719–4734. DOI:10.1007/s11269-015-1086-8.
- [8] **Tsoukalas I.**, Dimas P., and Makropoulos C., (2015). *Hydrosystem optimization on a budget: Investigating the potential of surrogate based optimization techniques*. *Mathematics in Engineering, Science and Aerospace (MESA)*, 6(4).

### 1.2 Other publications

- [9] Papaioannou G., A. Efstratiadis, L. Vasiliades, A. Loukas, S.M. Papalexiou, A. Koukouvinos, **I. Tsoukalas**, and P. Kossieris. (2018). *An operational method for Floods Directive implementation in ungauged urban areas*, *Hydrology*, 5(2), 24.
- [10] Psarrou, E., **Tsoukalas, I.**, and Makropoulos, C. (2018). *A Monte-Carlo based method for the optimal placement and operation scheduling of sewer mining units in urban wastewater networks*. *Water*, 10(2), 200.
- [11] **Tsoukalas, I. K.**, Makropoulos, C. K., and Michas, S. N. (2017). *Identification of potential sewer mining locations: a Monte-Carlo based approach*. *Water Science and Technology*, 76 (12), 3351-3357.
- [12] Tegos, A., Malamos, N., Efstratiadis, A., **Tsoukalas, I.**, Karanasios, A., and Koutsoyiannis, D. (2017). *Parametric Modelling of Potential Evapotranspiration: A Global Survey*. *Water*, 9(10), 795.
- [13] Makropoulos, C., Rozos, E., **Tsoukalas, I.**, Plevri, A., Karakatsanis, G., A., Karagiannidis, L., Makri, E., Lioumis, C., Noutsopoulos, C., Mamais, D., Rippis, C. and Lytras, E. (2017). *Sewer-Mining: A water reuse option supporting circular economy, public service provision and entrepreneurship*, *Journal of environmental management*. 216, 285-298.
- [14] Rozos E., **Tsoukalas I.**, Ripsis K., Smeti E., and Makropoulos, C., (2017) *Turning black into green: Ecosystem services from treated wastewater*, *Desalination and Water Treatment*.

## 2 Conference publications with full evaluation

### 2.1 Phd related

- [15] **Tsoukalas I.**, Efstratiadis A., and Makropoulos C., (2017). *Stochastic simulation of periodic processes with arbitrary marginal distributions*, Proceedings of the 15th International Conference on Environmental Science and Technology, Rhodes, Greece; 09/2017.
- [16] **Tsoukalas I.**, Michas S., and Makropoulos C., (2016). *A Monte-Carlo based method for the identification of potential sewer mining locations*, 13th IWA Specialized Conference on Small Water and Wastewater Systems, Athens, Greece.
- [17] **Tsoukalas I.**, Dimas P., and Makropoulos C., (2015). *Hydrosystem optimization on a budget: Investigating the potential of surrogate based optimization techniques*. Proceedings of the 14th International Conference on Environmental Science and Technology, Rhodes, Greece.

### 2.2 Other publications

- [18] Nikolopoulos, D., Makropoulos, C., Kalogeras, D., Monokrousou, K., and **Tsoukalas, I.** (2018). *Developing a stress-testing platform for cyber-physical water infrastructure*. In 2018 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater) (pp. 9-11). IEEE.
- [19] Psarrou E., **Tsoukalas I.**, and Makropoulos C., (2017). *A decision support tool for optimal placement of sewer mining units: Coupling SWMM 5.1 and Monte-Carlo simulation*, Proceedings of the 15th International Conference on Environmental Science and Technology, Rhodes, Greece; 09/2017.
- [20] Lykou A., Koutiva I., Karavokiros G., **Tsoukalas I.**, Pantazis C., Raspati G., Ugarelli R., Alves A., Sanchez Torres A., Vojinovic Z. and Makropoulos C. (2017). *The PEARL-toolbox: supporting the decision making process in selecting flood resilience strategies*. Proceedings of the 15th International Conference on Environmental Science and Technology, Rhodes, Greece; 09/2017.
- [21] Rozos E., **Tsoukalas I.**, and Makropoulos C., (2016). *Turning black into green: ecosystem services from treated wastewater*, 13th IWA Specialized Conference on Small Water and Wastewater Systems, Athens, Greece.

## 3 Conference publications and presentations with evaluation of abstract

### 3.1 Phd related

- [22] **Tsoukalas I.**, Kossieris P., Efstratiadis A., Makropoulos C., and Koutsoyiannis D., (2018). *CastaliaR: An R package for multivariate stochastic simulation at multiple temporal scales*, European Geosciences Union General Assembly 2018, Geophysical Research Abstracts, Vienna.
- [23] **Tsoukalas I.**, Kossieris P., Efstratiadis A., and Makropoulos C., (2015). *Handling time-expensive global optimization problems through the surrogate-enhanced evolutionary annealing-simplex algorithm*. European Geosciences Union General Assembly 2015, Geophysical Research Abstracts, Vienna.
- [24] Efstratiadis A., **Tsoukalas I.**, Kossieris P., Karavokiros G., Christofides A., Siskos A., Mamassis N., and Koutsoyiannis D., (2015). *Computational issues in complex water-energy optimization problems: Time scales, parameterizations, objectives and algorithms*. European Geosciences Union General Assembly 2015, Geophysical Research Abstracts, Vienna.

### 3.2 Other publications

- [25] Moustakis Y., Kossieris P., **Tsoukalas I.**, and Efstratiadis A., (2017). *Quasi-continuous stochastic simulation framework for flood modelling*. European Geosciences Union General Assembly 2017, Geophysical Research Abstracts, Vienna.
- [26] Kossieris P., Efstratiadis A., **Tsoukalas I.**, and Koutsoyiannis D., (2015). *Assessing the performance of Bartlett-Lewis model on the simulation of Athens rainfall*. European Geosciences Union General Assembly 2015, Geophysical Research Abstracts, Vienna.
- [27] Drosou, A., Dimitriadis P., Lykou A., Kossieris P., **Tsoukalas I.**, Efstratiadis A., and Mamassis N., (2015), *Assessing and optimising flood control options along the Arachthos river floodplain (Epirus, Greece)*. European Geosciences Union General Assembly 2015, Geophysical Research Abstracts, Vienna.

