Stochastics of Hydroclimatic Extremes

A Cool Look at Risk

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering, School of Civil Engineering
National Technical University of Athens

Edition 0 (for students)
Athens 2020

Stochastics of Hydroclimatic Extremes
A Cool Look at Risk

Edition 0 (for students)

Printed by the National Technical University of Athens

To my beloved wife Annouska
For standing by me and supporting me on hardship
For being my friend, comrade and spiritual guide
For criticizing me when she believes I am wrong
(Being herself wrong, of course)
For laughing with me and making me laugh

Contents

Foreword

by Zbigniew W. Kundzewicz (Polish Academy of Sciences)

(to appear in Edition 1)

τάν βροτοῖς δὲ πήματα
ἀκούσαθ', ὥς σφας νηπίους ὄντας τὸ πρὶν
ἔννους ἔθηκα καὶ φρενῶν ἐπηβόλους.
λέξω δέ, μέμψιν οὔτιν' ἀνθρώποις ἔχων,
ἀλλ' ὧν δέδωκ' εὔνοιαν ἐξηγούμενος·
οἳ πρῶτα μὲν βλέποντες ἔβλεπον μάτην,
κλύοντες οὐκ ἤκουον, ἀλλ' ὀνειράτων
ἀλίγκιοι μορφαῖσι τὸν μακρὸν βίον
ἔφυρον εἰκῇ πάντα, κοὔτε πλινθυφεῖς
δόμους προσείλους ἦσαν, οὐ ξυλουργίαν·
κατώρυχες δ' ἔναιον ὥστ' ἀήσυροι
μύρμηκες ἄντρων ἐν μυχοῖς ἀνηλίοις.
ἦν δ' οὐδὲν αὐτοῖς οὔτε χείματος τέκμαρ
οὔτ' ἀνθεμώδους ἦρος οὔτε καρπίμου
θέρους βέβαιον, ἀλλ' ἄτερ γνώμης τὸ πᾶν
ἔπρασσον, ἔστε δή σφιν ἀντολὰς ἐγὼ
ἄστρων ἔδειξα τάς τε δυσκρίτους δύσεις.
καὶ μὴν ἀριθμόν, ἔξοχον σοφισμάτων,
ἐξηῦρον αὐτοῖς, γραμμάτων τε
συνθέσεις,
μνήμην ἀπάντων, μουσομήτορ' ἐργάνην.
κἄζευξα πρῶτος ἐν ζυγοῖσι κνώδαλα
ζεύγλαισι δουλεύοντα σώμασίν θ' ὅπως
θνητοῖς μεγίστων διάδοχοι μοχθημάτων
γένοινθ', ὑφ' ἅρμα τ' ἤγαγον φιληνίους
ἵππους, ἄγαλμα τῆς ὑπερπλούτου χλιδῆς.
θαλασσόπλαγκτα δ' οὔτις ἄλλος ἀντ' ἐμοῦ
λινόπτερ' ἠῦρε ναυτίλων ὀχήματα.
τοιαῦτα μηχανήματ' ἐξευρὼν τάλας
βροτοῖσιν, αὐτὸς οὐκ ἔχω σόφισμ' ὅτῳ
τῆς νῦν παρούσης πημονῆς ἀπαλλαγῶ.

Αίσχύλος, Προμηθεὺς Δεσμώτης (442-471)*

But listen to the tale
Of human sufferings, and how at first
Senseless as beasts I gave men sense, possessed them
Of mind. I speak not in contempt of man;
I do but tell of good gifts I conferred.
In the beginning, seeing they saw amiss,
And hearing heard not, but, like phantoms huddled
In dreams, the perplexed story of their days
Confounded; knowing neither timber-work
Nor brick-built dwellings basking in the light,
But dug for themselves holes, wherein like ants,
That hardly may contend against a breath,
They dwelt in burrows of their unsunned caves.
Neither of winter's cold had they fix'd sign,
Nor of the spring when she comes decked with
flowers,
Nor yet of summer's heat with melting fruits
Sure token: but utterly without knowledge
Moiled, until I the rising of the stars
Showed them, and when they set, though much
obscure.
Moreover, number, the most excellent
Of all inventions, I for them devised,
And gave them writing that retaineth all,
The serviceable mother of the Muse.
I was the first that yoked unmanaged beasts,
To serve as slaves with collar and with pack,
And take upon themselves, to man's relief,
The heaviest labour of his hands: and I
Tamed to the rein and drove in wheelèd cars
The horse, of sumptuous pride the ornament.
And those sea-wanderers with the wings of cloth,
The shipman's waggons, none but me devised.
These manifold inventions for mankind
I perfected, who, out upon't, have none,—
No, not one shift—to rid me of this shame.

Aeschylus, Prometheus Bound (442-471),
Translated by G. M. Cookson†

* http://www.greek-language.gr/digitalResources/ancient_greek/library/browse.html?page=12&text_id=132

† https://en.wikisource.org/wiki/Four_Plays_of_Aeschylus_(Cookson)/Prometheus_Bound

Preface

A year ago, a flash flood killed 24 people in Mandra, a small town near Athens*. The losses are a result of lack of infrastructure for flood protection, while the natural stream network had been abused by urban development. If the storm had been predicted and if there were alert systems and evacuation plans, the consequences would not be that tragic. However, predictions for storms of small duration and extent, occurring at dry places, are difficult. This year some meteorologists in Greece (not the official meteorological service), perhaps envying the glory of American meteorologists who deal with storms of different type such as hurricanes—and at the same time facing the fact that in Greece there are no hurricanes—decided to give names to every meteorological depression entering Greece. The journalists received this initiative enthusiastically advertising the names in all media, while authorities started to react by closing schools in days of predicted (named) bad weather. At the very day I am writing these lines, the weather in Athens is wintry (as it should normally be in February). In the earlier days the meteorological predictions spoke about a "*historical snow event*" (ιστορικός χιονιάς) that has not happened before†. But it seems that, to forecasters' disappointment, this (named) historical snowfall did not happen even this time.

If meteorological predictions are difficult, especially those for a week after, what about climate predictions which are for really long time horizons? A few years ago, it was predicted that "*snowfalls are now just a thing of the past*"‡ Soon this prediction changed to the opposite one, "*Extreme snowfall is actually an expected consequence of a warmer world*".§ However, despite their variety, reaching self-contradiction, all these predictions have some things in common. For most people, they are scary. And in contrast to Cassandra's sorrowing prophesies, which were true but not believed by people, current prophesies of catastrophes usually are broadly believed but very often do not come true.

Apocalyptic prophesies have been common in history and were mostly connected—and owed their power—to religion. Modern ones are instead connected to the ideology of environmentalism and owe their power to scientists. However, they share several characteristics with the older prophesies, most prominently the people-scaring and world-saviour attitudes. Since 1970, several environmental scientists predicted lots of catastrophes, for which apparently God laughed and, as they did not come true, we too may laugh now**. I believe that bombarding people with negative predictions is detrimental to the society—and objectively it is contrary to the pretended world saviour attitude. It makes the society more vulnerable. This has been vividly expressed more than

---

* https://en.wikipedia.org/wiki/Mandra

† https://tvxs.gr/news/ellada/erxetai-istorikos-xionias

‡ https://web.archive.org/web/20150905124331/http://www.independent.co.uk/environment/snowfalls-are-now-just-a-thing-of-the-past-724017.html

§ http://www.bbc.com/earth/story/20160127-will-snow-become-a-thing-of-the-past-as-the-climate-warms

** Koutsoyiannis, D., Saving the world from climate threats vs. dispelling climate myths and fears, Invited Seminar, Lunz am See, Austria, doi:10.13140/RG.2.2.34278.42565, WasserCluster Lunz – Biologische Station GmbH, 2017.

2600 years ago in the Aesop's fable originally entitled "shepherd playing" («Ποιμήν παίζων»), better known in English as "the boy who cried wolf."* More than 2300 years ago, Epicurus pronounced science as the enemy of fear and of superstition. And a couple of centuries before Epicurus, other philosophers such as Plato and Aristotle clarified the meaning and the ethical value of science as the pursuit of the truth—pursuit that is not driven by political agendas and economic interests. For the latter, they used the term sophistry.

I believe what is needed is a cool look at risk. For risk exists—as it existed all the time in the past and will certainly exist in the future. Because the population has grown rapidly in the 20th century, increasing by an order of magnitude since 1800 and two orders of magnitude since the period of Plato and Aristotle, and becoming now a significant percentage (>7%) of the people that have ever lived on Earth, one would think that the risk, measured in terms of damages and human losses by natural hazards, has increased. This however is not the case. Thanks to substantial progress in engineering and technology the risk has decreased.†

Engineers' profession is tightly connected to risk. The infrastructure they build generally decreases risk from natural hazards but does not eliminate it. At the same time, infrastructure is subject to risk per se. The comedian and writer John Oliver gave it an interesting definition: "*Infrastructure: it's our roads, bridges, dams, levees, airports, power grids—basically anything that can be destroyed in an action movie.*"‡ Accordingly, the engineers' profession is socially sensitive and responsible at an enormous degree. Unlike Aesop's shepherd, an engineer cannot play with risk; the consequences, in case of a failure of infrastructure or its management, are not as ecologically friendly as wolves eating sheep.

Being an engineer, I have dealt with risk for decades. It is my intent to convey my experience to the readers of this book. Although I have published a lot of articles and made even more conference talks related to this subject, what is contained in the book is mostly new.

One important issue that I have consistently tried to communicate is my belief that the current standard methodologies underestimate substantially the probability of extreme events. I hope I have substantiated my claims in this book. The reasons of underestimation are basically two. The first is an inappropriate assumption of classical statistical methodologies: that the different events are independent of each other. They are not. This will repeatedly be illustrated in the book using long records of hydrometeorological processes, as well as invoking theoretical arguments. The second is the assumption that the distribution tails of those processes exhibit a rapid decay as we go to larger and larger values: an exponential descent, like in the exponential or even the normal distribution. The inappropriateness of both these assumptions has not been widely known because the

---

* https://en.wikisource.org/wiki/The_Shepherd%27s_Boy_and_the_Wolf; original Greek text: https://el.wikisource.org/wiki/Αισώπου_Μύθοι/Ποιμήν_παίζων.

† Related data are given in the last chapter of the book.

‡ Infrastructure: Last Week Tonight https://www.youtube.com/watch?v=Wpzvaqypav8.

relevant behaviours are hidden if the time series of observations are not long enough. And both assumptions are connected to each other and act synergistically to underestimate the probability of occurrence of extremes and hence the risk.

In particular, the independence assumption is virtually equivalent to a static climate. Accordingly, if we remove this assumption, we get a varying climate, which is consistent with the real-world climate. The latter statements may sound as counterintuitive or even wrong, because typically dependence is interpreted as *memory* rather than *change*. Nonetheless, the close relationship of dependence, particularly the long-range one, with change is illustrated in the book both empirically and theoretically. Given, on the one hand, the adherence to independence in typical studies of extremes and, on the other hand, the fact that independence entails a static climate, it is not a surprise to see that most recent studies try to remedy the consequences of the inappropriate assumptions by invoking climate change—or anthropogenic global warming, the global scapegoat of our era. Methods to embed climate change into studies dealing with occurrence probability vary, but all have several weaknesses—examples are provided in the book. I believe that just removing the independence assumption—and thus representing a changing climate without additional assumptions—resolves most of the underestimation problems.

The language used in this book is the language of stochastics. This may be inapprehensible at first glance but it is an effective language. The book tries to adhere on the rigorous use of stochastics, on the one hand, and to make its presentation both easy and self-contained, on the other hand. In this respect, the biggest part of the book is devoted to the theory of stochastics which is necessary for inferences about extremes. Stochastics is a scientific area broader than statistics—actually, according to the definition I adopt, statistics is part of stochastics. Another part is the theory of stochastic processes, in which *time* has a hypostasis that is typically absent in statistics. The direct analogy is dynamics vs. statics. This does not mean that statistics is underrepresented in the book. On the contrary, several new developments are presented—most notably the new tool of *knowable moments*, which have two relevant characteristics: they are closely connected to extremes and their estimation is unbiased in the framework of classical statistics or involves small bias in stochastic processes with dependence in time, whilst the bias in the estimation of classical statistical moments can be huge. As it will be seen in the book, knowable moments help to develop an extreme-oriented fitting methodology of probability distributions.

In parallel to being theoretical, the book is oriented to application. The new theoretical developments are supported by derivations and proofs, which to improve readability are contained in a number of Appendices in each chapter. The application is supported by several examples and illustrations, usually standing out as parenthetical sections or Digressions, as well as by tabulations of mathematical formulae that are used for each task.

Athens, 24 February 2019

Demetris Koutsoyiannis

Acknowledgments

The book is part of the unfunded project *"Seeking Theoretical Consistency in Analysis of Geophysical Data (Using Stochastics)*". Even though it is unfunded, my university, the National Technical University of Athens (NTUA), still provides my salary, albeit at a reduced rate by about 40% because of the Greek economic crisis, and I am thankful to it.

I am grateful to Zbyszek Kundzewicz (Polish Academy of Sciences), Harry Lins (World Meteorological Organization, Commission for Hydrology) and Alberto Montanari (University of Bologna and European Geosciences Union), all renowned scientists in the field, for agreeing to contribute with a Foreword or Afterword.

I thank my collaborators in the above project, Any Iliopoulou, Panayiotis Dimitriadis and Ioannis Tsoukalas from NTUA, and Federico Lombardo from the Sapienza University for their suggestions and corrections. I also thank my colleagues and friends George Karavokiros, Evangelos Rozos, Marina Pantazidou, Panagiotis Kossieris and Vassilios Zoukos, and my students Nikos Agatheris, Konstantinos Glynis, Faidon Diakomopoulos, Vasilis Kourakos, Nikos Tepetidis, Stelios Vavoulogiannis, who located errors in a draft of the book, made suggestions or tested some of the results in their theses or papers.

Even though I have discussed the content of this book with the people I thank and they suggested several corrections, it is certain that several errors remain, which is my responsibility, of course. Hopefully, I will be able to correct some more of them in a future edition.

I thank Rick Sanders for reminding me of Aeschylus and drawing my attention into the piece from Prometheus Bound that appears as a motto of the book.

I am grateful to the colleagues who have put online their huge data sets, as well as the data processing systems they developed. This has been the most important development in hydrology and geophysics since the time I entered academia several decades ago, to which I too have consistently tried to contribute in my country, with only partial success. In particular, I thank the Royal Netherlands Meteorological Institute (Koninklijk Nederlands Meteorologisch Instituut; KNMI) for their Climate Explorer (climexp) system, which provides access to (and processing of) a huge number of data sets. For the Uccle rainfall data, I thank the Royal Meteorological Institute of Belgium (RMIB) and Alexandre Dewalque for providing them at a 10-min time step and allowing their use for scientific or educational purposes.

Notational conventions

The book follows the *Guidelines for the use of units, symbols and equations in hydrology*[*]. In turn, these guidelines are based on (i) the Système International (SI) brochure[†]; (ii) the ISO 80000-2 Standard, *Mathematical Signs and Symbols to Be Used in the Natural Sciences and Technology*; and (iii) Unicode Technical Report #25, *Unicode Support for Mathematics*.[‡] We list some of the conventions here for the reader's convenience.

Physical dimensions and units

(a) All quantities are dimensionally consistent. In particular, arguments of functions such as exp( ) and ln( ) are dimensionless.
(b) We use s, min, h, and d for second, minute, hour and day respectively. We do not abbreviate week, month or year, which are non-SI units.[§]
(c) Multiplication of units is indicated by a space, e.g. N m, and division either by negative exponents (e.g. $\mathrm{m\,s^{-2}}$) or by use of the solidus (oblique line, e.g. $\mathrm{m/s^2}$); however repeated use of the solidus (e.g. m/s/s) is not permitted.
(d) Prefixes of units such as M (mega = 10^6) and μ (micro = 10^{-6}) have no space between (e.g. μs, MW). According to the SI, the prefix for kilo is lower case k (e.g. km—K is the symbol of the kelvin).
(e) For areas and volumes, we use $\mathrm{m^2}$ and $\mathrm{m^3}$; the hectare (ha) and the litre (L) are also allowed in SI. A million $\mathrm{m^2}$ is denoted as square kilometre ($1\ \mathrm{km^2} = 10^6\ \mathrm{m^2}$). A million $\mathrm{m^3}$ is denoted as cubic hectometre ($1\ \mathrm{hm^3} = 10^6\ \mathrm{m^3}$—not 1 $\mathrm{Mm^3}$ because $1\ \mathrm{Mm^3} = 10^{18}\ \mathrm{m^3}$; note that in SI any power to a unit applies also to the prefix); a billion $\mathrm{m^3}$ is denoted a cubic kilometre ($1\ \mathrm{km^3} = 10^9\ \mathrm{m^3}$).
(f) All units are typeset in upright (Roman) fonts, not italic or bold.
(g) Numerals are also typeset in upright fonts. The symbol for the decimal marker is the dot. To facilitate reading, numbers are divided in groups of three using a thin space (e.g. 12 345.6). (Note that neither dots nor commas are permitted as group separators). A space is used to separate the unit from the number (e.g. 10 m).

Symbols and equations

(a) We prefer single-letter variables (if necessary, with subscripts, e.g. E_{RMS}) over multi-letter ones. Single-letter variables or parameters and user-defined function symbols are italic (e.g. $x, Y, \beta, f(x)$). Multi-letter variables, if cannot be avoided, are typeset in upright, not italic (e.g. RMSE).
(b) Common, explicitly defined, functions are not italic, whether their symbols are single-letter (e.g. $\Gamma(x)$ for the gamma function, $\mathrm{B}(y, z)$ for the beta function) or multi-letter (e.g. $\ln x$, $\exp(x + y)$).
(c) Textual subscripts or superscripts are not italic (e.g. $x_{\max}$, $T_{\min}$ where 'max' and 'min' stand for maximum and minimum, respectively).
(d) Mathematical constants are upright (e.g. $\mathrm{e} = 2.718\ldots$, $\pi = 3.141\ldots$, $\mathrm{i}^2 = -1$). Also, mathematical operators are upright (e.g. $\mathrm{d}x$ in integrals and derivatives, $\Delta\gamma$ for the difference operator on γ).
(e) Vectors, matrices and vector functions are bold and, for single-letter variables, italic. In particular, vectors are usually denoted with lower case letters (e.g. $\boldsymbol{x}$, $\boldsymbol{\omega}$ as vectors; $\boldsymbol{f}(\boldsymbol{x})$ as a vector function of a vector variable) and matrices with upper case letters (e.g. $\boldsymbol{A}$ as matrix; $\boldsymbol{AB}$ as the product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A}^{\mathrm{T}}$ as the transpose of $\boldsymbol{A}$, $\det \boldsymbol{A}$ as the determinant of a square matrix $\boldsymbol{A}$).
(f) We use nested parentheses for grouping (e.g. $\ln(a\,(b + c))$ rather than $\ln[a\,(b + c)]$)
(g) To distinguish between stochastic variables from regular variables we use the Dutch convention[**], i.e., we underline the stochastic variables. Further, we use the curly brackets for sets (e.g. $P\{\underline{x} \le x\}$ for a scalar x or $P\{\underline{\boldsymbol{x}} \le \boldsymbol{x}\}$ for a vector $\boldsymbol{x}$; note that the argument of probability (P) is a set, not a number).
(h) We use square brackets for expectations, variances and other operators on stochastic variables (e.g. $\mathrm{E}[\underline{x}]$, $\mathrm{var}[\underline{x}]$, $\mathrm{cov}[\underline{x}, \underline{z}]$; note that $\mathrm{E}[\underline{x}]$ is not a function of $\underline{x}$ and thus it should not be denoted as $\mathrm{E}(\underline{x})$.)
(i) Definitions by mathematical equations are denoted using the symbols '$:=$' and '$=:$' (e.g. to define c as the sum of a and b we write $c := a + b$ or $a + b =: c$).

---

[*] Prepared by D. Koutsoyiannis and H.H.G. Savenije, 2013, doi: 10.13140/RG.2.2.10775.21922

[†] Ninth edition, http://www.bipm.org/en/si/si_brochure/

[‡] http://www.unicode.org/reports/tr25

[§] We avoid 'a' for year, because in SI 'a' is the prefix atto, meaning 10^{-18}; also it is the symbol of an 'are', a non-SI unit whose multiple hectare is accepted in SI ($1\ \mathrm{a} = 100\ \mathrm{m^2}$; $1\ \mathrm{ha} = 100\ \mathrm{a} = 10^4\ \mathrm{m^2} = 1\ \mathrm{hm^2}$).

[**] Hemelrijk, J., 1966. Underlining random variables. *Statistica Neerlandica*, 20(1), pp.1-7.

Main use of single-letter symbols

a	coefficients of stochastic generators	α	time scale parameter in stochastic processes
A		A	as Latin A
b		β	Background measure
B	(as a standard, the beta function B(,))	B	as Latin B
c	autocovariance	γ	climacogram (as a standard, Euler's constant, γ = 0.577216…)
C		Γ	cumulative climacogram (as a standard, the gamma function Γ( ) and the incomplete gamma function Γ_α( )).
d	(as a standard, the differential operation d)	δ	
D	time unit, discretization time step	Δ	
e	(as constant, e = 2.71828…)	ε	dimensionless location parameter in distributions
E		E	
f	probability density function	ζ	dimensionless shape parameter (lower-tail index) in distributions (as a standard, the Riemann zeta function ζ( ))
F	probability distribution function	Z	as Latin Z
g		η	time lag, dimensionless
G		H	as Latin H
h	time lag, dimensional	θ	angle (phase); also ombrian parameter
H	Hurst coefficient (also, $H_p \coloneqq \sum_{i=1}^{p} 1/i$ and $H_p^{(a)} \coloneqq \sum_{i=1}^{p} 1/i^a$ are the pth harmonic numbers of orders 1 and a, respectively.	Θ	Bias correction factor
i		ι	
I		I	as Latin I
j		κ	time scale, dimensionless (also cumulants)
J		K	as Latin K
k	time scale, dimensional	λ	state scale parameter in distributions
K	K-moment	Λ	Λ-coefficient
l		μ	mean, moment
L	Length of observation period	M	as Latin M
m	moment	ν	similar to n (size of sample or vector)
M	Mandelbrot coefficient	N	as Latin N
n	size of sample or vector	ξ	dimensionless shape parameter in distributions (tail index)
N	size of sample or vector	Ξ	
o		o	
O		O	as Latin O
p	moment order	π	(as constant, π = 3.14159…)
P	probability	Π	
q	moment order	ρ	standardized cross-climacogram
Q		P	as Latin P
r	correlation coefficient	σ	standard deviation
R		Σ	sum
s	power spectrum	τ	time, dimensionless
S		T	as Latin T
t	time, dimensional	υ	structure function
T	return period (as a superscript, 'T': transpose)	Y	as Latin Y
u		φ	entropy production
U		Φ	Entropy
v	white noise process	χ	
V		X	as Latin X
w	frequency	ψ	climacospectrum
W		Ψ	
x, y, z	stochastic variables and process or time series	ω	frequency, dimensionless
X, Y, Z	cumulative stochastic process or time series	Ω	

Chapter 1. An introduction by examples

1.1 General setting

We will start our journey to the hydroclimatic extremes with some illustrative examples. The purpose is to recognize the physical behaviours before we start discussing the mathematical and technical weaponry to tackle the problems about the risk related to the occurrence of extremes. In particular, by studying these examples we may understand how hard (perhaps infeasible) it is (and most probably will ever be) to deal with extremes using deterministic methods, while at the same time the theory of stochastics provides suitable means to quantify the extremes and the related risks. Generally speaking, deterministic approaches are popular as they are simple to understand and our education system is based on them, but methods offered by stochastics are much more powerful.

An interesting example of a spectacular failure of a deterministic approach in quantifying extremes is the notion of the so-called probable maximum precipitation, which is regarded to be an upper bound of precipitation that is physically feasible. Such an upper bound is philosophically and scientifically contradictory whilst the methods devised to determine it, even though they are thought to be physically-based deterministic methods, are in fact statistical methods using bad statistics. Therefore, we will not consider this type of approaches, while the reader interested to see the reasoning about the inconsistency of these methods is referenced to Koutsoyiannis (2007) and Koutsoyiannis and Papalexiou (2017).

But what is *stochastics*, the term appearing also in the title of the book? In the modern scientific vocabulary, it is used to collectively refer to (a) probability theory, (b) statistics and (c) stochastic processes. More loosely speaking, stochastics is the mathematics of stochastic variables and stochastic processes, which will formally be introduced in Chapter 2 and Chapter 3. However, the notion of stochastics, long before being implanted to the scientific vocabulary by Jacob Bernoulli, had originally appeared in ancient Greek philosophical texts. These appearances both enrich and elucidate the notion of stochastics and it is thus useful to trace back its different meanings through the history of philosophy and science. Relevant information is contained in Digression 1.A, which helps us to perceive the meaning of a stochastic approach, a rich meaning with several facets, including:

- probability theoretic;
- capable of quantifying the non-precise, the uncertain, or else the random;
- insightful—not superficial;
- aiming at prediction in a probabilistic sense, using information of the past;
- targeting at calculating the mean, or expectation, of uncertain quantities.

Naturally, once we have adopted a stochastic approach, we will deal with probabilities and expectations of extreme quantities, and our inferences will be based on past observations of the processes of interest. Thus, the examples below make use of the available information to make inferences of probabilistic type. But before we make

inferences of quantitative type, we need to (a) identify the most important characteristics of the process behaviour and (b) assume a model consistent with this behaviour. The examples discuss three types of models, namely (1) the classical probabilistic model according to which the different events are independent of each other, (2) a linear trend model and (3) a model assuming a certain type of dependence in time. By comparing these three models with the help of the examples, we will form a guide with directions that we should follow in studying extremes, which are also the directions to align the next chapters of this book.

Digression 1.A: The meaning of stochastics

Literally, *stochastics* is a term of Greek origin, stemming from the adjective '*stochasticos*' (στοχαστικός), or better its feminine gender, '*stochastice*' (στοχαστική). It is generated from the verb '*stochazesthai*' (στοχάζεσθαι), which in turn comes from the noun '*stochos*' (στόχος), meaning the target.

Aristotle, in his treatise *Nicomachean Ethics* (written ~350 BC) uses the term *stochastice* in its original meaning, related to the target, which, according to him, is the *mean*: "*virtue, therefore, is a balance* ['mesotes']*, in the sense that it is able to hit* [as a target – 'stochos'] *the mean*"[1]. Furthermore, in his treatise *Rhetoric* he uses the term with a metaphorical meaning, which could be translated into English as *guessing* or *guesswork*: "*men have a sufficient natural instinct for what is true, and usually do arrive at the truth. Hence the man who makes a good guess at truth is likely to make a good guess at probabilities* [stochastically]."[2]

However, it is Plato he who used the term with a meaning closer to the modern one, i.e., related to uncertainty. In his dialogue *Philebus* (written ~360 BC) he contrasts "*arithmetic and the sciences of measurement*" to *stochastics* and parallels the latter with music, which *"attains harmony by guesswork* […] *so that the amount of uncertainty mixed up in it is great, and the amount of certainty small."* [3]

The contrast between stochasticity and precision is made clear later by Galenus using the example of a city's clock: "*When a city is being built, let the following problem be set before those who will inhabit it: they want to expertly know, not stochastically but precisely, on an everyday basis, how much time has passed, and how much is left before sunset.*"[4]

The connection of stochastics with prediction or forecast becomes evident in an excerpt from Basilius Caesariensis who contrasts a prophet to a '*stochastes*' (στοχαστής, a noun usually translated into English as diviner): "*On the one hand, a prophet is he who foretells the future by revelation of the Spirit; on the other hand, a stochastes is he who infers the future by prudence, comparing similar states, and by the experience of forefathers.*"[5] It seems that this comment has influenced later scholars (e.g. Procopius) and perhaps determined the meaning of *stochastic* in modern Greek, which is imaginative, insightful, thoughtful, cogitative, contemplative, meditative.

The transplantation of *stochastics*, as an international scientific term, to the modern vocabulary is due to Jacob Bernoulli, evidently aware of the Greek language and literature, and in particular of the passage from Plato's Philebus mentioned above. In his famous book *Ars Conjectandi* (written in Latin in 1684-89 but published after his death, in 1713) he writes: "*To conjecture about something is to measure its probability. Therefore we define the art of conjecture, or stochastics, as the art of measuring the probabilities of things as exactly as possible, to the end that, in our judgments and actions, we may always choose or follow that which has been found to be better, more satisfactory, safer, or more carefully considered. On this alone turns all the wisdom of the philosopher and all the practical judgment of the statesman*."[6]

The term was revived by Bortkiewicz (1917; Russian economist and statistician of Polish ancestry) and also by Slutsky (1925, 1928a,b, 1929; Ukrainian/Russian/Soviet mathematical statistician and economist). It appears that the prevalence in USSR of the more sophisticated term *stochastic* (over the rather equivalent term *random*) must have been related to political and ideological reasons (incongruence or randomness with the *dialectic materialism*: models beyond strict deterministic were considered with a priori suspicion; see Mazliak 2018).

But it is Kolmogorov (1931) he who made the term popular and widespread, as he introduced the term *stochastic process* also clarifying that *process* means *change of a certain system*. Additionally, he used the term *stationary* to describe a probability density function that is unchanged in time (while at the same time the system state changes). Soon after, Kolmogorov (1933) introduced the modern and consistent definition of probability in an axiomatic manner, based on the measure theory (see section 2.1).

_____

[1] «*μεσότης τις ἄρα ἐστὶν ἡ ἀρετή, **στοχαστική** γε οὖσα τοῦ μέσου*» (Aristot. Nic. Eth. 1106b, translation into English adapted from that by H. Rackham. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. 1934). The notion of 'mesotes' (μεσότης), loosely translated as balance, middle, mean between a respective 'too much' and 'too little', is a key concept in Aristotle's ethical philosophy and thus to hit it as a target is important for him.

[2] «*οἱ ἄνθρωποι πρὸς τὸ ἀληθὲς πεφύκασιν ἱκανῶς καὶ τὰ πλείω τυγχάνουσι τῆς ἀληθείας: διὸ πρὸς τὰ ἔνδοξα **στοχαστικῶς** ἔχειν τοῦ ὁμοίως ἔχοντος καὶ πρὸς τὴν ἀλήθειάν ἐστιν*» (Aristot. Rh. 1.1, translation into English by W. Rhys Roberts, http://classics.mit.edu/Aristotle/rhetoric.1.i.html).

[3] The complete passage is: ΣΩΚΡΑΤΗΣ: «*οἷον πασῶν που τεχνῶν ἄν τις ἀριθμητικὴν χωρίζῃ καὶ μετρητικὴν καὶ στατικήν, ὡς ἔπος εἰπεῖν φαῦλον τὸ καταλειπόμενον ἑκάστης ἂν γίγνοιτο.* […] *τὸ γοῦν μετὰ ταῦτ᾽ εἰκάζειν λείποιτ᾽ ἂν καὶ τὰς αἰσθήσεις καταμελετᾶν ἐμπειρίᾳ καί τινι τριβῇ, ταῖς τῆς **στοχαστικῆς** προσχρωμένους δυνάμεσιν ἃς πολλοὶ τέχνας ἐπονομάζουσι, μελέτῃ καὶ πόνῳ τὴν ῥώμην ἀπειργασμένας.* […] *οὐκοῦν μεστὴ μέν που μουσικὴ πρῶτον, τὸ σύμφωνον ἁρμόττουσα οὐ μέτρῳ ἀλλὰ μελέτης **στοχασμῷ**, καὶ σύμπασα αὐτῆς αὐλητική, τὸ μέτρον ἑκάστης χορδῆς τῷ **στοχάζεσθαι** φερομένης θηρεύουσα, ὥστε πολὺ μεμειγμένον ἔχειν τὸ μὴ σαφές, σμικρὸν δὲ τὸ βέβαιον.*»

(SOCRATES: "*For example, if arithmetic and the sciences of measurement and weighing were taken away from all arts, what was left of any of them would be, so to speak, pretty worthless.* […] *All that would be left for us would be to conjecture and to drill the perceptions by practice and experience, with the additional use of the powers of guessing, which are commonly called arts and acquire their efficacy by practice and toil. […] Take music first; it is full of this; it attains harmony by guesswork based on practice, not by measurement; and flute music throughout tries to find the pitch of each note as it is produced by guess, so that the amount of uncertainty mixed up in it is great, and the amount of certainty small*" (Plat. Phileb. 55e, translation by Harold N. Fowler; Cambridge, MA, Harvard University Press; 1925.)

[4] «*πόλεως κτιζομένης προκείσθω τοῖς οἰκήσουσιν αὐτὴν ἐπίστασθαι βούλεσθαι, μὴ **στοχαστικῶς** ἀλλ᾽ ἀκριβῶς, ἑφ᾽ ἑκάστης ἡμέρας, ὁπόσον τε παρελήλυθεν ἤδη τοῦ χρόνου τοῦ κατ᾽ αὐτήν, ὁπόσον θ᾽ ὑπόλοιπόν ἐστιν ἄχρι δύσεως ἡλίου.*» (Γαληνοῦ Περὶ Διαγνώσεως καὶ Θεραπείας τῶν ἐν τῇ ἑκάστου Ψυχῇ Ἁμαρτημάτων — De Dignotione et Curatione cujusque Animi Peccatorum, 80, http://www.poesialatina.it/_ns/greek/testi/Galenus/De_animi_cuiuslibet_peccatorum_dignotione_et_curatione.html).

[5] «*Οὐκοῦν Προφήτης μέν ἐστιν, ὁ κατὰ ἀποκάλυψιν τοῦ Πνεύματος προαγορεύων τὸ μέλλον· **στοχαστὴς** δὲ, ὁ διὰ σύνεσιν ἐκ τῆς τοῦ ὁμοίου παραθέσεως, διὰ τὴν πεῖραν τῶν προλαβόντων, τὸ μέλλον συντεκμαιρόμενος.*» (Basilius, Ερμηνεία εις τον προφήτην Ησαΐαν —Enarratio in prophetam Isaiam, 3.102.26).

[6] "*Conjicere rem aliquam est metiri illius probabilitatem: ideoque Ars Conjectandi sive **Stochastice** nobis definitur ars metiendi quàm fieri potest exactissimè probabilitates rerum, eo fine, ut in judiciis & actionibus nostris semper eligere vel sequi possimus id, quod melius, satius, tutius aut consultius fuerit deprehensum; in quo solo omnis Philosophi sapientia & Politici prudential versatur*" (Bernoulli, 1713).

1.2 Introductory notes on the examples

The examples that follow make use for some of the longest available records of hydroclimatic observations. Only long records reveal the secrets of hydroclimate and its behaviours, which seem peculiar as they are very different from our perception of "random events". As we will see with the help of the examples:

1. While classical probability and statistics adhere to the assumption that different events are independent, this assumption is totally inappropriate when dealing

with hydroclimatic processes—and most other natural and artificial processes. An illustration of the difference is provided in Digression 1.B.

2. Popular "modern" approaches, such as those discovering "nonstationarity" are even more inappropriate. Models of this type identify mostly linear trends everywhere, trying to rectify, in an inappropriate and inefficient manner, the disagreement between natural behaviours and those resulting from the independence assumption.
3. Less popular approaches assuming dependence of events in time, in particular the type of dependence known as long-range dependence or persistence, can provide consistent quantification of extremes and the uncertainty thereof, which turns out to be much higher than captured by the other two alternatives.

One may think that an approach leading to high uncertainty is unsuccessful and, in this respect, models of type 2 are advantageous. Indeed, such approaches have been promoted as physically based and consistent with the popular anthropogenic global warming literature and with the industry of climate models and their predictions (or projections). If climate model information was really incorporated in the stochastic model and if this information was consistent with reality, then, indeed, the resulting nonstationary model, in which the trend was derived by a deterministic model, would be a progress. However, climate model outputs in their original form (without cosmetic reformations known as "bias correction" or "downscaling") are irrelevant to reality (Koutsoyiannis et al., 2008; Anagnostopoulos et al., 2010) particularly if we focus on extremes (Tsaknias et al., 2016). Attempts to incorporate climate model information within a stochastic framework in a consistent manner (Tyralis and Koutsoyiannis, 2017) lead to increased uncertainty or, in the best case, in indifferent results. For these reasons, we will not consider approaches based on climate model outputs in this book.

A necessary note about the examples which follow is that they do not refer to the details of the marginal distribution of extremes. Certainly, this is quite an important issue that will be studied in subsequent chapters—and of course there is a large body of publications that study it. However, it is equally important to study the variation of the occurrence of extremes in time, which actually is the focus of the examples. This problem, which severely influences modelling of hydroclimatic risk and decisions related to it, has not been given the deserved attention in the literature, or has been dealt with using naïve methods.

Digression 1.B Practical difference of dependence and independence

We assume that, based on observational data of river discharge, we have concluded that the probability of the event that the mean daily discharge at a certain location of a river exceeds 500 m^3/s is small, equal to 10^{-3}. Practically, this means that this event happens on the average one every 1000 days or once every 2.74 years. What is the probability that this event occurs for five consecutive days?

Even though we have not yet defined what independence formally is (this will be done in section 2.5 and 3.5 and Digression 3.B), we intuitively know that the probability of independent events occurring all together equals the product of the probabilities of the separate events. Thus, under independence, the probability sought is simply $(10^{-3})^5 = 10^{-15}$. This is an extremely low

probability: it means that we have to wait *on the average* 10^{15} days or 2.74 trillion years, or about 200 times the age of the universe, to see this event happen. However, such events (successive occurrences of extreme events for multiday periods) have been observed in several historical samples. This indicates that the independence assumption is not a justified assumption and yields erroneous results. Thus, we should avoid such an assumption if our target is to estimate probabilities for periods longer than the reference period. Methodologies admitting dependence, i.e. based on the theory of stochastic processes, are more appropriate for such problems and will result in probabilities much greater than 10^{-15}; these will be described in next chapters.

Now let us assume that for four successive days our extreme event has already occurred, i.e. that the mean daily river discharge was higher than 500 m^3/s in all four days. What is the probability that this event will also occur in the fifth day?

Many people, based on an unrefined intuition, may answer that the occurrence of the event already for four days will decrease the probability of another consecutive occurrence, and would be inclined to give an answer in between 10^{-3} and 10^{-15}. This is totally incorrect. If we assumed independence, then the correct answer would precisely be 10^{-3}; the past does not influence the future. If we assume positive dependence, which is a more correct assumption for natural processes, then the probability sought becomes higher, not lower, than 10^{-3}; it becomes more likely that a flood day will be followed by another flood day.

As we will see in next chapters, similar things happen if we move from the daily scale of the above example to the annual scale, or even larger. For example, if several warm winters have occurred in a series, then the probability that the next winter would be also warm is increased—not decreased. Ignorance of this simple truth may have severe consequences for those who aspire to predict the future and those who believe their predictions. A didactic historical example is the failed prediction of Hitler's meteorologist Franz Baur about the 1941-42 winter in Russia, which marked the Battle of Moscow. Quoting a fascinating paper by Neumann and Flohn (1987):

> *Baur was requested by the headquarters of the German Air Force to distribute his long-range forecasts to about 25 military offices. A forecast for winter 1941-42 was issued by him, probably at the end of October 1941, based on regional climatology and (supposed) sun-spot-climate relationships. The prediction called for a normal or a mild winter. Baur's main justification for this rested with the assertion that never in climatic history did more than two severe winters occur in a row. Since both of the preceding two winters, 1939-40 and 1940-41, were severe in Europe, he did not expect that the forthcoming winter would also be severe.*

However, that winter, in which the first major Soviet counteroffensive of the war was committed, turned out to be one of the coldest in record:

> *The cold outbreak of early December, coming after a cool to-cold October and November [...] gravely hit the German armies that were not appropriately clothed (Hitler expected to break the resistance of the USSR before the coming of winter) and which were not equipped with armaments, tanks, and motorized vehicles that could properly function even in a "normal" winter in the northern parts of the USSR, let alone in a winter as rigorous as that of 1941-42.*
> *On or about 8 December, K. Diesing, chief of the CWG and scientific adviser to the chief of the Weather Service of the Air Force (General Spang), asked Flohn to listen in on a second earphone to a telephone call to Baur. In the call, Diesing cited to Baur the reports of very low temperatures in the East and asked him if he maintains his seasonal forecast in face of the reports. Baur's response was "the observations must be wrong".*

1.3 Precipitation and its extremes as seen in a long record

Extreme behaviour in precipitation causes floods and droughts and therefore its study is very useful. Notably, even when flow records exist, rainfall probability has still a major role in engineering design; for instance, in major hydraulic structures, the design floods are generally estimated from appropriately synthesized design storms, which are rare extreme-rainfall events (e.g. U.S. Department of the Interior, 1987).

Therefore, we start our exploration of extremes with precipitation. In our example we study one of the longest daily rainfall records worldwide, that of Bologna, Italy (44.50°N, 11.35°E, 53.0 m). The time series of observations is available online in the frame of the Global Historical Climatology Network – Daily (GHCN-Daily; Menne et al., 2012)*. It is uninterrupted for the period 1813-2007, 195 years in total. For the most recent period, 2008-2018 daily data are provided by the online data repository Dext3r of ARPA Emilia Romagna, Rete di monitoraggio RIRER.† With these additional data, the record length becomes 206 years. The analyses that follow are based on the GHCN 195-year data set, while the most recent 11-year data are used for validation purposes.

Figure 1.1 depicts the daily time series as well the (right-aligned) moving averages and moving maxima for a time window of 10 years, representing the 10-year climatic values (for clarification of the meaning of *climatic* in our context see Digression 1.C). The most spectacular behaviour shown in the figure is the changing climate: The 10-year climatic average daily rainfall has been changing between a minimum of 1.2 mm (having occurred in the 1820s) and a maximum of 2.5 mm (having occurred at the decade ending in 1902)—more than twice the minimum. At the same time the 10-year climatic value of the maximum daily rainfall has varied between a minimum of 48.5 mm (having occurred in the 1820s) and a maximum of 155.7 mm (having occurred in the 1930s)—more than three times the minimum. These changes do not follow a linear pattern but have the form of long-term non-periodic fluctuations, up and down. In the most recent years, after 1950, there is a roughly increasing trend in both climatic indices, but such increasing trends were also observed before 1900, followed by drops thereafter.

A popular approach to deal with such changing patterns is to assume linear trends; publications adopting this approach abound (see Iliopoulou and Koutsoyiannis, 2020). A linear trend is a presumably a deterministic model (even though we use the data to fit it), as it describes the change of the mean of the process as a deterministic linear function of time. Here it is pertinent to recall the good practice of fitting deterministic models to data, which is typical for hydrological modelling, albeit commonly overlooked in fitting trend models. This practice is the so-called *split sample* testing, in which the available record is split into two segments one of which is used for calibration and the other for validation, as emphatically suggested by Klemes (1986).

We have applied the split-sample technique to the annual values of some indices extracted from the Bologna rainfall record. These are:

- the annual total precipitation, i.e., the sum of daily precipitation from all (wet) days of each year;

* GHCN Version 3; data retrieved on 2019-02-17 from https://climexp.knmi.nl/gdcnprcp.cgi?WMO=ITE00100550.

† Data retrieved on 2019-02-17 from http://www.smr.arpa.emr.it/dext3r/. In particular, the data from the station Bologna Idrografico (coordinates 44.499883°N, 11.346156°E, 84.0 m, practically the same as those given for the GHCN station (except a 31 m difference in the elevation, perhaps indicating that this particular station is located at the roof of a building), were used except for year 2008 for which no data are provided for this station. For this year, as well as for very few days with missing values in other years, the daily precipitation values of the station Bologna Urbana (44.500754°N, 11.328789°E, 78.0 m) were used instead.

- the annual maximum daily precipitation, i.e., the greatest of all daily rainfall depths over the (wet) days of a specific year;
- the probability dry, i.e., the ratio of the days with zero precipitation to the total number of a year's days (365 or 366); and
- the annual average wet-day precipitation, i.e., the ratio of the annual total precipitation to the number of wet days.

The annual maximum daily rainfall is related to the generation of floods. At the other end of extremes, as the annual minimum daily rainfall does not vary but it is always zero, an index of extreme behaviour is the probability dry, related to occurrence of droughts.

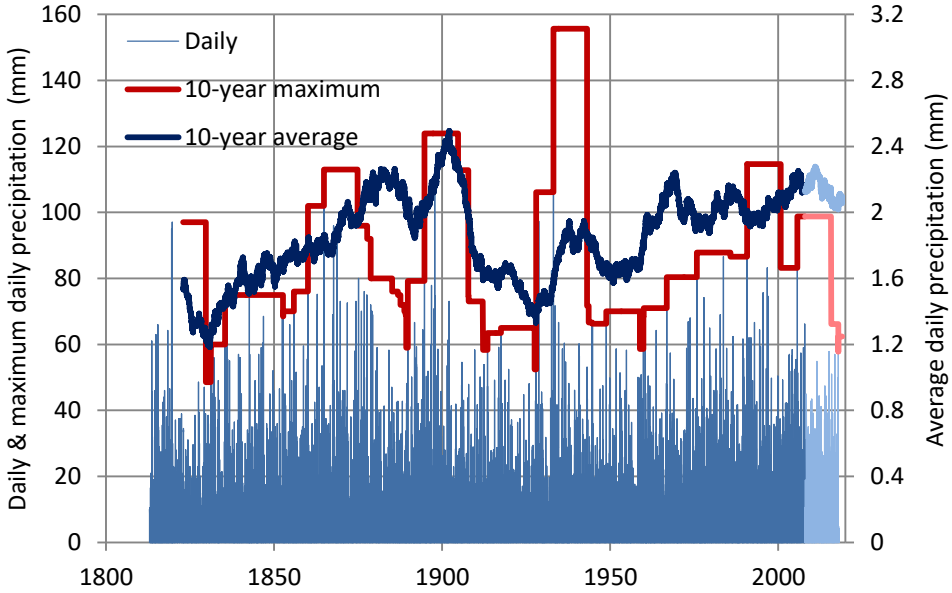

Figure 1.1 Plot of the time series of daily rainfall in Bologna, along with moving averages and moving maxima for a time window of 10 years (right-aligned, i.e., the value plotted at a specific year is the average or the maximum of the previous 10 years). The lines in darker colour represent the GHCN time series while those in lighter colour represent the newer data which are not included in the GHCN time series.

Plots of these annual indices are shown in Figure 1.2 along with fitted trends. Using the split-sample technique, we fitted the linear model on the mean of each index on the most recent part of the GHCN time series, namely the period 1950-2007. The linear trend model is

$$\mu(t) = a + bt \tag{1.1}$$

where μ is the mean of each process (index as a function of time), t is time and a and b are the parameters fitted by the standard linear regression method. As the simplest possible alternative, the constant mean model was used, i.e.,

$$\mu(t) = a = \text{constant} \tag{1.2}$$

(not shown in the graph).

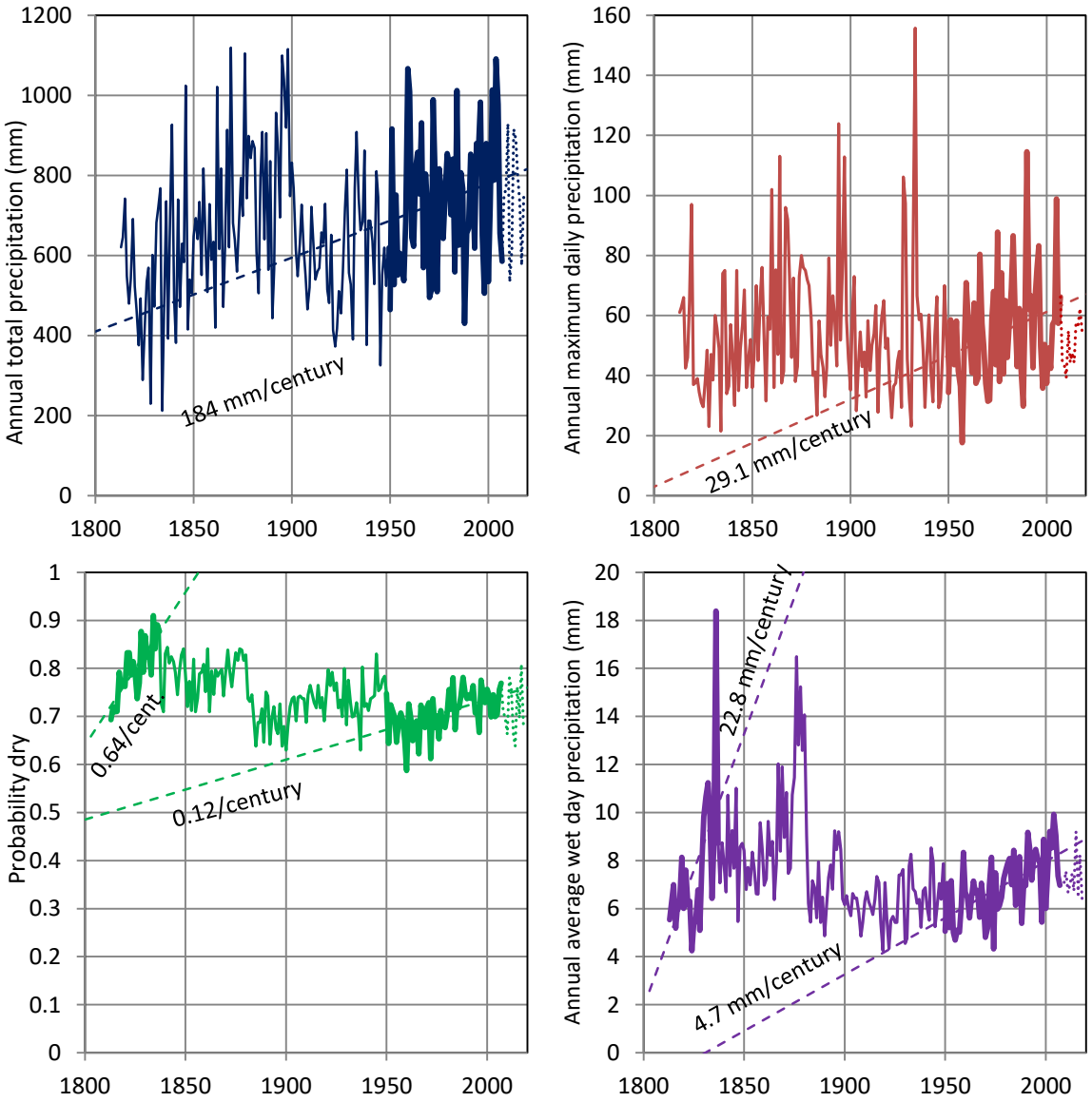

Figure 1.2 Plots of annual indices related to the (daily) rainfall process, namely annual total precipitation, annual maximum daily precipitation, probability dry and annual average wet day precipitation, with trends fitted on the most recent part of the GHCN time series, namely the period 1950-2007, for which the graphs are plotted with thicker lines. For the plots of the bottom row, namely the probability dry and the annual average wet day precipitation, trends are also plotted for the earliest 25-year period, 1813-1837. The newer data that are not included in the GHCN time series are plotted with dotted lines.

Two validation periods were used, namely the earlier period 1813-1949 of the GHCN time series, and the next period with the newer data of 2008-18, not contained in the GHCN time series. The comparison of the two models for each of the two validation periods is made in Table 1.1 in terms of the root mean square error in each case, defined as

$$E_{\mathrm{RMS}} := \sqrt{\frac{1}{n}\sum_{\tau=1}^{n}(x_\tau - \mu_\tau)^2} \tag{1.3}$$

where x_τ denotes the τth item of the observed time series and $\mu_\tau = \mu(\tau)$. Clearly, the comparison shows that the simpler, constant-mean model outperforms the linear model in all cases and in both validation periods. The inferior performance of the linear model is also seen visually in Figure 1.2. Therefore, we have no good reason to choose the linear-trend model.

Table 1.1 Root mean square error for the two validation periods and the two models, linear trend and constant mean, fitted to the calibration period (1950-2007).

	Annual total precipitation (mm)	Annual maximum daily precipitation (mm)	Probability dry (-)	Annual average wet-day precipitation (mm)
Validation period 1813-1949				
Assuming linear trend	206.9	36.8	0.194	6.12
Assuming constant mean	204.0	21.8	0.076	2.38
Validation period 2008-2018				
Assuming linear trend	138.3	16.3	0.064	1.54
Assuming constant mean	127.7	8.7	0.053	0.85

Actually, there are additional reasons why not to choose the latter, even if its performance was good. These are related to the poor (if any) logical background in using *time per se* as an exploratory variable in a natural process, as well as in fundamental concepts of stochastics, namely stationarity and ergodicity, which despite being fundamental are widely misunderstood. These concepts will be discussed in Chapter 3, while the reasons for excluding such models (including the exceptions in which such models are valid) are discussed elsewhere (Koutsoyiannis, 2011a; Montanari and Koutsoyiannis, 2014; Koutsoyiannis and Montanari, 2015). And even assuming that there were no theoretical obstacles and inferior performance, again we might adopt the constant mean model because of its parsimony (Iliopoulou and Koutsoyiannis, 2020). Specifically, philosophical reasoning (principle of parsimony, also known as Occam's razor) and practical considerations (model uncertainty) suggest preferring the more parsimonious model (Gauch, 2003). Quantification of such comparison of the model, which is not given here, is routinely done within stochastics—cf. the Akaike (1973, 1974) criterion and the Bayesian information criterion (Schwarz, 1978), as well as Serinaldi and Kilsby (2015), Serinaldi et al. (2018), and Iliopoulou and Koutsoyiannis (2020).

But even without these theoretical reasons, one can easily understand the absurdity of the linear-trend model by examining Figure 1.2. For example, if we assumed that the record of measurements did not go that back in time and we had adopted the linear-trend model for the annual maximum daily precipitation, we would conclude that about 1800 there was no intense rainfall at all, and that in the 18th century the precipitation was negative.

To go further with this example, let as make a thought experiment and assume that in the beginning of the 19th century there lived in the area three scientists, Drs A, B and C. Dr A kept records of the dry days of each year and Dr B observed the storm severity. In the 1830s, Dr A cast the prediction that rainfall will totally cease by 1850. In contrast Dr B

said that storms become more severe and by 1850 the storm activity will be tripled at least. Then came Dr C who reconciled the two theories stating that dry becomes drier and wet becomes wetter, and that the storms are much more severe while the regular rainfall events are becoming more and more rare, and will soon disappear. Now if we see again Figure 1.2, in particular the bottom panels where trends are also plotted for the 25-year period 1813-1837, we will understand that these claims would stand, if we were ready to accept the trend model as a decent one. Fortunately, however, scientists of our modern epoch do not use such naïve approaches to make groundless predictions, particularly of catastrophic or even apocalyptic events*.

Now having rejected the widespread practice of fitting linear trends, the question is, can we think a better alternative? Apparently, the answer is positive within stochastics. It would be a big failure otherwise, because the behaviour seen in our rainfall example is not a peculiarity of rainfall, nor one of Bologna. It is quite common everywhere and the main reasons we do not see it are two: (a) we do not have long enough records and (b) we are misled by the fact that we learnt probability by examples such are idealized (not even real) dice and roulettes.

In an idealized die the probabilities of the six possible outcomes are the same all the time, regardless of what happened in the previous throws. Macroscopically this simple system undergoes no change at all. That is, if we take the moving average of very many outcomes, we will have a flat line. In real-world processes the situation is different. There is change all the time and over all scales. Also, all events depend on each other. Dependence and change are closely related. We will see this relation later on (Digression 3.B). For now, we may take a note that dependence should not be interpreted as memory, as typically seen in literature, but as change. In particular, long-range dependence is not long-term memory but long-term change.

How is change quantified in stochastics? A simple way would be to describe some of the statistical characteristics as deterministic functions of time, but this is neither so effective nor rational, as we have seen in this Bologna rainfall example. Another option is to make this quantification in a stochastic, rather than deterministic, manner. In this case we view the change as variability across different time scales. In turn, the variability is quantified in terms of the variance.

Referring again to the annual time series of rainfall indices of Bologna for the entire 206-year period, which we denote as $x_1, x_2, \dots, x_{206}$, we take the following steps:

- We calculate the variance estimate $\hat{\gamma}(1)$, where '1' indicates the time scale of 1 year, as:

$$\hat{\gamma}(1) := \frac{(x_1 - \hat{\mu})^2 + \cdots + (x_n - \hat{\mu})^2}{n-1}, \qquad \hat{\mu} := \frac{x_1 + \cdots + x_n}{n} \tag{1.4}$$

* By the way, by examining the frequency of word usage in books with the help of Google's Ngram Viewer (https://books.google.com/ngrams/), we see that the word 'catastrophic' was practically not used in the 19th century but its frequency per million of words increased linearly in the 20th century to reach a recent peak value of 3.6. The usage of the word 'apocalyptic' peaked in 1808 with a frequency of 2.8 and again more recently in 1995 with a frequency of 3.6 (Koutsoyiannis, 2013b).

where $\hat{\mu}$ is the estimate of the mean and n = 206 is the sample size. The notion of estimate will be clarified later on, in Chapter 4.

- We form a time series at time scale 2 (years) by averaging pairs of consecutive items of the time series, i.e.:

$$x_1^{(2)} := \frac{x_1 + x_2}{2}, x_2^{(2)} := \frac{x_3 + x_4}{2}, \ldots, x_{103}^{(2)} := \frac{x_{205} + x_{206}}{2} \tag{1.5}$$

  and we calculate the estimate of the variance $\hat{\gamma}(2)$ in a similar manner.
- We repeat the same procedure to form time series at time scales 3, 4, …, up to scale 20 (1/10 of the record length) and calculate the variances $\hat{\gamma}(3), \hat{\gamma}(4), \ldots \hat{\gamma}(20)$.
- We plot (in double logarithmic axes) the variance $\hat{\gamma}(\kappa)$ as a function of time scale κ.

The function of the variance vs. time scale is called the *climacogram*[*] (Koutsoyiannis, 2010). If we have assumed a model for our process and we determine the variance, $\gamma(\kappa)$, from the model, we have the *theoretical climacogram*. If we estimate the variance, $\hat{\gamma}(\kappa)$, from a time series, then we have the empirical climacogram. Notably, if we have produced the times series from a model, the empirical climacogram will not necessarily coincide with the theoretical, because there is estimation bias. To make them coincide, we must subtract the bias from the theoretical climacogram. This is not difficult because, once we know the model, the bias is readily determined from that model by a simple and explicit relationship (see section 4.6).

Now, if the time series x_τ represented the so-called *white noise*, i.e., a pure random process, in which all events are independent of each other, the double logarithmic plot of the climacogram would be a straight line with slope –1; the proof is very easy (see equation (3.49)). In real-world processes, the slope is different from –1, designated as $2H - 2$, where H is the so-called Hurst parameter which takes on values in the interval (0, 1). We will see later on (section 3.7) that H is identical to the *entropy production in logarithmic time*. The case where this slope is constant for all time scales, corresponds to a simple scaling behaviour (e.g. Koutsoyiannis, 2006b), or the power law:

$$\gamma(\kappa) = \frac{\gamma(1)}{\kappa^{2-2H}} \tag{1.6}$$

which defines the *Hurst-Kolmogorov* (HK) process, a name giving credit to Hurst (1951), who was the first to discover this behaviour in natural processes, and to Kolmogorov (1940) who was the first to introduce the process as a mathematical object.

---

[*] The term *climacogram*, from the Greek *κλιμακόγραμμα*, deriving from *κλίμαξ* (climax = scale, as well as ladder; pl. *κλίμακες*) and *γράμμα* (gramma = written, drawn), was coined in Koutsoyiannis (2010) and could be translated in English as *scale(o)gram*, but the latter term is used for another concept. *Climacogram* should not be confused to *climatogram* which has another meaning related to the climatic regime of temperature and precipitation at a site or area. The latter term, deriving from the noun *κλίμα* (clima, originally meaning slope; pl. *κλίματα*), was first used in the Hellenistic period by the astronomer Hipparchus to describe climate (in relationship to the slope of the sun's rays) and is different from the other derivative noun *κλίμαξ*. Interestingly though, both *κλίμαξ* and *κλίμα* are eventually etymologized from the same verb *κλίνειν* (klinein = to slope).

It is easily seen that the value $H = 1/2$ corresponds to white noise as the slope is –1. High values of H (> 1/2) indicate *enhanced change* at large scales, else known as *long-term persistence*, or strong *clustering* (grouping) of similar values. This is quite common in natural processes. Low values H (< 1/2) indicate quite a different behaviour, called *antipersistence*. This is often confused with a periodic behaviour and hence called quasi-periodic (because the period of fluctuations is not constant). Such behaviour is much less frequent in hydroclimatic processes.

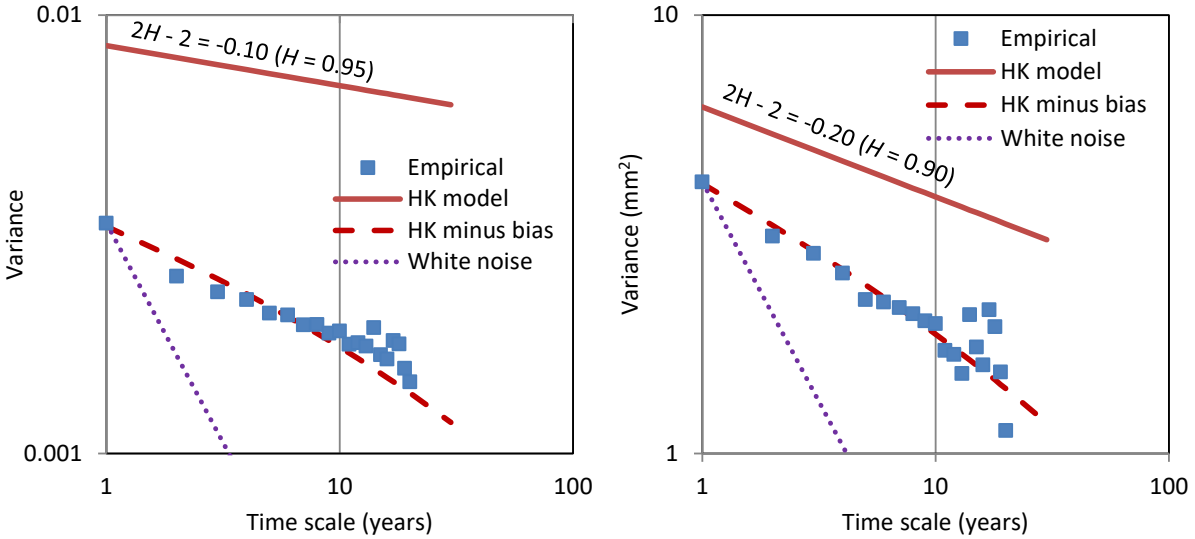

Figure 1.3 Empirical and theoretical climacograms of annual indices of daily rainfall at Bologna: **(left)** probability dry; **(right)** annual average wet-day precipitation.

Now we apply this method on the annual indices of daily Bologna rainfall. Figure 1.3 depicts the climacograms of the probability dry and the annual average wet-day precipitation. In both cases, the observed behaviour is spectacularly different from white noise while the Hurst-Kolmogorov behaviour is evident with Hurst parameter H as high as 0.95 for the probability dry and 0.90 for the wet-day precipitation. The situation is somewhat more complex for the annual total rainfall (not shown in Figure 1.3), in which the slope is different for small and large scales, an effect already known and analysed in Markonis and Koutsoyiannis (2016). The slope for large scales again suggests a strongly persistent behaviour with Hurst parameter $H = 0.86$. The annual maxima series tend to hide the Hurst behaviour, as explained in Iliopoulou and Koutsoyiannis (2019) and indeed the estimated H in this case is much smaller, ~0.60 (again not shown in Figure 1.3).

The Bologna precipitation example, as well as those that follow and many others, help shape a classification of change shown in the hierarchical chart of Figure 1.4. In simple systems (left part of the graph) the change is regular, either periodic or aperiodic. Regular change in simple cases is predictable in deterministic terms, using equations of dynamical systems. But this type of change is rather trivial. More interesting are the more complex systems at long time horizons (right part of the graph), where change is unpredictable in deterministic terms, or random. Pure randomness, like in classical statistics, where different variables are identically distributed and independent, is a useful model for idealized dice experiments, but in most natural systems it is inadequate. A structured randomness, like in the HK process, should be assumed instead. The structured

randomness is enhanced randomness, expressing enhanced unpredictability of enhanced multi-scale change.

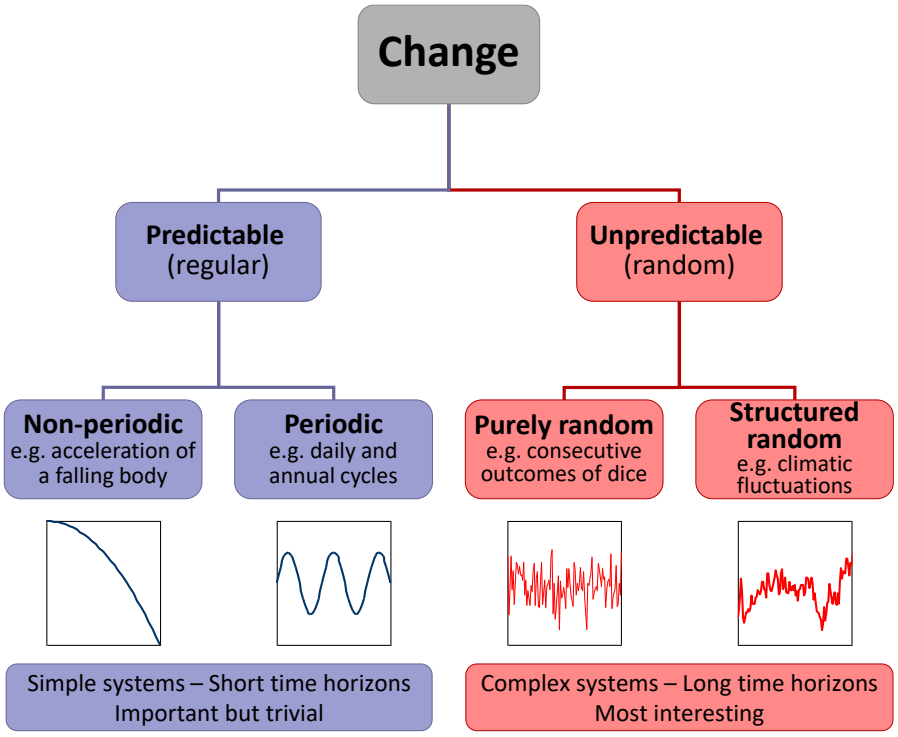

Figure 1.4 Classification of change (from Koutsoyiannis 2013b).

Digression 1.C: What is climate?

As happens with stochastics (Digression 1.A) the concept of climate is an old one. Aristotle in his *Meteorologica* describes the climates on Earth in connection with latitude but he uses a different term, *crasis* (*κρᾶσις*[1], literally meaning mixing, blending of things which form a compound, temperament).[2] The term *climate* (*κλίμα*, plural *κλίματα*) was coined as a geographical term by the astronomer Hipparchus[3] (190 –120 BC). He was the founder of trigonometry but is most famous for his discovery and calculation of *precession of the equinoxes* (μετάπτωσις ίσημεριών) by studying measurements on several stars. In the 20th century, this precession would be found to be related to the climate of Earth and constitutes one of the so-called *Milankovitch cycles*. The term *climate* originates from the verb *κλίνειν*, meaning 'to incline' and originally denoted the angle of inclination of the celestial sphere and the terrestrial latitude characterized by this angle (Shcheglov, 2007).

Hipparchus's *Table of Climates* is described by Strabo the Geographer (63 BC – AD 24), from whom it becomes clear that the *Climata* of that Table are just latitudes of several cities, from 16° to 58°N (see Shcheglov, 2007, for a reconstruction of the Table). However, Strabo himself uses the term climate with a meaning close to the modern one.[4] Furthermore Strabo, defined the five climatic zones, *torrid*, *temperate* and *frigid*, as we use them to date.[5]

The term climate was used with the ancient Greek geographical meaning until at least 1700 as imprinted in a dictionary of that era.[6] In contemporary times, a search on old books[7] reveals that the term *climatology* appears after 1800. With the increasing collection of meteorological measurements, the term climate acquires a statistical character as the average weather. Indeed, the geographer A.J. Herbertson (1907) in his book entitled "*Outlines of Physiography, an Introduction to the Study of the Earth*", gave the following definition of climate, based on, but also distinguishing it from, weather:

> *By climate we mean the average weather as ascertained by many years' observations. Climate also takes into account the extreme weather experienced during that period. Climate is what on an average we may expect, weather is what we actually get.*[8]

Herbertson also defined climatic regions of the world based on statistics of temperature and rainfall distribution, a work that was influential for the famous and most widely used Köppen (1918) climate classification; this includes six main zones and eleven climates which are on the same general scale as Herbertson's (Stamp, 1957). Herbertson's definition is kept virtually without essential changes till now; for example, Lamb (1972) states:

> *Climate is the sum total of the weather experienced at a place in the course of the year and over the years. It comprises not only those conditions that can obviously 'near average' or 'normal' but also the extremes and all the variations.*

Modern scientific glossaries also provide similar definitions of climate. We quote a few:

- By the USA National Weather Service:
  Climate – The composite or generally prevailing weather conditions of a region, throughout the year, averaged over a series of years.[9]
- By the Climate Prediction Center of the latter:
  Climate – The average of weather over at least a 30-year period. Note that the climate taken over different periods of time (30 years, 1000 years) may be different. The old saying is climate is what we expect and weather is what we get.[10]
- By the American Meteorological Society[11],
  Climate – The slowly varying aspects of the atmosphere–hydrosphere–land surface system. It is typically characterized in terms of suitable averages of the climate system over periods of a month or more, taking into consideration the variability in time of these averaged quantities. Climatic classifications include the spatial variation of these time-averaged variables. Beginning with the view of local climate as little more than the annual course of long-term averages of surface temperature and precipitation, the concept of climate has broadened and evolved in recent decades in response to the increased understanding of the underlying processes that determine climate and its variability.

  In turn, the climate system is defined as:
  The system, consisting of the atmosphere, hydrosphere, lithosphere, and biosphere, determining the earth's climate as the result of mutual interactions and responses to external influences (forcing). Physical, chemical, and biological processes are involved in the interactions among the components of the climate system.
- By the WMO (1992):
  C0850 climate – Synthesis of weather conditions in a given area, characterized by long-term statistics (mean values, variances, probabilities of extreme values, etc.) of the meteorological elements in that area.
  C0900 climate system – System consisting of the atmosphere, the hydrosphere (comprising the liquid water distributed on and beneath the Earth's surface, as well as the cryosphere, i.e. the snow and ice on and beneath the surface), the surface lithosphere (comprising the rock, soil and sediment of the Earth's surface), and the biosphere (comprising Earth's plant and animal life and man), which, under the effects of the solar radiation received by the Earth, determines the climate of the Earth. Although climate essentially relates to the varying states of the atmosphere only, the other parts of the climate system also have a significant role in forming climate, through their interactions with the atmosphere.
- By the IPCC (2013b):
  Climate – Climate in a narrow sense is usually defined as the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or millions of years. The classical period for averaging these variables is 30 years, as defined by the World Meteorological Organization. The

relevant quantities are most often surface variables such as temperature, precipitation and wind. Climate in a wider sense is the state, including a statistical description, of the climate system.

A useful observation is that all definitions use the term "average" (an exception is the definition by Lamb who uses the loose term *sum total* with the same meaning). Thus, by its definition, climate is a statistical concept. And since climate is not static but dynamic, it is better to think of it as a stochastic concept.

By scrutinizing the definitions, several questions may arise. A first is: Why "*at least a 30-year period*"? Is there anything special with the 30 years? Probably this reflects a historical belief that 30 years are enough to smooth out "random" weather components and establish a constant mean. In turn, this reflects a perception of a constant climate—and a hope that 30 years would be enough for a climatic quantity to get stabilized to a constant value. It can be conjectured that the number 30 stems from the central limit theorem (see section 2.17) and in particular the common (but not quite right) belief that the sampling distribution of the mean is normal for sample sizes over 30 (e.g. Hoffman, 2015). Such a perception roughly harmonizes with classical statistics of independent events. This perception is further reflected in the term *anomaly* (from the Greek *ανωμαλία*, meaning abnormality), commonly used in climatology to express the difference from the mean. Thus, the dominant idea is that a constant climate would be the norm and a deviation from the norm would be an abnormality, perhaps caused by an external agent. However, such belief is incorrect. The examples given in this chapter support the idea of an ever changing climate.

A second question inspired by Climate Prediction Center's definition is: Why the climate taken over 30 or 1000 years is different? The obvious reply is: Because different 30-year periods have different climate. This contradicts the tacit belief of constancy and harmonizes with the perception of an ever-changing climate. With the latter perception, Herbertson's idea (which the Climate Prediction Center refers to as an "old saying") that "*climate is what we expect, weather is what we get*" can be reformulated as "*weather is what we get immediately, climate is what we get if you keep expecting for a long time*" (Koutsoyiannis, 2011a).

As many of the above definitions refer to *weather*, it is useful to clarify its meaning, noting that it represents a popular notion, often used with respect to its effects upon life and human activities, rather than a rigour scientific one. Interestingly, in its colloquial use in Greek and Romance (Neo-Latin) languages, *weather* is almost indistinguishable from *time* (Greek: καιρός; Italian: tempo; French: temps, météo; Spanish: tiempo, clima; Portuguese: tempo, clima). On the other hand, in English and Greek, weather refers to short-scale variations in the atmosphere and is distinguished from climate; note however that in colloquial Spanish and Portuguese there is no such distinction. In scientific terms, the definition given by the WMO (1992) is this:

W0410 weather – State of the atmosphere at a particular time, as defined by the various meteorological elements.

Based on the above discussion, here we attempt to give a definition of climate, which is used in this book, in a hierarchical manner (avoiding circular logic) starting from the concept of climate system, as follows:

- *Climate system* is the system consisting of the atmosphere, the hydrosphere (including its solid phase—the cryosphere), the lithosphere and the biosphere, which mutually interact and respond to external influences and particularly those determining the solar radiation reaching the Earth, such as the solar activity, the Earth's motion and the volcano activity.
- *Climatic processes* are the physical, chemical and biological processes, which are produced by the interactions and responses of the climate system components through flows of energy and mass, and chemical and biological reactions.
- *Climate* is a collection of climatic processes at a specified area, stochastically characterized for a range of time scales.

According to this latter definition—and given that the term *process* means *change* (Kolmogorov, 1931), climate is changing by definition. Thus, there is no need to define or use the term *climate change*; actually, the latter term, which appeared in literature only after the 1970s,

serves non-scientific purposes (Koutsoyiannis, 2020c,d). Change occurs at all scales (Koutsoyiannis, 2013b), and there is nothing particular in any specific one, like the commonly assumed 30-year scale. By studying long observation series of atmospheric and hydrological processes, one would see that the only characteristic scale with clear physical meaning is the annual—beyond that there is no objective "border scale" that would support a different definition of climate. The above definition includes all scales beyond the annual, thus leaving out the smaller scales (e.g. of several minutes or days) to be associated to weather.

The stochastic characterization, appearing in the definition of climate, includes all statistics used in other definitions, such as averages, variability, extremes, etc., and collectively encompasses all related concepts of the scientific areas of probability, statistics and stochastic processes.

The main distinction between weather and climate is this. While weather, according to its definition by WMO (1992) which is kept unchanged here, refers to a particular time, climate refers to the entire climatic process, throughout all times.

As stated in the WMO (1992) definition of climate quoted above, the typical use of the term climate relates to the atmosphere only, leaving out the other parts of the climate system. However, since the climatic system includes the hydrosphere, there is no reason to exclude the hydrological processes from the climatic processes. Therefore, our definition includes them. Nevertheless, to give more emphasis on the inclusion of hydrological processes, the term *hydroclimatic* has been used even in the title of the book. This gives additional clarity or emphasis but it is rather a pleonasm as the hydrosphere is already included in the climate system.

_____

[1] The same root has the modern Greek word *κρασί* for *wine*. Yet the term is still in use today in Greek for derivative names related to climate such as *εύκρατος* (well-tempered, temperate) and *ευκρασία* (eucracy).

[2] [Aristot. Mete., 362b.17] «*…ὅ τε γὰρ λόγος δείκνυσιν ὅτι ἐπὶ πλάτος μὲν* [τὴν οἰκουμένην] *ὥρισται, τὸ δὲ κύκλῳ συνάπτειν ἐνδέχεται διὰ τὴν* ***κρᾶσιν****, -ού γὰρ ὑπερβάλλει τὰ καύματα καὶ τὸ ψῦχος κατὰ μῆκος, ἀλλ' ἐπὶ πλάτος, ὥστ' εἰ μή που κωλύει θαλάττης πλῆθος, ἅπαν εἶναι πορεύσιμον, —καὶ κατὰ τὰ φαινόμενα περί τε τοὺς πλοῦς καὶ τὰς πορείας·*»

"*… theoretical calculation shows that* [inhabited Earth] *is limited in breadth, but could as far as climate is concerned, extend round the Earth in a continuous belt; for it is not difference of longitude but of latitude that brings great variation of temperature, and if were not for the ocean which prevent it, the complete the complete circuit could be made. And the facts known to us from journeys by sea and land also confirm the conclusion…*" (English translation by H.D.P. Lee, Harvard University Press, Cambridge, Mass. USA, 1952).

[3] In his Commentary on Aratus (*Ιππάρχου των Αράτου και Ευδόξου φαινομένων εξηγήσεως*; Shcheglov, 2007).

[4] [Strab. 1.1] «*πάντες, ὅσοι τόπων ἰδιότητας λέγειν ἐπιχειροῦσιν, οἰκείως προσάπτονται καὶ τῶν οὐρανίων καὶ γεωμετρίας, σχήματα καὶ μεγέθη καὶ ἀποστήματα καὶ* ***κλίματα*** *δηλοῦντες καὶ θάλπη καὶ ψύχη καὶ ἁπλῶς τὴν τοῦ περιέχοντος φύσιν.*»

"*Every one who undertakes to give an accurate description of a place, should be particular to add its astronomical and geometrical relations, explaining carefully its extent, distance, degrees of latitude, and 'climate'—the heat, cold, and temperature of the atmosphere.*" (English translation by H.C. Hamilton, and W. Falconer, M.A., 1903)

[5] [Strab. 2.3] «*αὕτη δὲ τῷ εἰς τὰς [πέντε] ζώνας μερισμῷ λαμβάνει τὴν οἰκείαν διάκρισιν: αἵ τε γὰρ κατεψυγμέναι δύο τὴν ἔλλειψιν τοῦ θάλπους ὑπαγορεύουσιν εἰς μίαν τοῦ περιέχοντος φύσιν συναγόμεναι, αἵ τε* ***εὔκρατοι*** *παραπλησίως εἰς μίαν τὴν μεσότητα ἄγονται, εἰς δὲ τὴν λοιπὴν ἡ λοιπὴ μία καὶ διακεκαυμένη.*»

"*In the division into* [five] *zones, each of these is correctly distinguished. The two frigid zones indicate the want of heat, being alike in the temperature of their atmosphere; the temperate zones possess a moderate heat, and the remaining, or torrid zone, is remarkable for its excess of heat.*" (English translation by H.C. Hamilton, and W. Falconer, M.A., 1903). Notice the use of the Aristotelian *crasis* (*κρᾶσις*) in the term *εὔκρατοι* (*temperate*) zones.

[6] The following definition appears in Moxon (1700): "*Climate, From the Greek word Clima. of the same signification; it is a portion of the Earth or Heaven contained between two Parallels. And for distinction of Places, and different temperature of the Air, according to their situation; the whole Globe of Earth is divided into 24 Northern, and 24 Southern Climates, according to the half-hourly encreasing of the longest days; for under the Equator we call the first Climate: from thence as far as the Latitude extends, under which the longest*

day is half an hour more than under the Equator, viz. 12 hours and an half, is the second Climate: where it is encreased a whole hour, the third Climate: and so each Northerly and Southerly Climate respectively hath its longest day half an hour longer than the former Climate, till in the last Climate North and South, the Sun Sets not for half a year together, but moves Circularly above the Horizon."

[7] https://books.google.com/ngrams/graph?content=climatology.

[8] Thus Herbertson appears to be the father of the famous quotation "*climate is what we expect, weather is what we get*", often attributed to Marc Twain. What Twain has actually written, attributing it to an anonymous student, is "*Climate lasts all the time and weather only a few days*"; see https://quoteinvestigator.com/2012/06/24/climate-vs-weather/.

[9] https://w1.weather.gov/glossary/index.php?letter=c

[10] https://www.cpc.ncep.noaa.gov/products/outreach/glossary.shtml#C

[11] http://glossary.ametsoc.org/wiki/Climate

1.4 Temperature and its extremes as seen in a long record

Next, we study temperature data of the same site, Bologna, Italy (coordinates same as in the GHCN station above), again one of the longest temperature records worldwide, which has been thoroughly studied for that reason. The time series of average daily temperature is available online in the frame of the European Climate Assessment & Dataset (ECAD; Klein Tank et al., 2002).* It is uninterrupted for the period 1814-2003, 190 years in total. For the most recent period, 2004-2018, daily data are provided by the online data repository Dext3r, described above.† With these additional data, the record length becomes 205 years. The analyses that follow were based on the ECAD 190-year data set, while the most recent data were used for validation purposes. Additional time series for earlier periods that go back to 1715 have been compiled and made available online by Camuffo et al. (2017a,b), but they were not used in this study except as a background information.

Figure 1.5 shows plots of the time series of daily temperature, along with moving averages and moving maxima and minima for a time window of 10 years (right-aligned), representing the 10-year climatic values. We may first observe that the temperature has varied from −13 to 34.2 °C, a range of 47.2 °C, which would be much higher than 50 °C if we also considered the diurnal variation. The minimum value of −13 °C occurred on January 1830 and the maximum of 34.2 °C on August 2017. This latter value is thus not contained in the ECAD time series, whose maximum is 33.8 °C, occurring on August 1947. If we focus on the 10-year climatic values we will see again change, which however is small compared to the 47.2 °C range. Specifically, the 10-year climatic average daily temperature has been changing between 12.6 °C (for the 10-year period ending in 1861) and 15.6 (for 2007). At the same time, the 10-year climatic value of the maximum daily temperature has varied between 29.6 °C (for 1904) and 34.2 °C (in 2016 or 33.8 °C in

* Data retrieved on 2019-02-17 from https://climexp.knmi.nl/ecatemp.cgi?WMO=169.

† In particular, the average daily temperature values of the station Bologna Urbana (44.500754°N, 11.328789°E, 78.0 m) were used (note that no temperature data are provided for Bologna Idrografico, which was used for rainfall). The data at Bologna Urbana were adjusted by adding a constant temperature difference of 0.19 °C to become consistent with those of the ECAD station. To find this adjustment, as there is no common period of observation between the ECAD station and Bologna Urbana, a third station whose observations have common periods with both, namely the Bologna Meteo station (44.501223°N, 11.328197°E, 80.0 m) was used.

1947). Finally, the 10-year climatic value of the minimum daily temperature has varied between –13 °C (for 1830) and –2.4 °C (in 2007 or –3.8 °C in 1917).

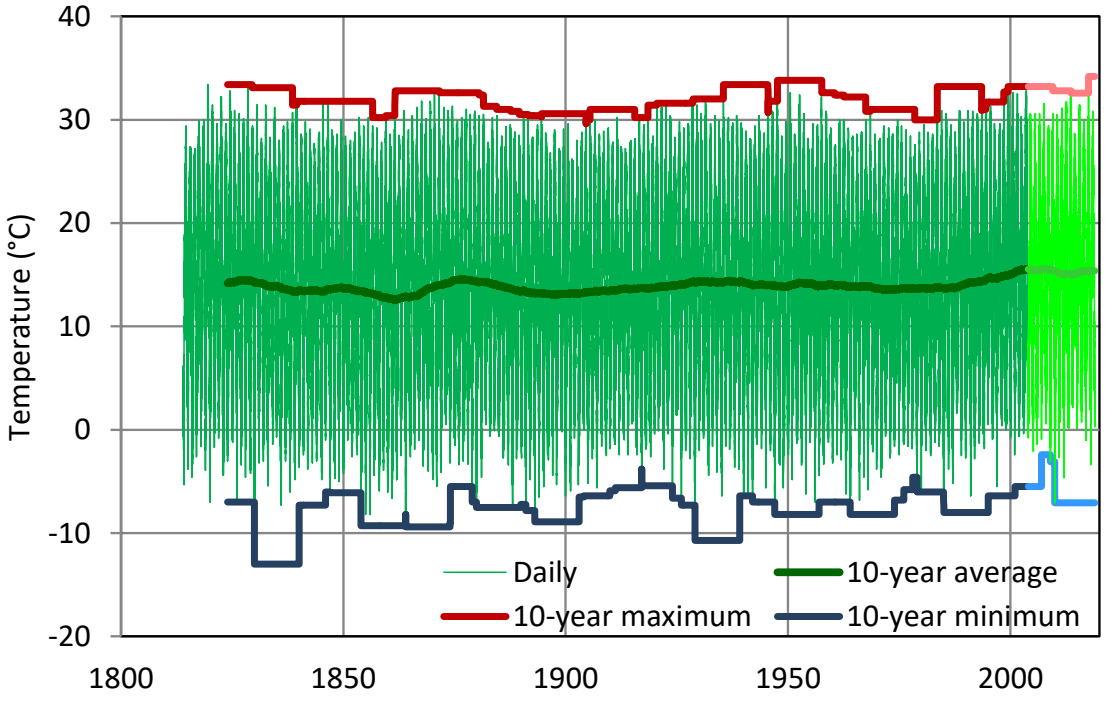

Figure 1.5 Plot of the time series of daily temperature in Bologna, along with moving averages and moving maxima and minima for a time window of 10 years (right-aligned). The lines in darker colour represent ECAD time series while those the lighter colour in the most recent years represent the newer data which are not included in the ECAD time series.

As in precipitation, the climatic changes of temperature do not follow a linear pattern but have the form of long-term non-periodic fluctuations, up and down. After 1970 the trends are increasing for average, maximum and minimum temperatures, but such increasing trends were also observed in other periods (most prominently after 1900), lasting several decades and followed by drops thereafter. As shown in Figure 1.6, the recent trends for the 35-year period 1969-2003 are very intense. Interestingly, by examining graphs of mean annual temperature for earlier periods, before 1814, published in Camuffo et al. (2017a,b), we note that there was an equally (or even more) intense increasing trend between 1740 and 1780, preceded by an even more rapid decreasing trend from 1720 to 1740. Thus, the minimum temperature in the last 300 years was observed in 1740.

However, if we follow the split-sample logic exposed in detail in section 1.3) we will reject the linear-trend model. Even the visual information in Figure 1.6 suffices to realize its bad performance for the early period, as well as the more recent period, after 2003. Furthermore, Figure 1.7 tells the same story as in precipitation (section 1.3): The Hurst behaviour is evident, with a Hurst parameter $H = 0.94$ for the annual average temperature and $H = 0.74$ for the annual maximum daily temperature.

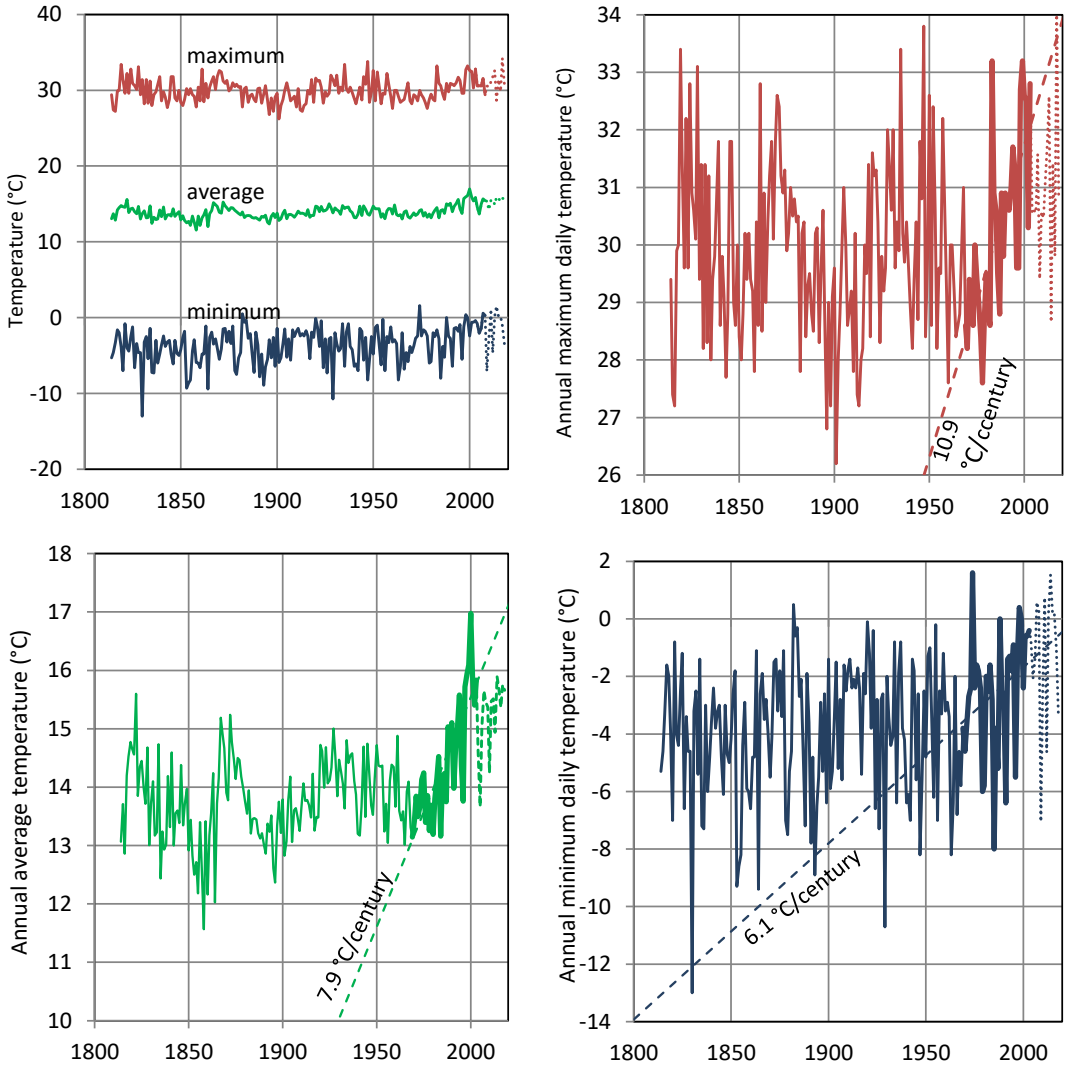

Figure 1.6 Plots of annual average, maximum and minimum daily temperature in Bologna, with trends fitted on the most recent 35-year part of the ECAD time series representing the most warming period 1969-2003, for which the graphs are plotted with thicker lines. The newer data that are not included in the ECAD time series are plotted with dotted lines.

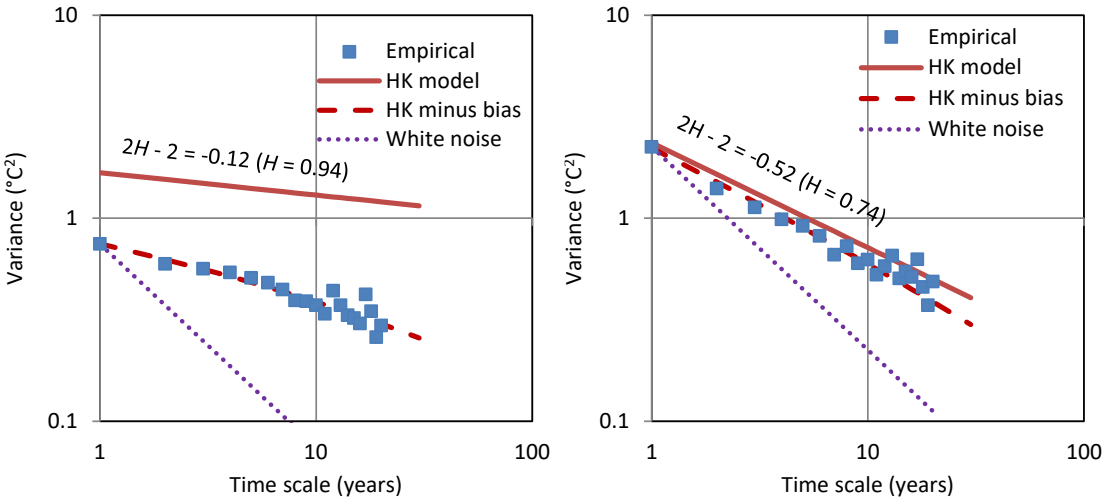

Figure 1.7 Empirical and theoretical climacograms of annual indices of daily temperature at Bologna: **(left)** annual average; **(right)** annual maximum daily.

1.5 A severe drought in a historical context

Discussions about droughts have been intense in the 21st century, triggered by climate change fears as well as by the severity of some droughts that have occurred: in Australia (2001-09), California (2011–17; Griffin and Anchukaitis, 2014) and Europe (2003, 2015; Hanel et al., 2018). Nonetheless, even though the 21st-century droughts in Europe have been broadly regarded as exceptionally severe, the Hanel et al. (2018) study shows that they were much milder in severity and areal extent in comparison to many older extensive drought events in Europe.

About a decade before these droughts, a prolonged and severe one hit Greece. It particularly influenced the Athens water supply system and shook Athenians. Despite that, the resulting water crisis is not as famous as the current economic crisis in Greece. Certainly, the reason for not being famous is the very successful management of the water crisis, in contrast to the economic crisis. Indeed, the entire campaign to handle the drought in Athens was very successful and despite the long (7-year) duration (1988-95) and severity of the drought there was neither one day of system failure (cf. Koutsoyiannis, 2011a); all inhabitants had water in their tap all the time.

Here we will study the hydrological conditions behind this water crisis using streamflow data for one of the major three catchments that supply water to Athens, namely, the Boeoticos Kephisos River at the Karditsa station (close to the outlet to Karditsa tunnel; catchment area 1930 km^2). The monthly runoff time series we use (compiled by Koutsoyiannis et al., 2007 and updated by Makropoulos et al., 2018 and Efstratiadis et al., 2019), is the longest streamflow time series in Greece, beginning in 1907 and uninterrupted since then (112 years up to 2016-19; note that the convention of a hydrological year is used, from October of previous year to September of the current year). In contrast to floods whose study requires high temporal resolution data, the monthly time scale is more than sufficient for studying droughts.

The 112-year monthly series of river discharge is shown in Figure 1.8, along with the 10-year moving average (right-aligned; left panel), as well as a linear trend fitted to the latest 50-year period before the beginning of the drought, i.e., the period 1937-87. It is seen in the left panel of the figure that, after the drought period, the climatic value of streamflow recovered (increased), but not to the level that was before the 1980s. The trend model would predict that the falling trend would continue.

Comparison of the two models introduced in section 1.3, the linear-trend model and the constant mean model, is given in Table 1.2 for two validation periods, before and after the calibration period. The constant-mean performs better. Furthermore, if, in spite that, we preferred the trend model and if we plan for a period of, say, 50 years in the future, we must think what we will do as we approach the end of the planning period. For extrapolation of the trend will give negative streamflow at 2060, forty years from now. This is similar to the early trend discussed in section 1.3, according to which the probability dry in Bologna would become 1 just after 1850. Therefore, it is again better not to trust the linear trend model. Later on (section 4.10), we will discuss how to make

better future predictions for a specified prediction horizon with the constant-mean model along with Hurst-Kolmogorov dynamics.

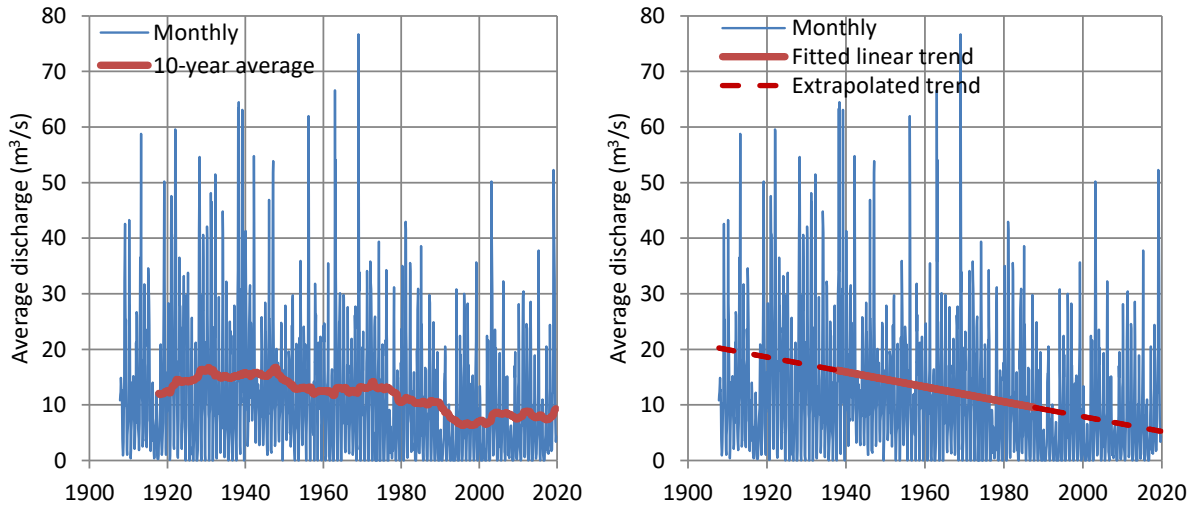

Figure 1.8 Plots of the time series of monthly average discharge of Boeoticos Kephisos, with **(left)** 10-year moving averages (right-aligned) and **(right)** trend fitted to the period 1937-87 (the 50-year period before the beginning of the drought).

Table 1.2 Root mean square errors (in m³/s) for the two validation periods for the linear-trend model and the constant-mean model, fitted to the calibration period (1937-87).

Validation period	1907-37	1987-2019
Assuming linear trend	13.4	12.7
Assuming constant mean	9.3	10.3

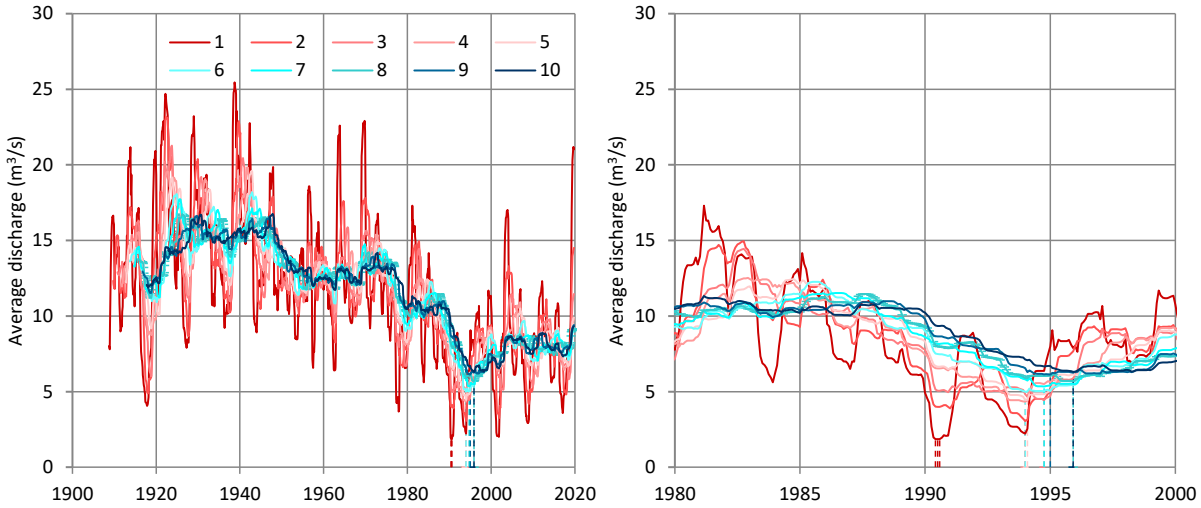

Figure 1.9 (left) Plot of the (right-aligned) moving average of the Boeoticos Kephisos discharge for the time scales noted in the legend; the time locations of the observed minima at each scale are also shown with dashed lines with the same colour as the corresponding moving-average time series. **(right)** Close up of the left panel for years 1980-2000.

It is useful to study in more detail the drought period. In contrast to a flash flood, a drought is not a rapid event but its evolution usually extends over many years. To characterize that evolution stochastically, we may use a multi-scale representation of the

time series, as we did to define the climacogram. Figure 1.9 shows such a representation at scales ranging from 1 to 10 years. The difference from the definition of the climacogram is that the values plotted in Figure 1.9 are constructed for a sliding window of length equal to the time scale, while in the standard definition of the climacogram the time windows are fixed in position. It is seen in the plots of the time series that the minima for all time scales for the entire period of observations are concentrated at that particular drought period. This is a characteristic of the HK behaviour; had the series been produced by a white noise model, that clustering would be quite improbable.

Indeed, the climacogram plotted in Figure 1.10 suggests Hurst behaviour of the process with Hurst parameter $H = 0.82$. Again, the difference from white noise is substantial. This difference is further illustrated in the right panel of Figure 1.10 in which the return periods of the lowest and highest observed average discharge over time scale 1 to 10 years.

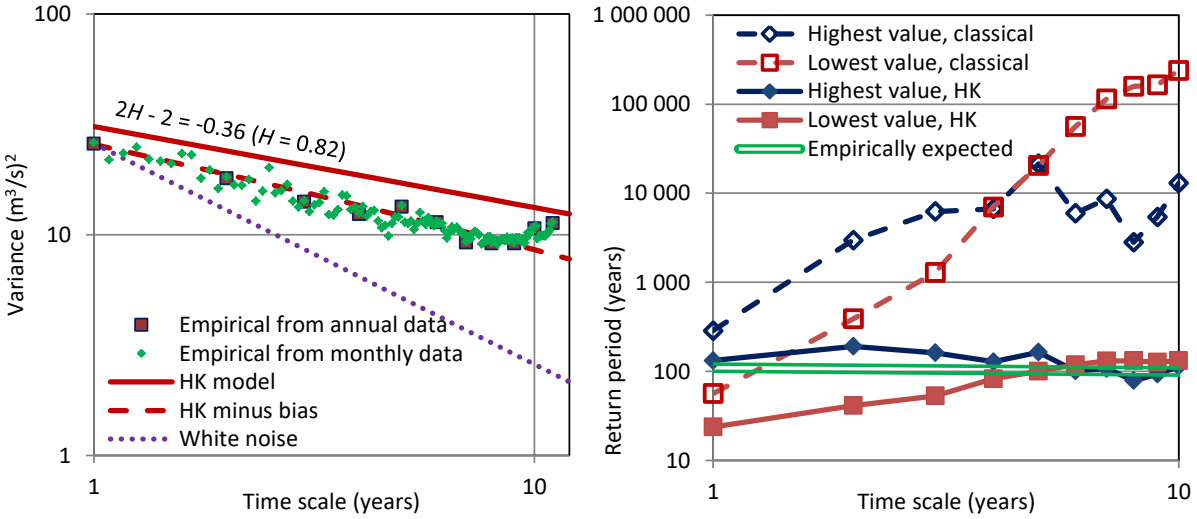

Figure 1.10 (left) Empirical and theoretical climacograms of the Boeoticos Kephisos discharge time series; **(right)** return periods of the lowest and highest observed average discharge over time scale 1 (annual scale) to 10 years (decadal scale) assuming normal distribution.

The concept of return period will be discussed in detail in Chapter 5. For the current discussion of our example, it suffices to say that, theoretically, the return period T of an event, which has probability P to occur in a time interval D, is related to P and D by the almost obvious relationship:

$$P = \frac{D}{T} \tag{1.7}$$

If we consider the highest or the lowest value that have been observed in a time period nD (where n is the sample size), then we can empirically assign to each of them a probability $P \approx 1/n$ and thus $T = nD$. If we change the time interval D to κD then the sample size of the observations becomes n/κ and again the empirical return period will be $T = nD$. Thus, in our record of 112 years (n = 112, D = 1 year) the empirical return period of the highest or the lowest observed value can for now be assumed to be 112

years, regardless of the time scale we consider. (A refinement of this technique will be discussed in Chapter 5 and Chapter 6.)

That is about the empirical return period. Now let us make a model for the process assuming normal marginal distribution with mean μ and standard deviation σ at time scale 1 (year), and time dependence consistent with the HK model. The estimates of these parameters for the HK model from the 110-year sample of annual values are μ = 11.69 m^3/s, σ = 5.56 m^3/s and H = 0.82. The method proposed in Koutsoyiannis (2003) was used for this estimation. For scales $\kappa > 1$ the normal distribution is preserved and so does the mean, while, according to equation (1.6), the standard deviation $\sigma(\kappa) = \sqrt{\gamma(\kappa)}$ will decrease according to $\sigma(\kappa) = \sigma/\kappa^{1-H}$. Therefore, for each scale we can determine the theoretical mean and standard deviation, find the theoretical probabilities of the highest and lowest values x_H and x_L, i.e., $P_H = P\{\underline{x} > x_L\}$ and $P_L = P\{\underline{x} \le x_L\}$, respectively, from the distribution function of the normal distribution, and determine the return period T from equation (1.7). The results of this exercise are visually shown in the right panel of Figure 1.10, where an agreement of theoretical and empirical distributions (T = 112 years) is seen. An underestimation of the theoretical return period of the lowest values for time scales 1-3 years is attributed to the fact that the normal distribution is not good enough for the lower distribution tail, as it is not bounded by 0, as it should; this deficiency ceases for larger scales, as the ratio σ/μ becomes smaller. All in all, the story told by the graph for the case that we assumed the HK model is that, in whatever time scale, the severe drought was as severe as expected for a 112-year period. Nothing more severe than expected.

Now let us assume that an expert on extremes, acting in 1995—around the end of the drought—was asked by water managers to assess the severity of the drought in terms of its return period. Further, let us assume that our extreme expert was ignorant of the HK behaviour and used classical statistics, as usually extreme experts do. Apart from that, let us assume that he adopted the same approach as above except the HK behaviour, which is equivalent to assuming H = 0.5. The expert at that time, based on the data and ignoring the estimation bias, which is absent in classical statistics, would estimate for the annual scale the mean as μ = 12.56 m^3/s and the standard deviation as σ = 5.01 m^3/s, which are not quite different from the estimates given before. However, by assuming independence and going to larger time scales, the standard deviation will differ substantially and, as a result, the return period will elevate. As seen in Figure 1.10, according to classical statistics, for time scales > 6 years the return periods of the lowest values exceed 100 000 years! Even for the largest values, high return periods are estimated, of the order of 10 000 years. Thus, the extreme expert would conclude that something extraordinary extreme has happened which requires an attribution study to most probably relate it to anthropogenic global warming. Evidently, such attributions differ substantially from similar ones in previous centuries. For example, after the great flood of the Arno River in Florence in November 1333 (the first recorded, which killed more than 3 000 people), it

was chronicled by Giovanni Villani* that "*the great debate in Florence was on whether the flood occurred for God's will or for natural causes.*"

Therefore, it is the Hurst-Kolmogorov dynamics that characterizes the natural changes, restores the estimates of extremes to reality and enables a cool look at extremes and their uncertainty, which is useful, if not absolutely necessary, for their management. And indeed, the HK behaviour has been the theoretical (stochastic) backing of the modelling and the successful handling of the Athens drought episode. On the other hand, it is striking that the name "Hurst" does not even appear in recent publications related to drought episodes (some of which have already been cited).

1.6 Maximum and minimum water level of the Nile

The longest instrumental record in history is that of the water level of the Nile. The observations were taken at the Roda Nilometer, near Cairo (Figure 1.11). Toussoun (1925) published the annual minimum and maximum water levels at the Roda Nilometer from AD 622 to 1921. The measurements were made available on the internet (Koutsoyiannis, 2013b)†, also converting water levels into water depths assuming a datum for the river bottom of 8.80 m. During 622-1470 (849 years), the record is almost uninterrupted but later there are large gaps. A few missing values of minima in the period before 1470 (namely, of the years 1285, 1297, 1303, 1310, 1319, 1363 and 1434) were filled in by Koutsoyiannis (2013b) using a simple method from Koutsoyiannis and Langousis (2011; p. 57), refined in Pappas et al. (2014).

The annual minimum and maximum water levels of this period are plotted in Figure 1.12 along with their climatic values given as 30-year averages. Due to the large extent of the Nile basin, the climatic fluctuation shown in the figure reflects the climate evolution of a very large area in the tropics and subtropics. We may notice that at the 780s the climatic (30-year) minimum value was 1.5 meters, while at AD 1110 and 1440 it was 4 meters, 2.5 times higher. In the lower panel of Figure 1.12 we can see a simulated series from a roulette wheel which has equal variance as the minimum water depth Nilometer series. Despite equal "annual" variability, the roulette wheel produces a static "climate", while the actual climate has varied substantially all the time.

Comparing the two Nilometer series, we observe that the series of maximum water depths exhibits much smaller variability than that of the minimum depths. This seems counterintuitive at first glance but we should bear in mind that, while the minimum depth refers to water confined in the river banks, the maximum one refers to a wide area inundated by the Nile water during flooding. One may express doubts about the accuracy of the measurements and record keeping in that era, several centuries ago, particularly in view of some points in the graph that look extraordinarily low or high outliers in each of the time series. However, the data can be crosschecked in some instances by historical information. As an example, the maximum water level in AD 967 is registered as 6.78 m

* Cronica, Tomo III, Libro XII, II; original text: "*D'una grande questione fatta in Firenze se 'l detto diluvio venne per iudicio di Dio o per corso naturale ...*"

† http://www.itia.ntua.gr/1351/.

which is the second lowest maximum water depth, 1.74 m below the local climatic average (see Figure 1.12). The historical information seems consistent with this extraordinarily low value: it is known that in AD 967 about a quarter of Egypt`s population (600 000 people) died of starvation and famine-related diseases (Fagan, 2008).

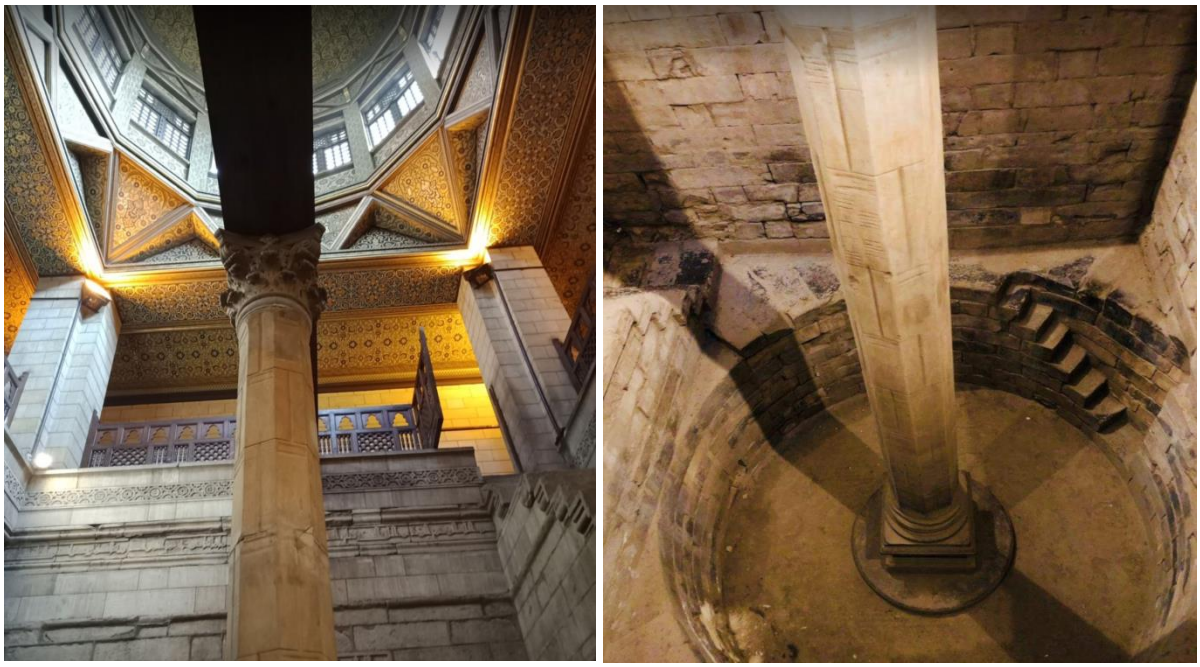

Figure 1.11 The Roda Nilometer, near Cairo. Water entered through three tunnels and filled the Nilometer chamber up to river level. The measurements were taken on the marble octagonal column (with a Corinthian crown) standing in the centre of the chamber; the column is graded and divided into 19 cubits (each slightly more than 0.5 m) and could measure floods up to about 9.2 m. A maximum level below the 16th mark could portend drought and famine and a level above the 19th mark meant catastrophic flood (Photos by Loai Samen and Mohamd Mubarak; Google maps, https://goo.gl/maps/T8NUgoDAorK2 and https://goo.gl/maps/dsdJHJYVv572).

While both decreasing and increasing trends appear in both time series, with most prominent the increasing trend in the series of maximum depths in the 14th and early 15th century, their alternating and aperiodic character defies a deterministic description. On the other hand, the stochastic description of the changes based on the HK dynamics is efficient. Indeed, Figure 1.13, which depicts the empirical and theoretical climacograms of the two Nilometer time series, shows that the natural changes are consistent with the HK behaviour.

The big length of these time series enables the validation of the HK hypothesis for a large range of time scales, from 1 to 84 (years). The difference from the popular white noise model (slope −1) is striking, as well as that of other popular models such as the Markov which will be discussed in section 3.11. The Hurst parameters are high, $H = 0.89$ for the series of minima and $H = 0.91$ for the series of maxima. Similar H values have been estimated from the contemporary, 131-year long, flow record of the Nile (naturalized) flows at Aswan (Koutsoyiannis and Georgakakos, 2006). The most notable deviation of the empirical behaviour and the HK model, shown in Figure 1.13, appears at scale 1 year for the series of maxima. The difference corresponds to the occurrence of extraordinarily

high or low maxima at isolated years. And as discussed above, these occurrences have been responsible for famines with thousands of lives lost.

In summary, the long Nilometer time series augments our confidence to the applicability in hydroclimatic processes of the HK behaviour which appeared in all our examples. According to this behaviour:

- long-term changes are more frequent and intense than commonly perceived;
- these changes are irregular and aperiodic, appear as alternating trends that can persist even for centuries, and are unpredictable per se;
- future states are much more uncertain and unpredictable on long time horizons than implied by pure randomness.

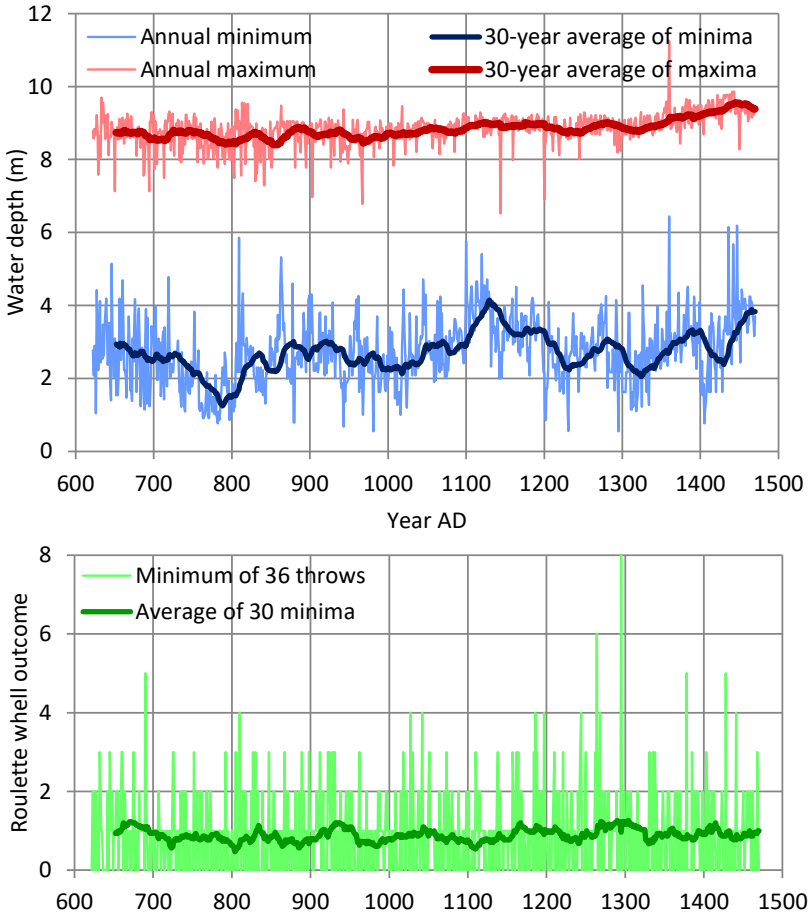

Figure 1.12 (upper) Nile River annual minimum and maximum water depth at Roda Nilometer (849 and 848 values, respectively, from Toussoun, 1925, as provided by Koutsoyiannis, 2013b). **(lower)** Synthetic time series, each value of which is the minimum of m = 36 roulette wheel outcomes; the value of m was chosen so that the standard deviation equals that of the minima of Nilometer series (where the latter is expressed in metres). In all series the climatic values, given as 30-year moving averages, are also plotted (right aligned).

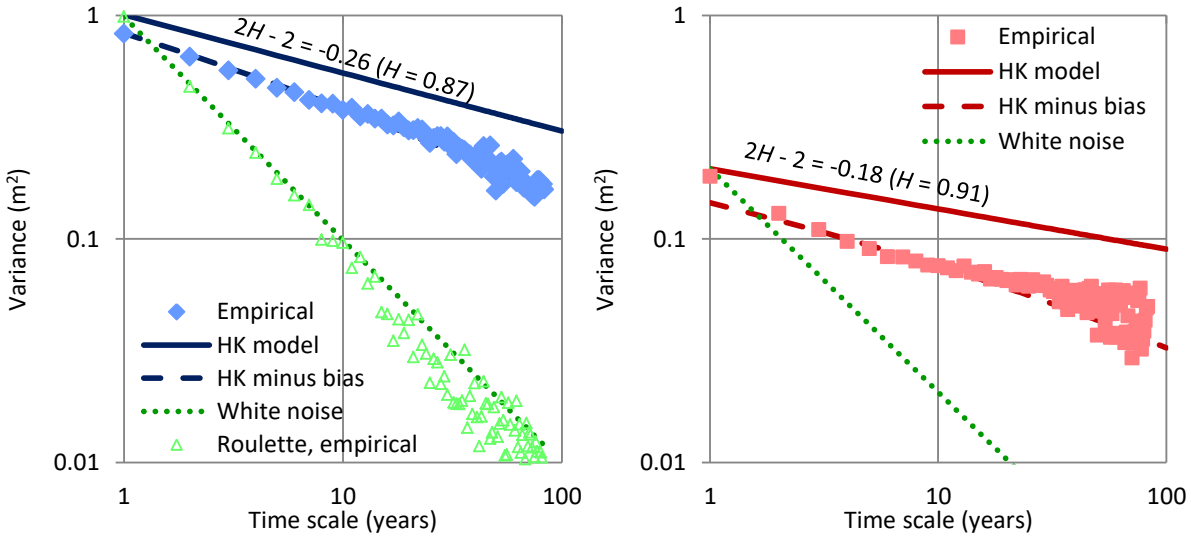

Figure 1.13 Empirical and theoretical climacograms of the two Nilometer series: **(left)** minimum and **(right)** maximum water depth; in the left graph the empirical climacogram of the roulette wheel time series is also shown, which is consistent with the white noise model, as expected.

Chapter 2. Basic concepts of probability with focus on extreme events

2.1 Definition of probability

For the understanding and the correct use of probability, it is very important to insist on the definitions and clarification of its fundamental concepts. Such concepts may differ from other, more familiar, arithmetic and mathematical concepts, and this may create confusion or even collapse of our cognitive construction, if we do not base it in concrete foundation. For instance, in our everyday use of mathematics, we expect that all quantities are expressed by numbers and that the relationship between two quantities is expressed by the notion of a function, which to a numerical input quantity associates (maps) another numerical quantity, a unique output. Probability too does such a mapping, but the input quantity is not a number but an event, which mathematically can be represented as a set. Probability is then a quantified likelihood that the specific event will happen. This type of representation was proposed by Kolmogorov (1933). There are other probability systems different from Kolmogorov's axiomatic system, according to which the input is not a set. Thus, in Jaynes (2003)* the input of the mapping is a logical proposition and probability is a quantification of the plausibility of the proposition. The two systems are conceptually different but the differences mainly rely on interpretation rather than on the mathematical results. Here we will follow Kolmogorov's system.

Table 2.1 Terminology correspondence in set theory and probability theory (adapted from Kolmogorov, 1933)

Set theory	Events
$A = \emptyset$	Event A is impossible
$A = \Omega$	Event A is certain
$AB = \emptyset$ (or $A \cap B = \emptyset$; disjoint sets)	Events A and B are incompatible (mutually exclusive)
$AB...N = \emptyset$	Events A, B, …, N are incompatible
$X := AB...N$	Event X is defined as the simultaneous occurrence of A, B, …, N
$X := A + B + ... + N$ (or $X := A \cup B \cup ... \cup N$)	Event X is defined as the occurrence of at least one of the events A, B, …, N
$X := A - B$	Event X is defined as the occurrence of A and, at the same time, the non-occurrence of B
$\overline{A}$ (the complementary of A)	The opposite event $\overline{A}$ consisting of the non-occurrence of A
$B \subseteq A$ (B is a subset of A)	From the occurrence of event B follows the inevitable occurrence of event A

Kolmogorov's approach to probability theory is based on the notion of *measure*, which maps *sets* onto *numbers*. The objects of probability theory, the *events*, to which probability

* Jaynes's book that we cite here was published after his death in 1998.

is assigned, are thought of as sets. For instance, the outcome of a roulette spin, i.e. the pocket in which the ball eventually falls on to the wheel, is one of 37 (in a European roulette pockets numbered 0 to 36 and coloured black or red except 0 which is coloured green). Thus, all sets {0}, {1}, … {36} are events (also called elementary events). But they are not the only ones. All possible subsets of Ω, including the empty set Ø, are events. The set Ω := {0, 1, …, 36} is an event too. Because any possible outcome is contained in Ω, the event Ω occurs in any case and it is called the *certain event*. The sets ODD := {1, 3, 5, …, 35}, EVEN := {2, 4, 6, …, 36}, RED := {1, 3, 5, 7, 9, 12, 14, 16, 18, 19, 21, 23, 25, 27, 30, 32, 34, 36}, and BLACK := Ω – RED – {0} are also events (in fact, betable). While events are represented as sets, in probability theory there are some differences from set theory in terminology and interpretation, which are shown in Table 2.1.

According to Kolmogorov's (1933) axiomatization, probability theory is based on three fundamental concepts and four axioms. The concepts form the triplet (Ω, Σ, P), called *probability space*, where:

1. Ω is a non-empty set, which Kolmogorov calls the *basic set* (sometimes also called *sample space* or the *certain event*), whose elements ω are called *elementary events* (also known as *outcomes* or *states*).
2. Σ is a set known as *σ-algebra* or *σ-field* whose elements E are subsets of Ω, known as *events*. Ω and Ø are both members of Σ, and, in addition, (a) if E is in Σ then the complement $\Omega - E$ is in Σ; (b) the union of countably many sets in Σ is also in Σ.
3. P is a function called *probability* that maps events (i.e., sets) to real numbers, assigning to each event E (member of Σ) a number between 0 and 1.

The four axioms, which define the properties of P, are:

I. *Non-negativity*: For any event A, $P(A) \geq 0$.
II. *Normalization*: $P(\Omega) = 1$.
III. *Additivity*: For any incompatible events A and B (i.e., $AB = \emptyset$), $P(A + B) = P(A) + P(B)$.
IV. *Continuity at zero*: If $A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n \supseteq \ldots$ is a decreasing sequence of events, with $A_1A_2 \ldots A_n \ldots = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$.

We note that in the case that Σ is finite, axiom IV follows from axioms I-III; however, for infinite fields it should be put as an independent axiom.

2.2 The concept of a stochastic variable

A *stochastic variable* or *random variable*[*] is a function that maps outcomes to numbers, i.e. enumerates the basic set Ω. More formally, according to Kolmogorov's (1933) definition, a real single-valued function $x(\omega)$, defined on the basic set Ω, is called a *random variable* if for each choice of a real number a the set $\{x(\omega) < a\}$ for all ω for which the inequality $x(\omega) < \alpha$ holds true, belongs to Σ. With the concept of the stochastic variable we can

[*] The two terms *stochastic variable* and *random variable* have identical meaning. Here we prefer the former, even though the latter is more common.

conveniently express events using basic mathematics. In most cases enumeration is done almost automatically. For instance, a stochastic variable that takes values 1 to 6 is intuitively assumed when we deal with a die throw experiment.

We must be attentive that a stochastic variable is not a number but a function. Intuitively, we could think of a stochastic variable as an object that represents *simultaneously* all possible outcomes and only them. The following analogy may help us to develop intuition about stochastic variables. Let us consider the equation $x^3(x-1)^2 = 0$. This has five roots, three of them being $x = 0$ and two being $x = 1$. What do we mean when we say "root of this equation"? Probably we mean both $x = 0$ and $x = 1$ and also we have in mind that there is not symmetry between the two; rather we would give a weight $3/5$ on the former and $2/5$ on the latter. Similar is the situation with a stochastic variable which takes on the values 0 and 1 with probabilities $3/5$ and $2/5$, respectively.

While formally a stochastic variable is a function $x(\omega)$, we usually omit reference to its argument ω and keep the symbol x. However, in this case we need to distinguish it symbolically from a regular variable; the best notation devised to this aim and used here is the so-called Dutch convention (see Hemelrijk, 1966, who mentions that it was introduced by D. Van Dantzig in 1947, i.e. later than Kolmogorov's foundation of probability). According to it, stochastic variables are underlined, i.e. $\underline{x}$. In this case the inequality $\{x(\omega) < a\}$ used for the formal definition of the stochastic variable is written as $\{\underline{x} < a\}$. Accordingly, $\{\underline{x} < a\}$ denotes an event (a subset of Ω), and therefore it has a probability, $P(\{\underline{x} < a\})$. For simplicity, in the latter notation we drop the parenthesis and we write $P\{\underline{x} < a\}$. Some texts drop the curly brackets instead of the parentheses, but this tactic misrepresents the important fact that the argument of probability is a set. This notation is further explained in Digression 2.A, along with its importance.

From a practical point of view, compared to a regular variable, a stochastic variable is a more abstract mathematical entity, which we use when a quantity of interest is something uncertain, unpredictable, unknown; this is the meaning of *stochastic* and *random* (cf. Koutsoyiannis, 2010; Dimitriadis et al., 2016). While a regular variable takes on one value at a time, a stochastic variable can be thought of as it takes on all its possible values at once, but not necessarily in a uniform manner; therefore, a probability distribution function, defined in next session, should always be associated with a stochastic variable. A stochastic variable becomes identical to a regular variable only if it can take on only one value.

When we perform an observation of a quantity that is modelled as a stochastic variable, then this observation is usually a regular variable. For example, we model a die throw with a stochastic variable $\underline{x}$ with possible values 1 to 6. After a specific throw of the dice and before we observe the outcome, we still have the same uncertainty as described by stochastic variable $\underline{x}$. When we observe the outcome, it becomes a regular variable x (e.g. $x = 5$). The particular value is called a *realization* of $\underline{x}$ and is denoted by the non-underlined symbol x. This happens when our observation is exact. Sometimes the observation is contaminated by error—our observations are not always exact (particularly those of real valued variables). Then we can use another stochastic variable

to describe the uncertain outcome. For example, if an observer has presbyopia combined with astigmatism (like the author) he may not be sure that the outcome was 5 or 4 and he could model it as a stochastic variable $\underline{z}$ with possible outcomes 4 and 5.

Considering a certain (deterministic) function $y = g(x)$, mapping the regular variable x to the regular variable y (e.g. $y = g(x) = x^2$), we can extend its meaning to apply to stochastic variables, i.e., $\underline{y} = g(\underline{x})$ (e.g. $\underline{y} = g(\underline{x}) = \underline{x}^2$). As implied by the notation, when the function's argument $\underline{x}$ is a stochastic variable, the result $\underline{y}$ is also a stochastic variable (formally, it is the composite function $y(\omega) = g(x(\omega))$. In other words, functions of stochastic variables are stochastic variables.

Digression 2.A: The importance of notation

The following simple example shows that the common practice of not distinguishing the notation of regular and stochastic variables is a bad practice. Let $\underline{x}$ and $\underline{y}$ represent the outcomes of each of two dice. What is the probability of the following cases?

$$\text{(a) } \left\{\underline{x} < \underline{y}\right\}, \text{ (b) } \{\underline{x} < y\}, \text{ (c) } \left\{x < \underline{y}\right\}, \text{ (d) } \{x < y\}.$$

(a) There are $6^2 = 36$ different possible combinations of outcomes of $\underline{x}$ and $\underline{y}$. In six of them $\underline{x} = \underline{y}$. Due to symmetry, in half of the remaining 30, $\underline{x} < \underline{y}$. Thus:

$$P\left\{\underline{x} < \underline{y}\right\} = \frac{15}{36} = \frac{5}{12}$$

(b) Now y is a number, not a stochastic variable. For convenience we assume that y is integer, even though it can also be assumed to be real. If $y > 6$ then obviously the event $\{\underline{x} < y\}$ is certain. If $y = 6$ then the probability of $\{\underline{x} < y\}$ is 5/6. Continuing like this we conclude that:

$$P\{\underline{x} < y\} = \max\left(0, \min\left(1, \frac{y-1}{6}\right)\right)$$

(c) Thinking as in (b) and noting that x is a number, assumed integer, and $\underline{y}$ a stochastic variable we find that:

$$P\left\{x < \underline{y}\right\} = \max\left(0, \min\left(1, 1 - \frac{x}{6}\right)\right)$$

(d) As both x and y are numbers the expression $\{x < y\}$ does not denote an event and therefore, strictly there is no probability associated with this expression. Loosely we may say that $P\{x < y\} = 1$ if $x < y$ and 0 otherwise.

Obviously, if we did not distinguish y from $\underline{y}$, we would even not be aware of the fact that $P\left\{\underline{x} < \underline{y}\right\}$ is a number while $P\left\{x < \underline{y}\right\}$ is a function of x.

Many texts (research articles and probability theory books) make the notational distinction of random and regular variables, but they use upper case letter for stochastic variables and lower case ones for regular variables. This may not be good enough a practice. If in our context we used another quantity denoted with the Greek letter χ (and actually χ is quite common in statistical texts—cf. the chi and chi-squared distributions), how would we distinguish the stochastic variables corresponding to x and χ? (In both cases the upper case letter is X, while in our convention $\underline{x}$ and $\underline{\chi}$ are distinguishable.) Furthermore, this would be too restrictive in our use of mathematical symbols. For example, the symbol H used in Chapter 1 (and many other chapters) to denote the Hurst parameter would be an incorrect notation if we adopted the upper- vs. lower-case notation. Another convention was used by Papoulis (1990, 1991) who denoted stochastic variables in bold letters. However, the typical use of bold letters is to denote vectors. Therefore,

the Dutch convention of underlining the stochastic variables is the most convenient, the clearest and the safest.

2.3 Distribution function

According to Kolmogorov's (1933) foundation* of probability theory, the function of the real variable x,

$$F(x) \coloneqq P\{\underline{x} \leq x\} \tag{2.1}$$

where $\underline{x}$ is a stochastic variable, is called the *distribution function*. We notice that the stochastic variable with which this function is associated is not an argument of the function. Even though we use the same letter for both $\underline{x}$ and x, the two are fundamentally different. For example, in a die throw, the stochastic variable $\underline{x}$ represents the whole numbers 1 to 6 and the regular variable x takes on any real value from −∞ to +∞. (The domain of $F(x)$ is not identical to the range of the stochastic variable $\underline{x}$; rather it is always the set of real numbers.) If there is risk of confusion (e.g. if we study a problem with many stochastic variables), the stochastic variable should also appear in the notation of the distribution function. Usually it is denoted as a subscript: $F_{\underline{x}}(x)$—but we can simplify the notation dropping the underscore in the subscript ($F_x(x)$) once we adopt the convention that the subscripts are stochastic variables.

Typically, $F(x)$ has a mathematical expression depending on some parameters. It is a non-decreasing function of x obeying the relationship:

$$0 = F(-\infty) \leq F(x) \leq F(+\infty) = 1 \tag{2.2}$$

For its non-decreasing attitude, in the English literature $F(x)$ is also known as *cumulative distribution function*, but here we adhere to the original Kolmogorov's (1933) terminology, which did not contain the adjective *cumulative*. In practical applications the distribution function is also known as *non-exceedance probability*. Likewise, the non-increasing function:

$$\overline{F}(x) \coloneqq P\{\underline{x} > x\} = 1 - F(x) \tag{2.3}$$

is known as, *tail function* (or survival function, survivor function) and represents *exceedance probability.*

The distribution function is always continuous on the right; however, if the basic set Ω is finite or countable, $F(x)$ is discontinuous on the left at all points x_i that correspond to outcomes ω_i, and it is constant between them (staircase-like). Such stochastic variable is called *discrete*. If $F(x)$ is a continuous function, then the stochastic variable is called *continuous*. A *mixed* case is also common; in this the distribution function has some discontinuities on the left, but is not staircase-like.

For continuous stochastic variables, the inverse function $F^{-1}(\,)$ of $F(\,)$ exists. Consequently, the equation $u = F(x)$ has a unique solution for x, called *u-quantile* of the variable $\underline{x}$, that is:

* We note that Kolmogorov used '<' in his definition but modern literature uses '≤' as in (2.1).

$$x_u = F^{-1}(u) \tag{2.4}$$

2.4 Probability mass and density function

In discrete stochastic variables, the probability of each event:

$$P_j := P\{\underline{x} = x_j\}, \qquad j = 1, \dots, J \tag{2.5}$$

where J is the number of possible outcomes (which can be infinite) is the *probability mass function*. It is easy then to see that the step (discontinuity) of the distribution function $F(x)$ at point x_j equals P_j.

In continuous variables there are no discontinuities and hence any particular value x has zero probability to occur. However, we can still tell which of two outcomes is more probable than the other and how much by examining the ratio of the two probabilities. As this is a $0/0$ expression, having in mind l'Hôpital's rule, we need to examine the ratio of derivatives of probabilities.

The derivative of the distribution function is called the *probability density function* (PDF) or simply *density*:

$$f(x) := \frac{\mathrm{d}F(x)}{\mathrm{d}x} \tag{2.6}$$

and its basic properties are:

$$f(x) \geq 0, \qquad \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1 \tag{2.7}$$

Obviously, the probability density function does not represent a probability; therefore, it can take on values higher than 1. Its relationship with probability is described by the following equation:

$$f(x) = \lim_{\Delta x \to 0} \frac{P\{x \leq \underline{x} \leq x + \Delta x\}}{\Delta x} \tag{2.8}$$

The distribution function can be calculated from the density function by:

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y \tag{2.9}$$

Some of the most common PDFs are shown in Table 2.2 along with their distribution functions. Additional distributions are shown in Table 2.3, along with their moments, while the derivation of these and other distributions in terms of the principle of maximum entropy is indicated in Table 2.4.

In discrete stochastic variables, the density is a sequence of Dirac δ functions (see definition of δ in equation (3.51)), while in mixed distributions Dirac δ functions appear at the points of discontinuity. This text mostly deals with continuous variables but mixed-type variables appear in several cases as will be discussed in Chapter 6 and Chapter 8.

As already discussed (section 2.2), the one-to-one mathematical transformation on $\underline{x}$, $\underline{y} = g(\underline{x})$ defines a new stochastic variable $\underline{y}$. The event $\{\underline{y} \leq y\}$ is identical to the event

$\{\underline{x} \le g^{-1}(y)\}$ where g^{-1} is the inverse function of g. Consequently, the distribution functions of $\underline{x}$ and $\underline{y}$ are related by:

$$F_y(y) = P\{\underline{y} \le y\} = P\{\underline{x} \le g^{-1}(y)\} = F_x(g^{-1}(y)) \tag{2.10}$$

In the case that the variables are continuous and the function g differentiable, it can be shown that the density functions of $\underline{x}$ and $\underline{y}$ are related by:

$$f_y(y) = \frac{f_x(g^{-1}(y))}{|g'(g^{-1}(y))|} \tag{2.11}$$

where g' is the derivative of g.

Table 2.2 Some of the simplest and most common distributions of continuous variables.

Name (and parameters)	Probability density function	Distribution function
Uniform in [0, 1]	$f(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$	$F(x) = \max(0, \min(x, 1))$
Exponential ($\mu > 0$)	$f(x) = \begin{cases} e^{-x/\mu} / \mu & x \ge 0 \\ 0 & x < 0 \end{cases}$	$F(x) = \begin{cases} 1 - e^{-x/\mu} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$
Normal ($\mu \in \mathbb{R}$, $\sigma > 0$)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du$

Digression 2.B: Illustration of distribution function by an example

For clarification of the basic concepts of probability theory, we give the following example of hydroclimatic interest. In particular we study (a) the occurrence of rainfall at a particular site and a specific time of the year, and (b) the rainfall depth at that site and time.

In (a) we are interested on the mathematical description of the possibilities that a certain day in the specified site and time and is wet or dry. These are the outcomes or states of our problem, so the basic set is:

$$\Omega = \{\text{wet, dry}\}$$

The field Σ contains all possible events, i.e.:

$$\Sigma = \{\emptyset, \{\text{wet}\}, \{\text{dry}\}, \Omega\}$$

To fully define probability on Σ it suffices to define the probability of one of the two states, say $P\{\text{wet}\}$. In fact, this is not easy – usually it is done by induction, and it needs a set of observations to be available and concepts of the *statistics* theory (see Chapter 4) to be applied. For the time being let us arbitrarily assume that $P\{\text{wet}\} = 0.2$. The remaining probabilities are obtained by applying the axioms. Clearly, $P(\Omega) = 1$ and $P(\emptyset) = 0$. Since wet and dry are incompatible, $P\{\text{wet}\} + P\{\text{dry}\} = P(\{\text{wet}\} + \{\text{dry}\}) = P(\Omega) = 1$, so $P\{\text{dry}\} = 0.8$.

We define a stochastic variable $\underline{x}$ based on the rule

$$x(\text{dry}) = 0, \qquad x(\text{wet}) = 1$$

We can now easily determine the distribution function of $\underline{x}$. For any $x < 0$,

$$F(x) = P\{\underline{x} \le x\} = 0$$

(because $\underline{x}$ cannot take negative values). For $0 \leq x < 1$,

$$F(x) = P\{\underline{x} \leq x\} = P\{\underline{x} = 0\} = 0.8$$

Finally, for $x \geq 1$,

$$F(x) = P\{\underline{x} \leq x\} = P\{\underline{x} = 0\} + P\{\underline{x} = 1\} = 1$$

The graphical depiction of the distribution function is shown in Figure 2.1 (left). The staircase-like shape reflects the fact that stochastic variable is discrete.

In (b) the state is described by the rainfall depth which can be zero or positive. Therefore, the basic set is the set $\mathbb{R}^+ \cup \{0\}$. The stochastic variable $\underline{x}$ is given by the rule $x(\omega) = \omega$. Again, the distribution function of $\underline{x}$ will be $F(x) = P\{\underline{x} \leq x\} = 0$ for $x < 0$ with a discontinuity at 0, so that $F(0^+) = P\{\underline{x} = 0\} = 0.8$. For $x \geq 0$ the distribution function will be continuous and increasing, approaching 1 as $x \to \infty$. To construct a plausible distribution function, without examining observations, we make an assumption that smaller values are more probable than higher and specifically that for two values x_1 and $x_2 > x_1$, the ratio of densities (expressing the ratio of probabilities according to l'Hôpital's rule) depends on the difference $x_2 - x_1$, i.e.,

$$\frac{f(x_1)}{f(x_2)} = g(x_2 - x_1)$$

where it is easy to see that the function $g(\,)$ should be given as $g(x) = f(0)/f(x)$. In turn, it can be shown (homework) that $f(x) = A\exp(-Bx)$ where A and B are constants. By integrating (according to equation (2.9)) we find:

$$F(x) = \frac{A}{B}(1 - \exp(-Bx)) + C$$

and, since $F(0^+) = 0.8$ and $F(\infty) = 1$, $C = 0.8$ and $A/B = 0.2$ and thus:

$$F(x) = 0.2(1 - \exp(-Bx)) + 0.8$$

where B can be any positive number. An example is depicted in Figure 2.1 (right) for $B = 1$. The result is a modified exponential distribution (see Table 2.2), where the modification resulted from the fact that the distribution is not continuous everywhere but mixed.

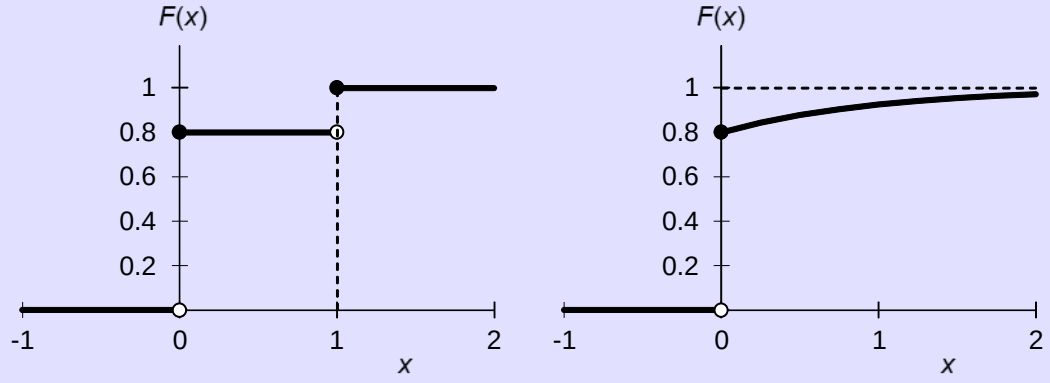

Figure 2.1 Distribution function of a stochastic variable representing events related to rainfall of a given day at a certain area and time of the year: **(left)** the dry or wet state; **(right)** the rainfall depth.

If this mathematical model is to represent a physical phenomenon, we must have in mind that all probabilities depend on a specific location and a specific time of the year. So, the model cannot be a global representation of the wet and dry state of a day, nor of the rainfall depth. The model as formulated here is extremely simplified. It does not make any reference to the succession of dry or wet states in different days. This is not an error; it simply diminishes the predictive capacity of the model. A better model would describe separately the probability of a wet day following a wet day, a wet day following a dry day (we anticipate that the latter should be smaller than the

former), etc. In addition, while the assumption on the rainfall depth leading to a mixed exponential distribution is plausible on first glance, it does not fully correspond to the empirically observed behaviour. There are better models than the exponential. We will discuss these issues in subsequent sections.

2.5 Conditional probability, independent and dependent events

By definition (Kolmogorov, 1933), conditional probability of the event A given B (i.e. under the condition that the event B has occurred) is the quotient:

$$P(A|B) := \frac{P(AB)}{P(B)} \tag{2.12}$$

Obviously, if $P(B) = 0$, this conditional probability cannot be defined. It follows that:

$$P(AB) = P(A|B)P(B) = P(B|A)P(A) \tag{2.13}$$

From this it follows that:

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)} \tag{2.14}$$

The latter equation is known as the *Bayes theorem*.

If it happens that $P(A|B) = P(A)$, i.e., the probability of A does not depend on whether or not B has occurred, then the events A and B are called *independent* (or *stochastically independent*). In this case from equation (2.12) it follows that:

$$P(AB) = P(A)P(B) \tag{2.15}$$

Otherwise A and B are called (*stochastically*) *dependent*.

The definition can be extended to many events. Thus, the events $A_1, A_2, \ldots$ are *independent* (or *mutually independent*) if for any finite set of distinct indices $i_1, i_2, \ldots, i_n$:

$$P\left(A_{i_1}A_{i_2}\ldots A_{i_n}\right) = P\left(A_{i_1}\right)P\left(A_{i_2}\right)\ldots P\left(A_{i_n}\right) \tag{2.16}$$

The handling of probabilities of independent events is thus easy. However, this is a special case because usually natural events are dependent. In the handling of dependent events the notion of *conditional probability* is vital.

It is easy to show that the generalization of (2.16) for dependent events takes the forms:

$$P(A_n \ldots A_1) = P(A_n|A_{n-1}\ldots A_1)P(A_2|A_1)P(A_1) \tag{2.17}$$

$$P(A_n \ldots A_1|B) = P(A_n|A_{n-1}\ldots A_1B)P(A_2|A_1B)P(A_1|B) \tag{2.18}$$

which are known as the *chain rules*. It can also be proved (homework) that if A and B are mutually exclusive, then

$$P(A+B|C) = P(A|C) + P(B|C) \tag{2.19}$$

$$P(C|A+B) = \frac{P(C|A)P(A) + P(C|B)P(B)}{P(A)+P(B)} \tag{2.20}$$

and if $A + B = \Omega$, so that $P(C|A+B) = 1$, then

$$P(C) = P(C|A)P(A) + P(C|B)P(B) \tag{2.21}$$

Digression 2.C: An example on the dependence of probability on information

We assume that at a certain place on Earth (say, in a city in the United Kingdom) and a certain period of the year a dry and a wet day are equiprobable and that in the different days the states (wet or dry) are independent. What is the probability that two consecutive days are wet under the following conditions? (a) Unconditionally. (b) If we know that the first day is wet. (c) If we know that the second day is wet. (d) If we know that one of the two days is wet.

We denote $A \coloneqq$ {first day wet}, $\bar{A} \coloneqq$ {first day dry}, $B \coloneqq$ {second day wet}, $\bar{B} \coloneqq$ {second day dry}. The basic set is $\{AB, A\bar{B}, \bar{A}B, \bar{A}\bar{B}\}$

(a) We seek to find $P(AB)$. Obviously, given the independence assumption, $P(AB) = P(A)P(B) = (1/2)^2 = 1/4$. Because of equiprobability and independence, each of the four events has probability $1/4$.

(b) Now the probability sought is $P(AB|A)$. Using the chain rule in equation (2.18) we find $P(AB|A) = P(A|AB)P(B|A) = 1 \times 1/2 = 1/2$.

(c) Like in (b) we find $P(AB|B) = 1/2$.

(d) The condition that one of the two days is the composite even $AB + A\bar{B} + \bar{A}B$. Thus, the probability sought is

$$P(AB|AB + A\bar{B} + \bar{A}B) = \frac{P\big(AB(AB + A\bar{B} + \bar{A}B)\big)}{P(AB + A\bar{B} + \bar{A}B)} = \frac{P(AB)}{P(AB + A\bar{B} + \bar{A}B)} = \frac{1/4}{3/4} = \frac{1}{3}$$

where we have used the definition of conditional probability and the fact that $AB, A\bar{B}, \bar{A}B$ are mutually exclusive.

To connect the example to the real world, let us assume that a friend travelled to this city for a specified couple of days. If we do not have any information except the specific dates, then to the event that she used her umbrella in both days we will assign probability $1/4$. If we have seen (e.g. in her social media posts) a photo showing her in the city holding an umbrella, then to the same event we will assign probability $1/3$. If, in addition, the photo has a time stamp on it, then we will change the probability to $1/2$. In other words, the information we have in a problem may introduce dependences in events that are initially assumed independent. More generally, the probability is not an invariant quantity, characteristic of physical reality in absolute terms, but a quantity that depends on our knowledge or information on the examined phenomenon. It may sound paradoxical that the probability depends on information, but it is not. The rules we are assigning probabilities are objective and theoretically consistent—there was nothing ambiguous in calculating the above probabilities, based on the information given each time. We may additionally recall that even in classical deterministic physics we are dealing with similar situations. For instance, the location and velocity of a moving particle are not absolute objective quantities. If we change the coordinate system, the numerical values of the coordinates and the velocity will also change.

Digression 2.D An example on dependent events

The independence assumption in the problem in Digression 2.C is obviously a poor representation of the physical reality. To make a slightly more realistic model, let us assume that the probability of today being wet (B) or dry ($\bar{B}$) depend on the state of yesterday (A or $\bar{A}$). It is reasonable to assume that the following inequalities hold:

$$P(B|A) > P(B) = 0.5, \qquad P(\bar{B}|\bar{A}) > P(\bar{B}) = 0.5$$

The problem now is more complicated than before. Let us arbitrarily assume that $P(B|A) = 0.6$. Then the probability that both days are wet is $P(AB) = P(B|A)P(A) = 0.6 \times 0.5 = 0.3 > 1/4$. For completeness we also calculate the probabilities of the other combinations. From (2.21), we get $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$, from which we find:

$$\begin{bmatrix} P(B|A) & P(B|\bar{A}) \\ P(\bar{B}|A) & P(\bar{B}|\bar{A}) \end{bmatrix} \begin{bmatrix} P(A) \\ P(\bar{A}) \end{bmatrix} = \begin{bmatrix} P(A) \\ P(\bar{A}) \end{bmatrix}, \quad \begin{bmatrix} P(A|B) & P(B|\bar{B}) \\ P(\bar{B}|A) & P(\bar{B}|\bar{A}) \end{bmatrix} \begin{bmatrix} P(B) \\ P(\bar{B}) \end{bmatrix} = \begin{bmatrix} P(B) \\ P(\bar{B}) \end{bmatrix}, \quad \begin{bmatrix} P(B) \\ P(\bar{B}) \end{bmatrix} = \begin{bmatrix} P(A) \\ P(\bar{A}) \end{bmatrix}$$

where for convenience we have used matrix/vector representation. Thus.

$$P(B|\bar{A}) = \frac{P(B) - P(B|A)P(A)}{P(A)} = \frac{0.5 - 0.6 \times 0.5}{0.5} = 0.4$$

Hence $P(\bar{A}B) = P(B|\bar{A})P(\bar{A}) = 0.4 \times 0.5 = 0.2 < 1/4$. Because of symmetry $P(\bar{A}\bar{B}) = 0.3$ and $P(A\bar{B}) = 0.2$. Thus, the dependence resulted in higher probabilities that the consecutive events are similar and smaller probabilities that they are dissimilar. This corresponds to a general natural behaviour (see also Chapter 3).

2.6 Random number generation for stochastic simulation

One of the important scientific advances offered by stochastics in the last several decades is the Monte Carlo method, else known as stochastic simulation. It was originally developed for the numerical solution of integro-differential equation in Los Alamos in the framework of the Manhattan Project (Metropolis and Ulam, 1949). It can be easily shown (e.g. Niederreiter, 1992) that high dimensional numerical integration (specifically for a number of dimensions $d > 4$), a stochastic (Monte Carlo) integration method (in which the function evaluation points are taken at random) is more accurate (for the same total number of evaluation points) than classical numerical integration (based on a grid representation of the integration space). This gave importance to the much older concept of random numbers, whose first appearance in a scientific publication was Tippett's (1927) table, with 41 600 random digits taken from a 1925 census report. Before that (and even after; see Digression 3.F) random sampling was performed by means of dice and cards. Thus, Galton (1890) invented a set of three modified dice to generate samples from a normal distribution. "Student" (pseudonym of W.S. Gosset) in 1908 performed simulation experiments using 3000 cards (in 750 groups of size 4) to find the distribution of the *t*-statistic and of the correlation coefficient (see more information in Stigler 2002).

With today's meaning, a *sequence of random numbers* is a sequence of numbers x_i whose every statistical property is consistent with that of realizations from a sequence of independent identically distributed stochastic variables $\underline{x_i}$ (adapted from Papoulis, 1990). In turn, a *random number generator* is a device (typically computer algorithm) which generates a sequence of random numbers x_i with given distribution $F(x)$. Random number generation is also known as Monte Carlo sampling.

The basis of practically all random generators is the uniform distribution in [0,1] (see Table 2.2). A typical procedure is the following:

- We generate a sequence of integers q_i from the recursive algorithm $q_i = (k\, q_{i-1} + c) \bmod m$ where k, c and m are appropriate integers (e.g. k = 69 069, c = 1, $m = 2^{32}$ = 4 294 967 296 or $k = 7^5$ = 16 807, c = 0, $m = 2^{31} - 1$ = 2 147 483 647; Ripley, 1987, p. 39).

- We calculate the sequence of random numbers u_i with uniform distribution in [0,1] by $u_i = q_i/m$.

Obviously, this is a simple algorithm, purely deterministic. Why the numbers it generates are regarded as random? The answer is simple: Because if we do not know the algorithm and the initial condition (q_0 or q_{i-1}) we cannot predict these numbers. As most algorithms, like this one, are purely deterministic, sometimes the numbers are called pseudorandom. But this implies the idea that there exists another category of true or genuine random numbers. Even though in the literature references to true random numbers abound, this may reflect a misunderstanding of the notion of randomness and a dichotomic view of natural processes (cf. Koutsoyiannis, 2010; Dimitriadis et al., 2016). In any process of the macroscopic world, if we were able to know the "algorithm" (the system dynamics), and the initial conditions with full precision, the situation would be the same as with the simple algorithm described. The fact that we are not able to know precisely the algorithm of a physical process and the initial conditions does not make the numbers of different type.

A more recent and better algorithm for random number generation with uniform distribution is the so-called *Mersenne twister*, which is available in most computer languages and software packages*.

Once we have a random generator for the uniform distribution, we can make one for any distribution $F(x)$. A direct (but sometimes time demanding) algorithm to produce random numbers x_i from *any* distribution $F(x)$ is given by:

$$x_i = F^{-1}(u_i) \tag{2.22}$$

where u_i is the sequence of random numbers with uniform distribution in [0,1].

2.7 Expectation

Expectation is a key concept of stochastics, enabling a macroscopic view of a phenomenon or process in which the details are intentionally neglected.

For a discrete stochastic variable $\underline{x}$, taking on the values $x_1, x_2, \dots, x_J$ (where J could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{x = x_j\}$, if $g(\underline{x})$ is an arbitrary function of $\underline{x}$ (so that $g(\underline{x})$ is a stochastic variable per se), we define the *expectation* or *expected value* or *mean* of $g(\underline{x})$ as:

$$\mathrm{E}\big[g(\underline{x})\big] := \sum_{j=1}^{J} g(x_j)P(x_j) \tag{2.23}$$

Likewise, for a continuous stochastic variable $\underline{x}$ with density $f(x)$, the expectation is defined as:

* For example, for Excel (which by default includes the function rand) the Mersenne twister algorithm, called NtRand, can be found in www.ntrand.com/download/.

$$\mathrm{E}\left[g\left(\underline{x}\right)\right] \coloneqq \int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x \tag{2.24}$$

Expected values are regular variables: for example, $\mathrm{E}\left[\underline{x}\right]$ and $\mathrm{E}\left[g\left(\underline{x}\right)\right]$ are constants—neither functions of x nor of $\underline{x}$. That justifies the notation $\mathrm{E}\left[\underline{x}\right]$ instead of $\mathrm{E}\left(\underline{x}\right)$ or $\mathrm{E}(x)$ which would imply functions of $\underline{x}$ or x.

2.8 Moments and cumulants

For certain types of functions $g\left(\underline{x}\right)$ we get very commonly used statistical parameters, as specified below:

- The *noncentral moment* of order p (or the p*th moment about the origin*):

$$g\left(\underline{x}\right) = \underline{x}^p, \qquad \mu'_p \coloneqq \mathrm{E}[\underline{x}^p] \tag{2.25}$$

- The *mean* (or the first moment):

$$g\left(\underline{x}\right) = \underline{x}, \qquad \mu \coloneqq \mu'_1 = \mathrm{E}[\underline{x}] \tag{2.26}$$

- The *central moment of order p*:

$$g\left(\underline{x}\right) = (\underline{x} - \mu)^p, \qquad \mu_p \coloneqq \mathrm{E}[(\underline{x} - \mu)^p] \tag{2.27}$$

  For p = 0 and 1 the central moments are respectively 1 and 0.

- The *variance*:

$$g\left(\underline{x}\right) = (\underline{x} - \mu)^2, \qquad \gamma \coloneqq \mathrm{E}[(\underline{x} - \mu)^2] \coloneqq \sigma^2 \tag{2.28}$$

  The variance is also denoted as $\mathrm{var}\left[\underline{x}\right]$; its square root σ (also denoted as $\mathrm{std}\left[\underline{x}\right]$) is called the standard deviation.

Among the moments of order higher than two, most commonly used are the third and fourth. If we standardize them by appropriate powers of σ to make them dimensionless, we get, respectively, the *coefficients of skewness* and *kurtosis*:

$$C_\mathrm{s} \coloneqq \frac{\mu_3}{\sigma^3}, \qquad C_\mathrm{k} \coloneqq \frac{\mu_4}{\sigma^4} \tag{2.29}$$

Another dimensionless index is the ratio:

$$\frac{1}{C_\mathrm{v}} = \frac{\mu}{\sigma}, \qquad C_\mathrm{v} \coloneqq \frac{\sigma}{\mu} \tag{2.30}$$

where the former case is always meaningful, while the latter (the inverse of the former) is meaningful for nonnegative stochastic variables and is called *coefficient of variation*.

Central and noncentral moments are related to each other by:

$$\mu'_p = \sum_{i=0}^{p} \binom{p}{i} \mu^{p-i} \mu_i, \qquad \mu_p = \sum_{i=0}^{p} \binom{p}{i} (-\mu)^{p-i} \mu'_i \tag{2.31}$$

where $\mu_0 = \mu_0' = 1, \mu_1 = 0, \mu_1' = \mu$. Proof of these relationships is given in Appendix 6-II. For small p they take the following forms:

$$\mu_2' = \sigma^2 + \mu^2, \qquad \mu_3' = \mu_3 + 3\sigma^2\mu + \mu^3, \qquad \mu_4' = \mu_4 + 4\mu_3\mu + 6\sigma^2\mu^2 + \mu^4 \tag{2.32}$$

and can be inverted as follows:

$$\sigma^2 = \mu_2' - \mu^2, \qquad \mu_3 = \mu_3' - 3\mu_2'\mu + 2\mu^3, \qquad \mu_4 = \mu_4' - 4\mu_3'\mu + 6\mu_2'\mu^2 - 3\mu^4 \tag{2.33}$$

Table 2.3 Moments of some common distributions of continuous variables.

Name, parameters, domain	Probability density or distribution function	Moments
Uniform in $[a, b]$, $a \le x \le b$	$f(x) = \dfrac{1}{b-a}$	$\mu_1' = \dfrac{a+b}{2}, \quad \mu_2 = \dfrac{(b-a)^2}{12}, \quad \mu_p' = \dfrac{b^{p+1} - a^{p+1}}{(p+1)(b-a)}$
Exponential $\mu > 0, x \ge 0$	$f(x) = \mathrm{e}^{-x/\mu}/\mu$	$\mu_1' = \mu, \qquad \mu_2 = \mu^2, \qquad \mu_p' = p!\,\mu^p$
Gamma $\zeta > 0, \lambda > 0, x \ge 0$	$f(x) = \dfrac{(x/\lambda)^{\zeta-1}\mathrm{e}^{-x/\lambda}}{\lambda\,\Gamma(\zeta)}$	$\mu_1' = \zeta\lambda, \qquad \mu_2 = \zeta\lambda^2, \qquad \mu_p' = \dfrac{\Gamma(p+\zeta)}{\Gamma(\zeta)}\lambda^p$
Weibull $\zeta > 0, \lambda > 0, x \ge 0$	$F(x) = 1 - \exp\left(-\left(\dfrac{x}{\lambda}\right)^{\zeta}\right)$	$\mu_1' = \Gamma\left(1 + \dfrac{1}{\zeta}\right)\lambda, \quad \mu_2 = \left(\Gamma\left(1 + \dfrac{2}{\zeta}\right) - \Gamma\left(1 + \dfrac{1}{\zeta}\right)^2\right)\lambda^2$ $\mu_p' = \Gamma\left(1 + \dfrac{p}{\zeta}\right)\lambda^p$
Normal $\mu \in \mathbb{R}, \sigma > 0$, $x \in \mathbb{R}$	$f(x) = \dfrac{\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}$	$\mu_1' = \mu, \qquad \mu_2 = \sigma^2, \qquad \mu_p = \begin{cases} 0, & p \text{ odd} \\ \sigma^p(p-1)!!, & p \text{ even} \end{cases}$
Lognormal ($\ln \underline{x}$ is $\mathrm{N}(\ln\lambda, \sigma)$) $\sigma > 0, \lambda > 0, x \ge 0$	$f(x) = \dfrac{\exp\left(-\dfrac{1}{2\sigma^2}\left(\ln\left(\dfrac{x}{\lambda}\right)\right)^2\right)}{\sqrt{2\pi}\,\sigma x}$	$\mu_1' = \mathrm{e}^{\frac{\sigma^2}{2}}\lambda, \qquad \mu_2 = \mathrm{e}^{\frac{\sigma^2}{2}}\left(\mathrm{e}^{\frac{\sigma^2}{2}} - 1\right)\lambda^2, \qquad \mu_p' = \mathrm{e}^{\frac{p\sigma^2}{2}}\lambda^p$
Pareto[1] $\xi > 0, \lambda > 0, x \ge 0$	$F(x) = 1 - \left(1 + \xi\dfrac{x}{\lambda}\right)^{-\frac{1}{\xi}}$	$K_1' = \dfrac{\lambda}{1-\xi}, \qquad \mu_2 = \dfrac{\lambda^2}{(1-\xi)^2(1-2\xi)}$ $\mu_p' = \dfrac{p\mathrm{B}\left(\frac{1}{\xi} - p, p\right)}{\xi^p}\lambda^p$
Pareto-Burr-Feller (PBF)* $\zeta > 0, \xi > 0$, $\lambda > 0, x \ge 0$	$F(x) = 1 - \left(1 + \zeta\xi\left(\dfrac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\xi\zeta}}$	$\mu_p' = \dfrac{p\mathrm{B}\left(\frac{1}{\xi\zeta} - \frac{p}{\zeta}, \frac{p}{\zeta}\right)}{\zeta(\xi\zeta)^{\frac{p}{\zeta}}}\lambda^p$
Dagum* $\zeta > 0, \xi > 0$, $\lambda > 0, x \ge 0$	$F(x) = \left(1 + \dfrac{\xi}{\zeta}\left(\dfrac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)^{-\zeta}$	$\mu_p' = \zeta\left(\dfrac{\xi}{\zeta}\right)^{\xi p} p\mathrm{B}(1 - \xi p, \xi p + \zeta)\lambda^p$
Extreme value type I (EV1) $\lambda > 0, x \in \mathbb{R}$	$F(x) = \exp\left(-\mathrm{e}^{-\frac{x}{\lambda}}\right)$	$\mu_1' = \gamma\lambda, \qquad \mu_2 = \dfrac{\pi^2\lambda^2}{6}, \qquad \kappa_p = (p-1)!\,\zeta(p)\lambda^p$
Extreme value type II (EV2)* $\xi > 0, \lambda > 0, x \ge 0$	$F(x) = \exp\left(-\xi\left(\dfrac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)$	$\mu_1' = \xi^{\xi}\Gamma(1-\xi)\lambda, \mu_2 = \xi^{2\xi}(\Gamma(1-2\xi) - \Gamma(1-\xi)^2)\lambda^2$ $\mu_p' = \xi^{p\xi}\Gamma(1 - p\xi)\lambda^p$

[1] The moments exist (have finite values) only for order $p < 1/\xi$; for larger p they are infinite.

For ready reference, Table 2.3 provides the analytical expressions of the moments of some common distribution functions.

Another useful expectation is formed by choosing $g(\underline{x}) = e^{t\underline{x}}$ for any t. The logarithm of the resulting expectation is called the *cumulant generating function*:

$$K(t) \coloneqq \ln \mathrm{E}[e^{t\underline{x}}] \tag{2.34}$$

The power series expansion of the cumulant generating function i.e.:

$$K(t) = \sum_{p=1}^{\infty} \kappa_p \frac{t^p}{p!} \tag{2.35}$$

defines the *cumulants* κ_p. These are related to noncentral moments of similar order by (Smith, 1995):

$$\mu'_p = \sum_{i=0}^{p-1} \binom{p-1}{i} \kappa_{p-i} \mu'_i, \qquad \kappa_p = \mu'_p - \sum_{i=1}^{p-1} \binom{p-1}{i} \kappa_{p-i} \mu'_i \tag{2.36}$$

For small p they take the following forms:

$$\kappa_0 = \mu_1 = 0, \qquad \kappa_1 = \mu'_1 = \mu, \qquad \kappa_2 = \sigma^2, \qquad \kappa_3 = \mu_3, \qquad \kappa_4 = \mu_4 - \sigma^4 \tag{2.37}$$

The importance of cumulants results from their homogeneity and additivity properties. Namely, for a stochastic variable that is the weighted sum of r independent variables $\underline{v}_i$, i.e., $\underline{x} = a_1\underline{v}_1 + \cdots + a_r\underline{v}_r$, the cumulants of $\underline{x}$ are given as

$$\kappa_p = a_1^p \, \kappa_p^{(v_1)} + \cdots + a_r^p \, \kappa_p^{(v_r)} \tag{2.38}$$

where $\kappa_p^{(v_l)}$ is pth cumulant of $\underline{v}_l$. This property is quite useful in stochastic simulation (see Chapter 7).

Digression 2.E: Illustration of the first four moments and statistical characteristics other than moments

The geometrical meaning of the four first moments is visualized in Figure 2.2. Essentially, the first moment, i.e. the mean, describes the location of the centre of gravity of the shape defined by the probability density function and the horizontal axis (Figure 2.2a). It is also equivalent with the static moment of this shape about the vertical axis (given that the area of the shape equals 1). Often, the following quantities are alternatively used as location parameters:

- The *mode*, or most probable value, x_{p}, is the value of x for which the density $f(x)$ becomes maximum, if the stochastic variable is continuous, or, for discrete variables, the probability mass becomes maximum. If $f(x)$ has one, two or many maxima, we say that the distribution is unimodal, bi-modal or multi-modal, respectively.
- The *median*, $x_{0.5}$, is the value for which $P\{\underline{x} \le x_{0.5}\} \ge 1/2$ and $P\{\underline{x} \ge x_{0.5}\} \ge 1/2$. Thus, for a continuous stochastic variable, a vertical line at the median separates the graph of the density function into two equivalent parts each having an area of $1/2$.

Generally, the mean, the mode and the median are not identical unless the density has a symmetrical and unimodal shape.

The variance of a stochastic variable and its square root, the standard deviation, which has same dimensions as the stochastic variable, describe a measure of the scatter or dispersion of the probability density around the mean. Thus, a small variance shows a concentrated distribution

(Figure 2.2b). The variance cannot be negative, as the lowest possible value is zero. This corresponds to a variable that takes one value only (the mean) with absolute certainty. Geometrically, the variance is equivalent with the moment of inertia about the vertical axis passing from the centre of gravity of the shape defined by the probability density function and the horizontal axis.

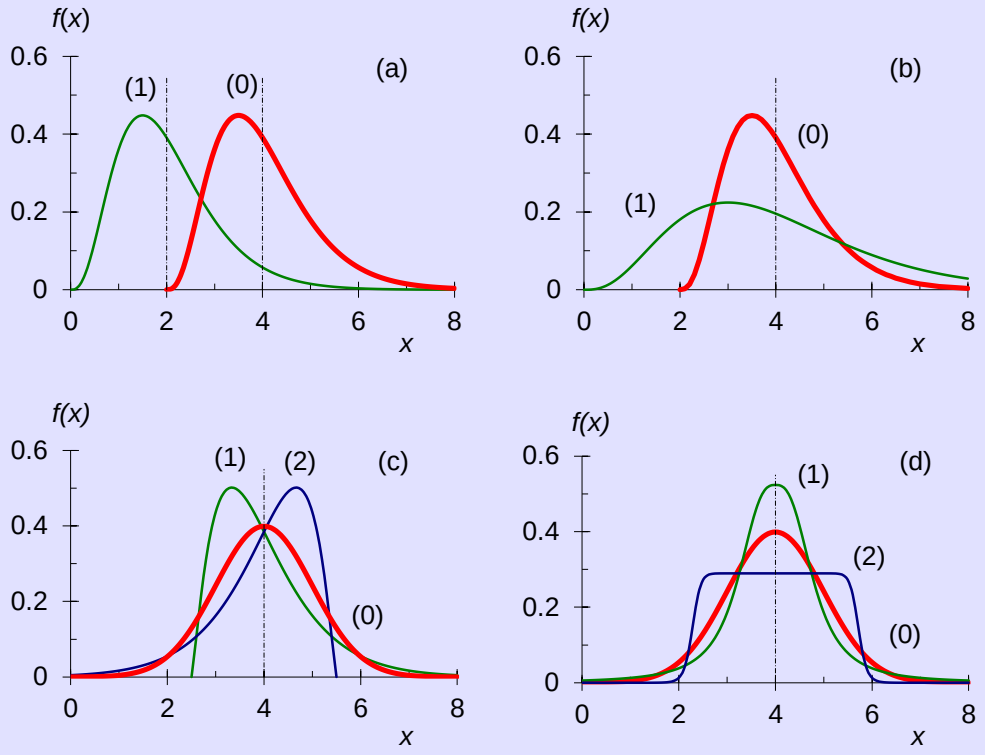

Figure 2.2 Graphical illustration of the geometrical interpretation of moments of a stochastic variable: **(a)** *Effect of the mean.* Curves (0) and (1) have means 4 and 2, respectively, whereas they both have standard deviation 1, coefficient of skewness 1 and coefficient of kurtosis 4.5. **(b)** *Effect of the standard deviation.* Curves (0) and (1) have standard deviation 1 and 2 respectively, whereas they both have mean 4, coefficient of skewness 1 and coefficient of kurtosis 4.5. **(c)** *Effect of the coefficient of skewness.* Curves (0), (1) and (2) have coefficients of skewness 0, +1.33 and −1.33, respectively, but they all have mean 4 and standard deviation 1; their coefficients of kurtosis are 3, 5.67 and 5.67, respectively. **(d)** *Effect of the coefficient of kurtosis.* Curves (0), (1) and (2) have coefficients of kurtosis 3, 5 and 2, respectively, whereas they all have mean 4, standard deviation 1 and coefficient of skewness 0.

Alternative measures of dispersion are provided by the so-called interquartile range, defined as the difference $x_{0.75} - x_{0.25}$, i.e. the difference of the 0.75 and 0.25 quantiles (or upper and lower quartiles) of the stochastic variable (they define an area in the density function equal to 0.5).

The third central moment is used as measures of skewness. A zero value indicates that the density is symmetric. This can be easily verified from the definition of the third central moment. If the third central is positive or negative, we say that the distribution is positively or negatively skewed respectively (Figure 2.2c). In a positively skewed unimodal distribution, the following inequality holds true: $x_{\mathrm{p}} \leq x_{0.5} \leq \mu$; the reverse holds for a negatively skewed distribution.

The fourth central moment is used as measure of kurtosis, a term which describes the "peakedness" of the probability density function around its mode. A reference value for kurtosis is provided by the normal distribution, which has $C_{\mathrm{k}} = 3$. Distributions with kurtosis greater than the reference value are called *leptokurtic* (acute, sharp) and have typically fat tails (see below), so that more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations. Distributions with kurtosis less than the reference values are called *platykurtic* (flat; Figure 2.2d).

2.9 Definition and importance of entropy

The enumeration of the basic set and hence the definition of a stochastic variable entails arbitrary choices and one could think of different options. In turn, expectations and moments depend on the option chosen. One may think of defining the function $g(\,)$ whose expectation is sought, in terms of the probability per se, i.e. $g(\underline{x}) = h(P(\underline{x}))$ for a discrete variable or $g(\underline{x}) = h\left(f(\underline{x})\right)$ for a continuous variable, where $h(\,)$ is any specified function. Among the several choices of $h(\,)$, most useful is the logarithmic function, which results in the definition of entropy. The emergence of the logarithm in the definition of entropy follows some postulates originally set up by Shannon (1948). Assuming a discrete stochastic variable $\underline{x}$ taking on values x_j with probability mass function $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, $j = 1,\ldots,J$, which satisfies the obvious relationship:

$$\sum_{j=1}^{J} P_j = 1 \tag{2.39}$$

the postulates, as reformulated by Jaynes (2003, p. 347), are:

(a) It is possible to set up a numerical measure Φ of the *amount of uncertainty* which is expressed as a real number.
(b) Φ is a continuous function of P_j.
(c) If all the P_j are equal ($P_j = 1/J$) then Φ should be a monotonic increasing function of J.
(d) If there is more than one way of working out the value of Φ, then we should get the same value for every possible way.

Quantification of postulate (d) is given, among others, in Robertson (1993, p. 3) and Uffink (1995; theorem 1), and is related to refinement of partitions to which the probabilities P_j refer.

From these general postulates about uncertainty, a unique (within a multiplicative factor) function Φ results, which serves as the definition of entropy:

$$\Phi[\underline{x}] \coloneqq \mathrm{E}[-\ln P(\underline{x})] = -\sum_{j=1}^{J} P_j \ln P_j \tag{2.40}$$

Shannon's work leading to the above definition was on information theory, but followed the works of Boltzmann and Gibbs in thermodynamics. Additional notes on the historical evolution of the entropy concept are given in Digression 2.F. We note that in classical thermodynamics, entropy is denoted by S (the original symbol used by Clausius), while probability texts use the symbol H. Here Φ was preferred as a unifying symbol for information and thermodynamic entropy, under the interpretation that the two are

essentially the same thing* (see Koutsoyiannis, 2013a, 2014a, even though others have different opinion).

Extension of the above definition for the case of a continuous stochastic variable $\underline{x}$ with probability density function $f(x)$, where:

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1 \tag{2.41}$$

is possible, although not contained in Shannon's (1948) original work. This extension involves some additional difficulties. Specifically, if we discretize the domain of x into intervals of size δx, then (2.40) would give an infinite value for the entropy as δx tends to zero (the quantity $-\ln P = -\ln(f(x)\,\delta x)$ will tend to infinity). However, if we involve a (so-called) *background measure* with density $\beta(x)$ and take the ratio $(f(x)\delta x)/(\beta(x)\delta x) = f(x)/\beta(x)$, then the logarithm of this ratio will generally converge. This allows the definition of entropy for continuous variables as (see e.g. Jaynes, 2003, p. 375, Uffink, 1995):

$$\Phi[\underline{x}] := \mathrm{E}\left[-\ln\frac{f(\underline{x})}{\beta(\underline{x})}\right] = -\int_{-\infty}^{\infty} \ln\frac{f(x)}{\beta(x)} f(x)\mathrm{d}x \tag{2.42}$$

The background measure $\beta(x)$ can be any probability density, proper (with integral equal to 1, as in (2.41)) or improper (meaning that its integral diverges); typically it is an (improper) Lebesgue density, i.e. a constant with dimensions $[\beta(x)] = [f(x)] = [x^{-1}]$, so that the argument of the logarithm function be dimensionless. It is easily seen that for both discrete and continuous variables the entropy $\Phi[\underline{x}]$ is a dimensionless quantity. For discrete variables it can only take positive values, while for continuous variables it can be either positive or negative, depending on the assumed $\beta(x)$. In contrast to the discrete variables where the entropy for a specified probability mass function is a unique number, in continuous variables the value of entropy depends on the assumed $\beta(x)$.

The importance of the entropy concepts relies in the *principle of maximum entropy* (Jaynes, 1957). This postulates that the entropy of a stochastic variable $\underline{x}$ should be at maximum, under some conditions, formulated as constraints, which incorporate the information that is given about this variable. This principle can be used for *logical inference* as well as for *modelling physical systems*; for example, the tendency of entropy to become maximal (Second Law of thermodynamics), a tendency that is the driving force of natural change, can result from this principle.

Digression 2.F: The meaning of entropy

Entropy is etymologized from the ancient Greek έντροπία (from the verb έντρέπειν, to turn into, to turn about) but was introduced as a scientific term by Rudolf Clausius only in 1865, although the concept appears also in his earlier works (as he describes in Clausius, 1872). The rationale for

* The reasons for this preference include historical ones: for long time, entropy used to be denoted by Φ (Perry, 1903; Swinburne, 1904; Ewing, 1920), and this is still echoed in the term tephigram (T-Φ-gram) used in meteorology.

introducing the term is explained in his own words (Clausius, 1867, p. 358, which indicates that he was not aware of the existence of the word ἐντροπία in ancient Greek):

> *We might call S the transformational content of the body [...]. But as I hold it to be better to borrow terms for important magnitudes from the ancient languages, so that they may be adopted unchanged in all modern languages, I propose to call the magnitude S the entropy of the body, from the Greek word τροπή, transformation. I have intentionally formed the word entropy so as to be as similar as possible to the word energy; for the two magnitudes to be denoted by these words are so nearly allied their physical meanings, that a certain similarity in designation appears to be desirable.*

In addition to its semantic content, this quotation contains a very important insight: the recognition that entropy is related to transformation and change and the contrast between entropy and energy, where the latter is a quantity that is conserved in all changes. This meaning has been more clearly expressed in Clausius' famous aphorism (Clausius, 1865):

> *Die Energie der Welt ist konstant.*
> *Die Entropie der Welt strebt einem Maximum zu.*
> *(The energy of the world is constant.*
> *The entropy of the world strives to a maximum).*

In other words, entropy and its ability to increase (as contrasted to energy and other quantities that are conserved) is the driving force of change. This property of entropy has seldom been acknowledged; instead in common perception entropy is typically identified with disorganization and deterioration as if change can only have negative consequences.

Mathematically, the thermodynamic entropy, S, is defined in the same Clausius' texts through the equation $\mathrm{d}S = \mathrm{d}Q/T$, where Q and T denote heat and temperature. The definition, however, applies to a reversible process only. The fact that in an irreversible process $\mathrm{d}S > \mathrm{d}Q/T$ makes the definition imperfect and affected by circular reasoning, as, in turn, a reversible process is one in which the equation holds.

Two decades later (in 1877) Ludwig Boltzmann (1877; see also Swendsen, 2006) gave entropy a statistical content as he related it to probabilities of statistical mechanical system states, thus explaining the Second Law of thermodynamics as the tendency of the system to run toward more probable states, which have higher entropy. The statistical concept of entropy was advanced later in the works of Gibbs in thermodynamics and von Neumann in quantum mechanics.

Shannon (1948) used an essentially similar, albeit more general, entropy definition to describe the information content, which he also called entropy at von Neumann's suggestion (Robertson, 1993; Koutsoyiannis, 2011b). According to the latter definition, entropy is a probabilistic concept, a measure of information or, equivalently, uncertainty. A decade later, Jaynes (1957) introduced the principle of maximum entropy thus equipping the entropy concept with a powerful tool for logical inference.

More than half a century later, the meaning of entropy is still debated and a diversity of opinion among experts is encountered (Swendsen, 2011). In particular, despite having the same name, probabilistic (or information) entropy and thermodynamic entropy are still regarded by many as two distinct notions having in common only the name. The classical definition of thermodynamic entropy (as above) does not give any hint about similarity with the probabilistic entropy. The fact that the latter is a dimensionless quantity and the former has units (J/K) has been regarded as an argument that the two are dissimilar. Even Jaynes (2003), the founder of the maximum entropy principle, states:

> *We must warn at the outset that the major occupational disease of this field is a persistent failure to distinguish between the information entropy, which is a property of any probability distribution, and the experimental entropy of thermodynamics, which is instead a property of a thermodynamic state as defined, for example by such observed quantities as pressure, volume, temperature, magnetization, of some physical system. They should never have been called by the same name; the experimental entropy makes no reference to any probability distribution, and the*

information entropy makes no reference to thermodynamics. Many textbooks and research papers are flawed fatally by the author's failure to distinguish between these entirely different things, and in consequence proving nonsense theorems.

However, the units of thermodynamic entropy are only an historical accident, related to the arbitrary introduction of temperature scales (Atkins, 2007). In a recent book, Ben-Naim (2008) has attempted to replace altogether the concept of entropy with the concept of information. However, such a replacement is unnecessary or even meaningless if we accept that the two concepts are identical. As has recently been shown (Koutsoyiannis, 2013a, 2014a), the formal probability theory can produce the thermodynamic entropy of gases without difficulties and without the need of strange assumptions (e.g. the indistinguishability of particles). The logical basis of the latter study includes the following points.

- The classical definition of thermodynamic entropy is not necessary; it can be abandoned and replaced by the probabilistic definition.
- The thus defined entropy is the fundamental thermodynamic quantity, which supports the definition of all other derived ones. For example, the temperature is defined as the inverse of the partial derivative of entropy with respect to the internal energy (see Digression 10.D).
- The entropy retains its dimensionless character even in thermodynamics, thus rendering the unit of kelvin an energy unit, a multiple of the joule (i.e., 1 K = 0.138 065 05 yJ = $1.380\ 650\ 5\times10^{-23}$ J).
- The entropy retains its probabilistic interpretation as a measure of uncertainty, leaving aside the traditional but obscure 'disorder' interpretation.
- The tendency of entropy to reach a maximum is the driving force of natural change. This tendency is formalized as the principle of maximum entropy, which can be regarded both as a physical (ontological) principle obeyed by natural systems, as well as a logical (epistemological) principle applicable in making inference about natural systems.

Examples of deductive reasoning in deriving thermodynamic laws from the formal probabilistic principle of maximum entropy have been provided in Koutsoyiannis (2014a). Notable among them is the derivation of the law of phase transition of water (Clausius-Clapeyron equation) by maximizing entropy, i.e. uncertainty, at the microscopic level, yet leading to an expression that is virtually certain at a macroscopic level (see Digression 10.D).

Digression 2.G: Illustration of the principle of maximum entropy

Here we illustrate the maximum entropy (ME) principle in a few simple cases. The examples may look trivial. However, we must have in mind that, as already mentioned in Digression 2.F, we can infer with the same reasoning more interesting things, such as the saturation vapour pressure in the atmosphere (Digression 10.D). The logic is the same: we maximize the uncertainty with respect to state of a die or a water molecule.

(a) We thus start from the simple example of determining the probabilities of the outcomes of a die throw. For the die the entropy is:

$$\Phi = \mathrm{E}[-\ln P(\underline{z})] = -P_1 \ln P_1 - P_2 \ln P_2 - P_3 \ln P_3 - P_4 \ln P_4 - P_5 \ln P_5 - P_6 \ln P_6$$

Considering also the equality constraint:

$$P_1 + P_2 + P_3 + P_4 + P_5 + P_6 = 1$$

we form the objective function to maximize as:

$$A := -P_1 \ln P_1 - P_2 \ln P_2 - P_3 \ln P_3 - P_4 \ln P_4 - P_5 \ln P_5 - P_6 \ln P_6 + a(P_1 + P_2 + P_3 + P_4 + P_5 + P_6 - 1)$$

where a is a Lagrange multiplier. We find the partial derivatives with respect to each of the variables and equate them to zero, obtaining:

$$\frac{\partial A}{\partial P_1} = -1 - \ln P_1 - a = 0\ , \qquad \ldots\ , \qquad \frac{\partial A}{\partial P_6} = -1 - \ln P_6 - a = 0$$

Obviously, the solution of these equations yields the single maximum:

$$P_1 = P_2 = P_3 = P_4 = P_5 = P_6 = 1/6$$

The entropy is $\Phi = -6\ (1/6)\ \ln\ (1/6) = \ln 6$. In general, the entropy for J equiprobable outcomes is:

$$\Phi = \ln J \tag{2.43}$$

It is noted that entropy and information are complementary to each other. When we know (observe) that the outcome is i ($P_i = 1$, $P_j = 0$ for $j \neq i$), the entropy is zero.

In the above case of a fair die throw, the application of the ME principle is equivalent to the *principle of insufficient reason* (attributed to Bernoulli and Laplace). However, while the former is a variational law (equivalent to the solution of an optimization problem), the latter is formulated in terms of equations. A single variational law is always much more powerful than very many equations. Actually, from a variational law we derive as many equations as there are unknowns (even an infinite number of equations). And as we showed, in this case the principle of insufficient reason results from the variational ME principle, and thus there is no need at all to postulate the former as an additional philosophical or scientific principle.

(b) To illustrate that the variational ME principle is more powerful, we consider the following variant of the problem in which uniformity is a priori excluded. Specifically, we assume that the die is loaded and that we have prior information that $P_6 = 2P_1$. What is the probability that the outcome of a die throw will be i in this case? For the ME optimization we only need to take into account the additional constraint, by adding to the objective function the term $b(P_6 - 2P_1)$ where b is another additional Lagrange multiplier. The solution of the optimization problem is a single maximum, $P_2 = P_3 = P_4 = P_5 = P_6 = 0.1698$ (slightly >1/6), $P_1 = 0.1069$, $P_6 = 0.2139$. The entropy is $\Phi = 1.7732$, smaller than in the case of equiprobability, where $\Phi = \ln 6 = 1.792$. The decrease of entropy in the loaded die derives from the additional information incorporated in the constraints.

(c) In another example we consider a roulette wheel which is not divided into pockets but its outcome is a real number measured on a circular scale graded 0 to J. In this case our stochastic variable $\underline{x}$ is of continuous type. Assuming background measure $\beta(\underline{x}) = 1$, the entropy is

$$\Phi[\underline{x}] = -\int_0^J \ln f(x)\, f(x)\mathrm{d}x$$

Considering also the constraint (2.41) with a Lagrange multiplier a, we should maximize:

$$A := -\int_0^J \ln f(x)\, f(x)\mathrm{d}x - a\left(\int_0^J f(x)\mathrm{d}x - 1\right)$$

Finding the partial derivative with respect f and equating it to zero we obtain:

$$\frac{\partial A}{\partial f} = -1 - \ln f - a = 0$$

Hence $f = \exp(-1 - \lambda) =$ constant and from the constraint we obtain that the entropy maximizing density is:

$$f(x) = 1/J \tag{2.44}$$

and the entropy is:

$$\Phi = \ln J \tag{2.45}$$

This is the uniform distribution and for $J = 1$ it takes the form given in Table 2.2. Notice that the expression of maximum entropy for a discrete stochastic variable (equation (2.43)) is identical to that of a continuous stochastic variable (equation (2.45)).

(d) If in the uniform distribution the upper bound J tends to ∞, it becomes improper ($f(x) = 0$). Therefore, in this case we need an additional constraint to find a proper distribution. The simplest one that we can think of is that the distribution has a specified mean μ, i.e.:

$$\int_0^\infty xf(x)\mathrm{d}x = \mu$$

The expression of the entropy is the same as in the example (c), but the objective function to maximize becomes:

$$A \coloneqq -\int_0^\infty \ln f(x)\,\mathrm{f}(x)\mathrm{d}x - a\left(\int_0^\infty f(x)\mathrm{d}x - 1\right) - b\left(\int_0^\infty xf(x)\mathrm{d}x - \mu\right)$$

Thus,

$$\frac{\partial A}{\partial f} = -1 - \ln f - \lambda - bx = 0$$

and

$$f(x) = B\exp(-bx)$$

where from the two constraints we find, after the algebraic operations, that $B = b = 1/\mu$. This is the exponential distribution given in Table 2.2. It is very common in physics, as the mean constraint, from which it results, is omnipresent. For example, if $\underline{x}$ represents the kinetic energy of one of many particles moving in a box, we do not know the exact energy of each particle (which may change due to collisions, assumed to be elastic) but we may know the average μ, which is preserved according to the related physical principle. Consequently, the distribution of the kinetic energy is exponential.

(e) If in the above example of moving particles, we limit the motion on a straight line and we choose as stochastic variable $\underline{x}$ not the kinetic energy but the velocity, which can take on both positive and negative values, the kinetic energy constrained is written as

$$\int_0^\infty x^2 f(x)\mathrm{d}x = \gamma$$

where γ is twice the average kinetic energy. The objective function to maximize becomes:

$$A \coloneqq -\int_0^\infty \ln f(x)\,\mathrm{f}(x)\mathrm{d}x - a\left(\int_0^\infty f(x)\mathrm{d}x - 1\right) - b\left(\int_0^\infty x^2 f(x)\mathrm{d}x - \gamma\right)$$

Thus,

$$\frac{\partial A}{\partial f} = -1 - \ln f - a - bx^2 = 0$$

and

$$f(x) = B\exp(-bx^2)$$

where from the two constraints we find, after the algebraic operations, that $B = \sqrt{2\pi\gamma}$, $b = 1/2\gamma$. This is the normal distribution given in Table 2.2, with $\mu = 0$ and $\sigma = \sqrt{\gamma}$.

2.10 Maximum entropy distributions

In Digression 2.G we illustrated several simple cases of entropy maximization, in which we determined the entire probability mass or density function based on one or two constraints. We can generalize the result for a number of constraints of the form:

$$\mathrm{E}[g_i(x)] = \gamma_i \Leftrightarrow \int_{-\infty}^{\infty} g_i(x) f(x) \mathrm{d}x - \gamma_i = 0 \tag{2.46}$$

and for any background measure β. In this case, after incorporating the constraints to the entropy with Lagrange multipliers, the expression whose maximization is sought is:

$$A \coloneqq \int_{-\infty}^{\infty} \ln \frac{f(x)}{\beta(x)} f(x) \mathrm{d}x - a\left(\int_{-\infty}^{\infty} f(x) \mathrm{d}x - 1 \right) - \sum_i b_i \left(\int_{-\infty}^{\infty} g_i(x) f(x) \mathrm{d}x - \gamma_i \right) \tag{2.47}$$

Taking the partial derivative with respect to f and equating it to zero we find

$$\ln \frac{f(x)}{\beta(x)} + 1 - a - \sum_i b_i g_i(x) = 0 \tag{2.48}$$

and thus, the entropy maximizing density is:

$$f(x) = A\, \beta(x) \exp\left(- \sum_i b_i g_i(x) \right) \tag{2.49}$$

As we have seen in Digression 2.G, some of the most typical distributions which are used in a variety of scientific fields can be derived by entropy maximization using a simple constraint. Here we will try to get a plethora of distributions again using a single constraint but both with a Lebesgue background measure and a generalized one.

The background measure $\beta(x)$ determines the way of measuring the distances d between values of x; the Lebesgue measure corresponds to the Euclidean distance, $d(x, x') = |x - x'|$. However, most hydrometeorological variables are non-negative physical quantities unbounded from above (e.g., precipitation, streamflow, temperature—absolute, expressed in kelvins). In positive physical quantities, often the Euclidean distance is not a proper metric; sometimes we use a logarithmic distance $d(x, x') = |\ln(x'/x)|$, as shown in the example below referring to precipitation depth:

	Euclidean distance	Logarithmic distance
x = 0.1 mm, x' = 0.2 mm	0.1 mm	ln 2
x = 100 mm, x' = 100.1 mm	0.1 mm	ln 1.001
x = 100 mm, x' = 200 mm	100 mm	ln 2

Which of the second and third pairs of points is equidistant to the first one? In an attempt to merge (or unify) the Euclidean and logarithmic distance, we heuristically introduce (see Koutsoyiannis, 2014a) a background measure for nonnegative variables that is based on the hyperbola:

$$\beta(x) = \frac{1}{\lambda + x} \tag{2.50}$$

where λ is a characteristic scale parameter, which also serves as a physical unit for x. We will refer to it as the *hyperbolic background measure* and we note that for $\lambda \to \infty$, it tends to the Lebesgue measure. According to this measure, the distance of any point x from 0 is:

$$B(x) = \int_0^x \lambda\, \beta(s) \mathrm{d}s = \lambda \ln\left(1 + \frac{x}{\lambda}\right) \tag{2.51}$$

An example plot for $B(x)$ is given in Figure 2.3. Its limiting properties are:

$$\lim_{x\to 0} B(x) = \lim_{\lambda\to\infty} B(x) = x, \qquad \lim_{x\to\infty}\left(\frac{B(x)}{\lambda} + \ln\lambda\right) = \lim_{\lambda\to 0}\left(\frac{B(x)}{\lambda} + \ln\lambda\right) = \ln x \tag{2.52}$$

The distance between any two points x and x' is:

$$d(x, x') = |B(x') - B(x)| = \lambda \left|\ln\left(\frac{1 + x'/\lambda}{1 + x/\lambda}\right)\right| \tag{2.53}$$

For small x values, i.e., $x < x' \ll \lambda$, the distance is $d(x, x') = \lambda \ln(1 + (x' - x)/(\lambda + x)) \approx x' - x$ (Euclidean distance). For large values, $\lambda \ll x < x', d(x, x') \approx \lambda \ln(x'/x)$ (logarithmic distance). We notice that both $B(x)$ and $d(x, x')$ have the same units as x (physical consistency).

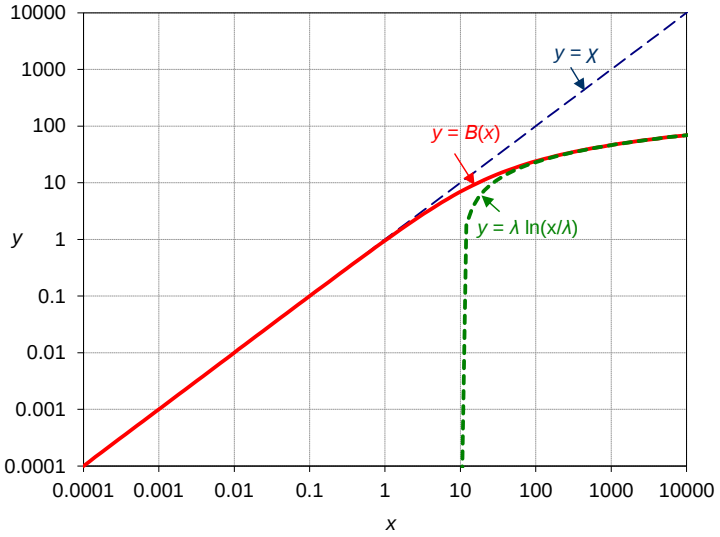

Figure 2.3 Illustration of the distance function $B(x)$; the example plot of $y = B(x)$ is for $\lambda = 10$ and shows that for small x ($< \lambda/10$) $B(x)$ is indistinguishable from x, while for large x ($> 10\lambda$) $B(x)$ becomes a linear function of $\ln x$.

In the general solution (2.49) we use a single constraint for $g(x) \equiv B(x)$, that is $\mathrm{E}[B(x)] = \gamma$, where we have assumed dimensions $[B(x)] = [x] = [\lambda]$. We note that $\beta(x) = B'(x)/\lambda$, where the derivative $B'(x)$ is dimensionless. Thus from (2.49) we get:

$$f(x) = A\frac{B'(x)}{\lambda}\exp\big(-b_1\, B(x)\big) = \frac{A}{\lambda}\exp\left(-b\frac{B(x)}{\lambda} + \ln\big(B'(x)\big)\right) \tag{2.54}$$

where $b = b_1\lambda$. We may notice that all quantities in the big parenthesis are dimensionless. Now we make the following generalizations by raising the following quantities in powers:

$$(x/\lambda) \to (x/\lambda)^c, \qquad B(x)/\lambda \to (B(x)/\lambda)^d, \qquad B'(x) \to \big(B'(x)\big)^e \tag{2.55}$$

and get

$$f(x) = \frac{A}{\lambda}\exp\left(-b\left(\frac{B(x)}{\lambda}\right)^d + e\ln\big(B'(x)\big)\right) \tag{2.56}$$

where

$$B(x) = \lambda\ln\left(1+\left(\frac{x}{\lambda}\right)^c\right), \qquad B'(x) = c\left(\frac{x}{\lambda}\left(1+\left(\frac{x}{\lambda}\right)^{-c}\right)\right)^{-1} \tag{2.57}$$

After the algebraic operations we find the generalized maximum entropy distribution:

$$f(x) = \frac{A'}{\lambda}\left(\frac{x}{\lambda}\right)^{(c-1)e}\left(\left(1+\left(\frac{x}{\lambda}\right)^c\right)\right)^{-e}\exp\left(-b\left(\ln\left(1+\left(\frac{x}{\lambda}\right)^c\right)\right)^d\right) \tag{2.58}$$

where $A' \coloneqq Ae^c$.

As a special case, when $\lambda \to \infty$, the hyperbolic background measure approaches the Lebesgue measure and the quantities in (2.57) become:

$$B(x) = \lambda\left(\frac{x}{\lambda}\right)^c, \qquad B'(x) = c\left(\frac{x}{\lambda}\right)^{c-1} \tag{2.59}$$

Hence, the density of (2.56) becomes

$$f(x) = \frac{A}{\lambda}\exp\left(-b\left(\left(\frac{x}{\lambda}\right)^c\right)^d + e\ln\left(c\left(\frac{x}{\lambda}\right)^{c-1}\right)\right) \tag{2.60}$$

or

$$f(x) = \frac{A'}{\lambda}\left(\frac{x}{\lambda}\right)^{(c-1)e}\exp\left(-b\left(\frac{x}{\lambda}\right)^{cd}\right) \tag{2.61}$$

The densities (2.61) and (2.58) contain as special cases most common distributions used in stochastics, including hydroclimatic stochastics. These special cases are listed in Table 2.4 in terms of their densities $f(x)$ and tail functions $\bar{F}(x) = 1 - F(x)$. In particular, the density (2.61), which is derived from the Lebesgue background measure, corresponds to a generalized gamma distribution also listed in Table 2.4, after suitable transformation of its parameters. The density (2.58), which is derived from the hyperbolic background measure, does not yield a closed expression for $F(x)$ in its general case, and therefore is not listed in Table 2.4. In this case, a sufficiently general form with a closed expression of $F(x)$ is derived if we set $d = 1$; this is listed as the generalized (power transformed) beta prime distribution (where the standard beta prime corresponds to $c = 1$). The generalized gamma and generalized beta prime distributions were also studied in Koutsoyiannis (2005a,c, where additional information for some of their characteristics are provided) and Papalexiou and Koutsoyiannis (2012).

Table 2.4 Special cases of maximum entropy distributions given by equations (2.61) and (2.58).

Name	Parameters	$f(x)$	$\bar{F}(x) = 1 - F(x)$
Lebesgue background measure			
Exponential	$b = c = d = 1$	$\frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right)$	$\exp\left(-\frac{x}{\lambda}\right)$
Gamma[1]	$d = 1/c, b = 1,$ $\zeta = (c-1)e + 1$	$\frac{1}{\lambda\,\Gamma(\zeta)}\left(\frac{x}{\lambda}\right)^{\zeta-1}\exp\left(-\frac{x}{\lambda}\right)$	$\frac{\Gamma_{x/\lambda}(\zeta)}{\Gamma(\zeta)}$
Weibull[2]	$b = d = e = 1$ $\zeta = c$	$\frac{\zeta}{\lambda}\left(\frac{x}{\lambda}\right)^{(\zeta-1)}\exp\left(-\left(\frac{x}{\lambda}\right)^{\zeta}\right)$	$\exp\left(-\left(\frac{x}{\lambda}\right)^{\zeta}\right)$
Half*- normal	$c = 1, d = 2,$ $b = 1/2$	$\frac{2}{\lambda\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x}{\lambda}\right)^{2}\right)$	$1 - \mathrm{erf}\left(\frac{x}{\sqrt{2}\lambda}\right)$
Extended half* normal[3]	$b = d = 1, c = 2,$ $\zeta = (e+1)/2$	$\frac{1}{\lambda\,\Gamma(\zeta)}\left(\left(\frac{x}{\lambda}\right)^{2}\right)^{\zeta-\frac{1}{2}}\exp\left(-\left(\frac{x}{\lambda}\right)^{2}\right)$	$\frac{\Gamma_y(\zeta)}{\Gamma(\zeta)},\ y = \left(\frac{x}{\lambda}\right)^{2}$
Generalized gamma[4]	$\zeta = cd, b = 1,$ $\zeta' = \big((c-1)e+1\big)/cd$	$\frac{1}{\lambda\,\Gamma(\zeta')}\left(\frac{x}{\lambda}\right)^{\zeta\zeta'-1}\exp\left(-\left(\frac{x}{\lambda}\right)^{\zeta}\right)$	$\frac{\Gamma_y(\zeta')}{\Gamma(\zeta')},\ y = \left(\frac{x}{\lambda}\right)^{\zeta}$
Hyperbolic background measure			
Pareto[5]	$b = c = d = 1, \xi = 1/e$ $c = d = e = 1, \xi = 1/b$	$\frac{1}{\lambda\xi}\left(1+\frac{x}{\lambda}\right)^{-\frac{1}{\xi}-1}$	$\left(1+\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}$
Pareto-Burr-Feller (PBF)[6]	$d = e = 1, \xi = 1/bc$	$\frac{1}{\lambda\xi}\left(\frac{x}{\lambda}\right)^{\zeta-1}\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\zeta\xi}-1}$	$\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\zeta\xi}}$
Half lognormal	$c = e = 1, d = 2,$ $b = 1/2$	$\frac{2}{\sqrt{2\pi}\,\lambda}\frac{\exp\left(-\frac{1}{2}\left(\ln\left(1+\frac{x}{\lambda}\right)\right)^{2}\right)}{1+x/\lambda}$	$1 - \mathrm{erf}\left(\frac{1}{\sqrt{2}}\ln\left(1+\frac{x}{\lambda}\right)\right)$
Generalized log-gamma[7]	$b = e = 1$ $\zeta = c$	$\frac{\zeta d}{\Gamma(1/d)\,\lambda}\frac{\exp\left(-\left(\ln\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)\right)^{d}\right)}{\left(\frac{x}{\lambda}\right)^{1-\zeta}\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)}$	$\frac{\Gamma_z(1/d)}{\Gamma(1/d)},$ $z = \left(\ln\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)\right)^{d}$
Half Student[8]	$c = 2, d = 1, e = 0$ $\xi = 1/(b+e-1)$	$\frac{2}{\lambda\,\mathrm{B}\left(\frac{1}{2},\frac{1}{2\xi}\right)}\left(1+\left(\frac{x}{\lambda}\right)^{2}\right)^{-\frac{1}{2}-\frac{1}{2\xi}}$	$\frac{\mathrm{B}_{\frac{y}{1+y}}\left(\frac{1}{2},\frac{1}{2\xi}\right)}{\mathrm{B}\left(\frac{1}{2},\frac{1}{2\xi}\right)},\ y = \left(\frac{x}{\lambda}\right)^{2}$
Half extended Student	$c = 2, d = 1,$ $\zeta = 1/(e+1),$ $\xi = 1/(b+e-1)$	$\frac{2\left(\left(\frac{x}{\lambda}\right)^{2}\right)^{\frac{1}{2\zeta}-\frac{1}{2}}\left(1+\left(\frac{x}{\lambda}\right)^{2}\right)^{-\frac{1}{2\zeta}-\frac{1}{2\xi}}}{\lambda\,\mathrm{B}\left(\frac{1}{2\zeta},\frac{1}{2\xi}\right)}$	$\frac{\mathrm{B}_{\frac{y}{1+y}}\left(\frac{1}{2\zeta},\frac{1}{2\xi}\right)}{\mathrm{B}\left(\frac{1}{2\zeta},\frac{1}{2\xi}\right)},\ y = \left(\frac{x}{\lambda}\right)^{2}$
Generalized beta prime (GBP)[9]	$d = 1, \zeta = c$ $\zeta' = (1/(c-1)e+1),$ $\xi = 1/(b+e-1)$	$\frac{\zeta\left(\frac{x}{\lambda}\right)^{\frac{1}{\zeta'}-1}\left(1+\left(\frac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\zeta\zeta'}-\frac{1}{\zeta\xi}}}{\lambda\,\mathrm{B}\left(\frac{1}{\zeta\zeta'},\frac{1}{\zeta\xi}\right)}$	$\frac{\mathrm{B}_{\frac{y}{1+y}}\left(\frac{1}{\zeta\zeta'},\frac{1}{\zeta\xi}\right)}{\mathrm{B}\left(\frac{1}{\zeta\zeta'},\frac{1}{\zeta\xi}\right)},\ y = \left(\frac{x}{\lambda}\right)^{\zeta}$

Note: Distributions named "half" have their "full" version whose density $f(x)$ and exceedance $\bar{F}(x)$ is obtained by dividing those given in the table by 2. The "half" version given here corresponds to $x \geq 0$, while the "full" version is supported on the whole real line, except for the full lognormal distribution in which $x \geq -\lambda$.

[1] Special cases: *Chi-squared* and *Erlang*.

[2] Special case: *Rayleigh*. Antisymmetric case (in which $F(x) \leftarrow \bar{F}(x)$): *Fréchet*.

[3] Also known as *Chi*.

[4] Special cases: *Maxwell–Boltzmann*, *Maxwell–Jüttner*, *Nakagami*. Antisymmetric cases: *Inverse-chi-squared*, *Inverse-gamma*, *Lévy*.

[5] More precisely, *Pareto II* or *Lomax*.

[6] Also known as *Pareto III and IV*, *Burr XII* and *Feller*. Antisymmetric case: *Log-logistic*.

[7] For $d = 1$ becomes *PBF* with tail index $\xi = 1/c$. For $d > 1$, $\xi = 0$ (all moments exist). For $d < 1$, $\xi = \infty$ (no moment exists).

[8] Also known as *Tsallis* or *1-particle kappa distribution* (Olbert, 1968; Livadiotis and McComas 2013). Special case: Cauchy.

[9] Special cases: Beta prime, F. Antisymmetric case: *Dagum*—often referred to in hydrology as the kappa distribution (Mielke, 1973; Mielke and Johnson, 1973; Hosking, 1994) but it is totally different from the kappa distribution of footnote 8.

The distributions and the special cases resulting from equations (2.61) and (2.58) correspond to nonnegative stochastic variables, $\underline{x} \geq 0$. However, in some of the cases, in which the variable x appears in Table 2.4 raised to power 2, the extension to the whole real line is direct. The distributions of this type are earmarked as "half" in the table, and their "full" versions (valid for all real numbers) are derived by dividing the expressions given in the table by 2; this case includes the normal and Student distributions.

2.11 Tails, heavy-tailed and light-tailed distributions

There is a substantial difference between the distributions corresponding to equation (2.61) on the one hand and (2.58) on the other hand. Specifically, the former are light-tailed and latter heavy-tailed. In heavy-tailed distributions for any $t > 0$ (however small) the following limit diverges to infinity:

$$\lim_{x\to\infty} \mathrm{e}^{tx}\bar{F}(x) = \infty \tag{2.62}$$

In turn, a heavy-tailed distribution is characterized by the so-called *tail index* (or, in case of ambiguity, *higher-tail index*), defined to be that number ξ for which the following asymptotic equation holds true:

$$\lim_{x\to\infty} x^{1/\xi}\bar{F}(x) = l \tag{2.63}$$

where l is nonzero and finite. Those distributions listed in Table 2.4 under the title *Hyperbolic background measure* are heavy tailed. Those of them in which a parameter ξ appears have tail index ξ (e.g. Pareto, Pareto-Burr-Feller). The remaining (e.g. lognormal) have tail index zero (except a specific case of the generalized log-gamma, shown in the table footnotes, whose tail index is infinite). At the same time, the moments of heavy-tailed distributions also diverge beyond a certain order, i.e., $\mathrm{E}[\underline{x}^p] = \infty$ for all $p > 1/\xi$, where ξ is the tail index. The distributions with zero tail index, such as the lognormal distribution, have all their moments finite. For that reason, they are often regarded as light-tailed. However, the lognormal distribution clearly satisfies (2.62) and therefore according to this definition is heavy tailed.

In a similar manner we can define a *lower-tail index* whenever the domain of the distribution is the entire line of real numbers (we must replace ∞ with –∞ and $x^{1/\xi}$ with $|x|^{1/\xi}$). However, usually we deal with nonnegative quantities and, in this case, we need a different manner to define the lower-tail index. Specifically, the lower-tail index is that number ζ for which the following asymptotic equation holds true:

$$\lim_{x\to 0} x^{-\zeta}F(x) = l' \tag{2.64}$$

where l' is again nonzero and finite*. Those distributions listed in Table 2.4 in which a parameter ζ appears have lower-tail index ζ (e.g. Gamma, Weibull, Pareto-Burr-Feller).

* It would be more natural to use $1/\zeta$ instead of ζ in (2.64) so that it be more consistent with (2.63). However, we used that convention in order for the parameterization of common distributions, such as Gamma and Weibull, to be similar to the one dominating in the statistical literature.

Using l'Hôpital's rule, we see that $\lim_{x\to 0} x^{-\zeta}F(x) = \lim_{x\to 0} x^{1-\zeta}f(x)/\zeta$ and thus if, $\zeta < 1$, the density $f(x)$ should necessarily be a decreasing function, at least close to the origin, with $\lim_{x\to 0} f(x) = \infty$. In contrast when $\zeta > 1$, the density $f(x)$ is increasing close to the origin, with $f(0) = 0$, and is usually bell-shaped. The particular case $\zeta = 1$ is characteristic of the exponential and Pareto distributions, in which $f(0)$ is finite and the density $f(x)$ is a decreasing function.

Later we will discuss how both indices ξ and ζ can be visualized in a probability plot (see Digression 5.A).

Digression 2.H: The hydrometeorological importance of heavy-tailed distributions

In classical statistical mechanics the Lebesgue measure is used as background distribution. As a consequence, a constrained mean results in exponential distribution which notably has coefficient of variation σ/μ = 1. However, in several hydrometeorological processes, most notably rainfall, when the time scale is small (e.g. daily or hourly) the empirical σ/μ is greater than one, which means that the exponential distribution is not suitable. One may think that adding one more constraint would fix the problem. The natural choice seems to be to constrain entropy maximization by both the mean μ and the variance σ^2. However, this does not work as for nonnegative stochastic variable entropy maximization with Lebesgue background measure cannot yield $\sigma/\mu > 1$. In other words, the exponential distribution is the upper limit.

The next solution to try is either to use a trickier (less natural) constraint, to change the definition of entropy (using a generalized definition) or to change the background measure. The first two cases have been studied in Koutsoyiannis (2005a) and Papalexiou and Koutsoyiannis (2012) and the last one in Koutsoyiannis (2017). Whatever the choice may be, the result is practically the same: a heavy-tailed distribution. The easiest way to derive that distribution is by the framework described above, using the hyperbolic background measure and a single constraint, the mean of the distance function. The resulting Pareto distribution has $\sigma/\mu = 1/\sqrt{1-2\xi} > 1$.

In other words, by changing the background measure from Lebesgue to hyperbolic, the light-tailed exponential distribution changes to the heavy-tailed Pareto one. The theoretical framework otherwise remains unaffected—the same probabilistic definition of entropy is used in both cases. But the change in the derived distribution may have important consequences in the design and management related to extreme events. To illustrate this based on real world data we use the daily rainfall data of Bologna, a data set already studied in section 1.3.

During the 206 years of observations there were 19 426 rain days, all of which are used in the modelling. The nonzero rainfall depths of all 19 426 days are plotted against their empirical return periods in Figure 2.4. Following the initial discussion of the concept of return period in section 1.5, the return period of an observed value x is related to the probability of exceedance by $T(x) = D/\bar{F}(x)$, where D = 1 d. More accurate and detailed discussion of return period will be provided in Chapter 5.

The 19 426 values range between 0.1 and 155.7 mm, with a mean of 7.2 mm. In the exponential distribution the single parameter λ equals the mean, which allows plotting the theoretical curve corresponding to it in Figure 2.4. Clearly, the comparison with the empirical points of the figure indicates a bad performance of the exponential model. In contrast, the Pareto model, also plotted in Figure 2.4 looks suitable. It is admirable that a model with only two parameters (the tail index ξ and the scale parameter λ) can make such a good fitting on so many observations of 206 years. The parameter values, ξ = 0.11 and λ = 7.78 mm), have been estimated by a least squares method to minimize the error between the empirical and theoretical return period. The empirical return period has been assigned by the method described in section 5.6). The good performance of the

Pareto distribution suggests that the hypothesis of a hyperbolic background measure, along with the principle of maximum entropy, leads to a good predictive capacity.

Now, comparing the behaviour of the light-tailed exponential distribution with the heavy-tailed Pareto distribution, and both with the empirical distribution, we clearly see that the former underestimates severely the magnitude of the extremes. Notably, for a return period of 10 000 years, which is typically used in the engineering design of major projects such as dams, Figure 2.4 shows that the exponential distribution predicts a rainfall depth of ~100 mm, a value that was in fact exceeded seven times in the 206-year available record. On the other hand, the Pareto distribution predicts a value of ~250 mm, 2.5 times higher (and as we will see in Digression 6.E, it becomes even higher if we also take into account the dependence structure of rainfall). Thus, inappropriate model choice, based on inappropriate theoretical considerations, may have substantial consequences in practical application. Sooner or later, nature per se will reveal the inappropriateness (e.g. by frequent exceedances of the design values). In such cases, one could re-examine the theory (even though an alternative more popular practice is to blame external agents and find good scapegoats).

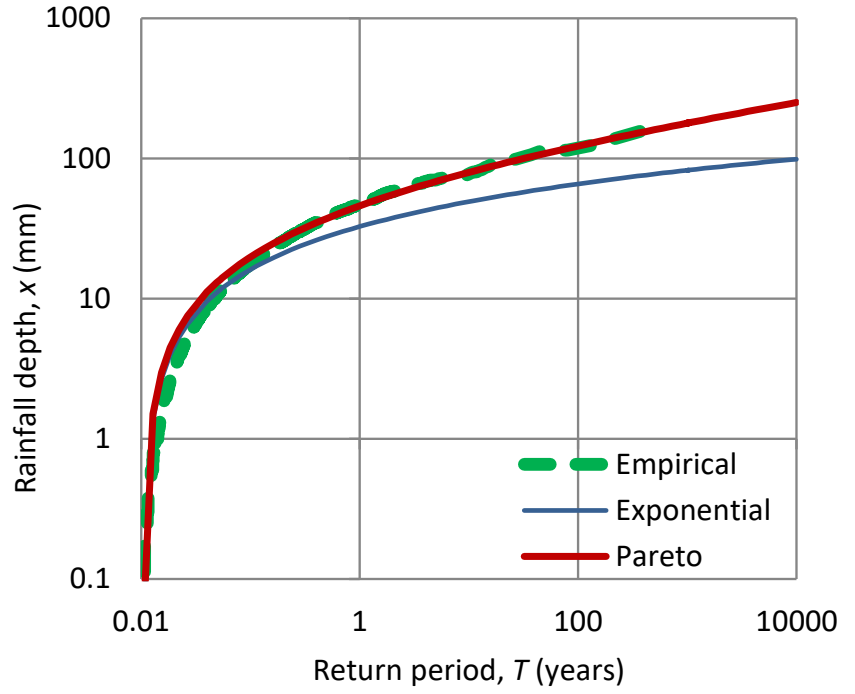

Figure 2.4 Rainfall depth vs. return period for Bologna based on 19 426 daily rainfall depths observed throughout the observation period of 206 years.

Indeed, in the 20th century, the light-tailed distributions constituted the dominant theoretical model in research and engineering practice. And given the substantial underestimation of extremes by this model, its failure (and its severe consequences) should not be regarded a surprise. By now, both theoretical advances and accumulated empirical evidence have shaken this model and pointed to heavy-tailed distributions. More details will be provided in Digression 8.F.

In addition, Koutsoyiannis (2004a, 2005a, 2007) discussed several theoretical reasons that favour the heavy tailed distributions over the exponential case, which are consistent to the above discourse related to the hyperbolic background measure. Furthermeore, the already discussed (Chapter 1) omnipresence of change and the non-static climate are consistent with heavy-tailed distributions, as will be further illustrated in Digression 3.H.

2.12 Two variables: joint distribution and joint moments

The above sections have been devoted to concepts of probability pertaining to the analysis of a single variable $\underline{x}$. Often, however, the simultaneous modelling of two (or more)

variables is necessary. Let the pair of stochastic variables $(\underline{x}, \underline{y})$ be defined on two basic sets (Ω_x, Ω_y), respectively. The intersection (simultaneous occurrence) of the two events $\{\underline{x} \le x\}$ and $\{\underline{y} \le y\}$, denoted as $\{\underline{x} \le x, \underline{y} \le y\}$ is an event of the sample space $\Omega_{xy} = \Omega_x \times \Omega_y$. Based on the latter event, we can define the *joint probability distribution function* of $(\underline{x}, \underline{y})$ as a function of the real variables (x, y):

$$F_{xy}(x, y) \coloneqq P\{\underline{x} \le x, \underline{y} \le y\} \tag{2.65}$$

The subscripts xy can be omitted if there is no risk of ambiguity.

If F_{xy} is differentiable, then the function:

$$f_{xy}(x, y) \coloneqq \frac{\partial^2 F_{xy}(x, y)}{\partial x \, \partial y} \tag{2.66}$$

is the *joint probability density function* of the two variables. Obviously, the following equation holds:

$$F_{xy}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{xy}(\xi, \omega) \mathrm{d}\omega \mathrm{d}\xi \tag{2.67}$$

The functions:

$$F_x(x) \coloneqq P\{\underline{x} \le x\} = \lim_{y \to \infty} F_{xy}(x, y), \qquad F_y(y) \coloneqq P\{\underline{y} \le y\} = \lim_{x \to \infty} F_{xy}(x, y) \tag{2.68}$$

are called the *marginal probability distribution functions* of $\underline{x}$ and $\underline{y}$, respectively. Also, the *marginal probability density functions* can be defined, from

$$f_x(x) = \int_{-\infty}^{\infty} f_{xy}(x, y) \mathrm{d}y, \qquad f_y(y) = \int_{-\infty}^{\infty} f_{xy}(x, y) \mathrm{d}x \tag{2.69}$$

Similar to the univariate case, we can define the expected value of any given function $g\left(\underline{x}, \underline{y}\right)$ of the stochastic variables $\underline{x}$ and $\underline{y}$ by

$$\mathrm{E}\left[g\left(\underline{x}, \underline{y}\right)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \; f_{xy}(x, y) \mathrm{d}x \mathrm{d}y \tag{2.70}$$

In this manner, we define the (*noncentral* or *about the origin*) *joint moment* of orders p, q as:

$$\mu'_{pq} \coloneqq \mathrm{E}\left[\underline{x}^p \underline{y}^q\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \, f_{xy}(x, y) \mathrm{d}x \mathrm{d}y \tag{2.71}$$

as well as the *joint central moment* of orders p, q:

$$\mu_{pq} \coloneqq \mathrm{E}\left[\left(\underline{x} - \mu_x\right)^p \left(\underline{y} - \mu_y\right)^q\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\underline{x} - \mu_x\right)^p \left(\underline{y} - \mu_y\right)^q f_{xy}(x, y) \mathrm{d}x \mathrm{d}y \tag{2.72}$$

If $p = 0$ or $q = 0$, we get the *marginal moments* (e.g. means, $\mu_x \coloneqq \mu'_{10}, \mu_y \coloneqq \mu'_{01}$; variances, $\mathrm{var}[\underline{x}] \coloneqq \mathrm{E}\left[(\underline{x} - \mu_x)^2\right] \equiv \mu_{20} \equiv \gamma_x \equiv \sigma_x^2$, $\mathrm{var}\left[\underline{y}\right] = \mu_{02} \equiv \gamma_y \equiv \sigma_y^2$, etc.). The lowest-order joint central moment is the *covariance*:

$$\mathrm{cov}\left[\underline{x}, \underline{y}\right] \coloneqq \mathrm{E}\left[(\underline{x} - \mu_x)\left(\underline{y} - \mu_y\right)\right] \equiv \mu_{11} \equiv \sigma_{xy} = \mathrm{E}\left[\underline{x}\,\underline{y}\right] - \mathrm{E}[\underline{x}]\mathrm{E}\left[\underline{y}\right] \tag{2.73}$$

A dimensionless index derived from covariance is the *correlation coefficient*:

$$r_{xy} \coloneqq \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{2.74}$$

It is common knowledge (and easy to prove*) that $-1 \leq r_{xy} \leq 1$, where the values –1 and 1 indicate *fully anti-correlated* (*fully negatively correlated*) and *fully (positively) correlated* variables. The particular case where:

$$\sigma_{xy} = r_{xy} = 0 \Leftrightarrow \mathrm{E}\left[\underline{x}\,\underline{y}\right] = \mathrm{E}[\underline{x}]\mathrm{E}\left[\underline{y}\right] \tag{2.75}$$

defines *uncorrelated variables*. The *independent variables* are necessarily uncorrelated, but independence is a stricter concept whose definition is:

$$F_{xy}(x, y) = F_x(x)F_y(y), \qquad f_{xy}(x, y) = f_x(x)f_y(y) \tag{2.76}$$

The joint entropy is defined in an analogous manner with that in the univariate case (section 2.9). For discrete stochastic variables the entropy is:

$$\Phi\left[\underline{x}, \underline{y}\right] \coloneqq \mathrm{E}\left[-\ln P\left(\underline{x}, \underline{y}\right)\right] = -\sum_{i,j} P_{ij} \ln P_{ij} \tag{2.77}$$

where $P_{ij} \coloneqq P\left\{\underline{x} = x_i, \underline{y} = y_j\right\}$. For continuous stochastic variables it is:

$$\Phi\left[\underline{x}, \underline{y}\right] \coloneqq \mathrm{E}\left[-\ln \frac{f\left(\underline{x}, \underline{y}\right)}{\beta\left(\underline{x}, \underline{y}\right)}\right] = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ln \frac{f(x, y)}{\beta(x, y)} f(x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{2.78}$$

2.13 Conditional densities and expectations

Of particular interest are the so-called *conditional probability distribution function* and *conditional probability density function* of $\underline{x}$ for a specified value of $\underline{y} = y$; these are given by:

---

* From the obvious $\mathrm{E}\left[\left(\underline{x} + \underline{y}\right)^2\right] = \mathrm{E}[\underline{x}^2] + \mathrm{E}\left[\underline{y}^2\right] + 2\mathrm{E}\left[\underline{x}\,\underline{y}\right]$, observing that the two sides are nonnegative quantities and assuming, without loss of generality, $E[\underline{x}] = E\left[\underline{y}\right] = 0$, so that $\mathrm{E}[\underline{x}^2] = \sigma_x^2, \mathrm{E}\left[\underline{y}^2\right] = \sigma_y^2$, $\mathrm{E}\left[\underline{x}\,\underline{y}\right] = \sigma_{xy}$, we get $\sigma_x^2 + \sigma_y^2 + 2\sigma_{xy} \geq 0$ or $\sigma_{xy}/\sigma_x\sigma_y \geq -(1/2)\left(\sigma_x/\sigma_y + \sigma_y/\sigma_x\right) = -(1/2)(a + 1/a)$, where $a \coloneqq \sigma_x/\sigma_y > 0$. It is easy to see that $(a + 1/a)$ has minimum value 2, so that $\sigma_{xy}/\sigma_x\sigma_y \geq -1$. Furthermore, $\mathrm{E}\left[\left(\underline{x} - \underline{y}\right)^2\right] = \mathrm{E}[\underline{x}^2] + \mathrm{E}\left[\underline{y}^2\right] - 2\mathrm{E}\left[\underline{x}\,\underline{y}\right]$ and, likewise, $\sigma_{xy}/\sigma_x\sigma_y \leq (1/2)\left(\sigma_x/\sigma_y + \sigma_y/\sigma_x\right) = (1/2)(a + 1/a) \leq 1$.

$$F_{x|y}(x|y) \coloneqq \frac{\int_{-\infty}^{x} f_{xy}(\xi, y)\mathrm{d}\xi}{f_y(y)}, \qquad f_{x|y}(x|y) \coloneqq \frac{f_{xy}(x,y)}{f_y(y)} \tag{2.79}$$

respectively. Switching $\underline{x}$ and $\underline{y}$ we obtain the conditional functions of $\underline{y}$.

The *conditional expected value* of a function $g(\underline{x})$ for a specified value of $\underline{y} = y$ is defined by

$$\mathrm{E}[g(\underline{x})|y] \coloneqq \mathrm{E}\left[g(\underline{x})\middle|\underline{y} = y\right] = \int_{-\infty}^{\infty} g(x) f_{x|y}(x|y)\mathrm{d}x \tag{2.80}$$

An important quantity of this type is the conditional expected value of $\underline{x}$:

$$\mathrm{E}[\underline{x}|y] \coloneqq \mathrm{E}\left[\underline{x}\middle|\underline{y} = y\right] = \int_{-\infty}^{\infty} x f_{x|y}(x|y)\mathrm{d}x \tag{2.81}$$

Likewise, the conditional variance is

$$\mathrm{var}[\underline{x}|y] \coloneqq \mathrm{E}\left[\left(\underline{x} - \mathrm{E}[\underline{x}|y]\right)^2\middle|\underline{y} = y\right] = \int_{-\infty}^{\infty} \left(x - \mathrm{E}[\underline{x}|y]\right)^2 f_{x|y}(x|y)\mathrm{d}x \tag{2.82}$$

and can be also written as

$$\mathrm{var}[\underline{x}|y] \coloneqq \mathrm{E}[\underline{x}^2|y] - \left(\mathrm{E}[\underline{x}|y]\right)^2 \tag{2.83}$$

It is obvious from the definition (2.80) that the conditional expectation $\mathrm{E}[g(\underline{x})|y]$ is function of the real variable y, call it $h(y)$, rather than constant. If we do not specify in the condition the value y of the stochastic variable $\underline{y}$, then the quantity $\mathrm{E}\left[g(\underline{x})\middle|\underline{y}\right] = h(\underline{y})$ becomes function of the stochastic variable $\underline{y}$. Hence, it is a stochastic variable itself. Its own expected value is:

$$\mathrm{E}\left[\mathrm{E}\left[g(\underline{x})\middle|\underline{y}\right]\right] = \int_{-\infty}^{\infty} \mathrm{E}[g(\underline{x})|y] f_y(y)\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x) f_{xy}(x,y)\,\mathrm{d}x\mathrm{d}y \tag{2.84}$$

where we have utilized (2.80) and (2.79). As a result,

$$\mathrm{E}\left[\mathrm{E}\left[g(\underline{x})\middle|\underline{y}\right]\right] = \mathrm{E}[g(\underline{x})] \tag{2.85}$$

This can be readily generalized for a function of two stochastic variables, i.e.,

$$\mathrm{E}\left[\mathrm{E}\left[g\left(\underline{x},\underline{y}\right)\middle|\underline{y}\right]\right] = \mathrm{E}\left[g\left(\underline{x},\underline{y}\right)\right] \tag{2.86}$$

Entropy, as formally defined for the univariate case in section 2.9 and for the bivariate case in equations (2.77) and (2.78), is an expectation and thus we can also define

conditional forms of entropy which are quite useful. Thus, for a specified value of $\underline{y} = y$ and for a discrete stochastic variable the entropy is:

$$\Phi[\underline{x}|y] \coloneqq \mathrm{E}[-\ln P(\underline{x}|y)] = -\sum_{i,j} P_{i|j} \ln P_{i|j} \tag{2.87}$$

where $P_{i|j} \coloneqq P\left\{\underline{x} = x_i | \underline{y} = y_j\right\}$, and for a continuous stochastic variable it is:

$$\Phi[\underline{x}|y] \coloneqq \mathrm{E}\left[-\ln \frac{f(\underline{x}|y)}{\beta(\underline{x})}\right] = -\int_{-\infty}^{\infty} \ln \frac{f(x|y)}{\beta(x)} f(x|y)\mathrm{d}x \tag{2.88}$$

These quantities depend on the specified conditioning value y of $\underline{y}$. However, we can define a global conditional entropy, for an unspecified value of $\underline{y}$.

$$\mathrm{E}\left[\Phi\left[\underline{x}|\underline{y}\right]\right] \coloneqq \mathrm{E}\left[\mathrm{E}\left[-\ln P\left(\underline{x}|\underline{y}\right)|\underline{y}\right]\right] = -\sum_{j}\sum_{i,j} P_{i|j} \ln P_{i|j}\, P_j \tag{2.89}$$

For continuous stochastic variables it is:

$$\mathrm{E}\left[\Phi\left[\underline{x}|\underline{y}\right]\right] \coloneqq \mathrm{E}\left[\mathrm{E}\left[-\ln \frac{f\left(\underline{x}|\underline{y}\right)}{\beta(\underline{x})}\right]|\underline{y}\right] = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ln \frac{f(x|y)}{\beta(x)} f(x|y) f(y)\mathrm{d}x\,\mathrm{d}y \tag{2.90}$$

A relationship analogous to (2.85) does not hold in this case. This is easy to verify as

$$\mathrm{E}\left[\Phi\left[\underline{x}|\underline{y}\right]\right] = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ln \frac{f(x|y)}{\beta(x)} f(x,y)\mathrm{d}x\,\mathrm{d}y \neq -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ln \frac{f(x)}{\beta(x)} f(x,y)\mathrm{d}x\,\mathrm{d}y = \Phi[\underline{x}] \tag{2.91}$$

In fact, the true relationship between the (global) conditional entropy and the marginal one is an inequality (e.g. Papoulis, 1991, p. 564):

$$\mathrm{E}\left[\Phi\left[\underline{x}|\underline{y}\right]\right] \leq \Phi[\underline{x}] \tag{2.92}$$

Another distinction we have to stress is that:

$$\mathrm{E}\left[\Phi\left[\underline{x}|\underline{y}\right]\right] \neq \Phi[\underline{x}|y] \tag{2.93}$$

because the latter quantity is generally a function of y while the former is not. An interesting exception is the case of a bivariate normal distribution in which $\Phi[\underline{x}|y]$ turns out to be a constant rather than a function of y ($\Phi[\underline{x}|y] = \Phi_{\mathrm{c}}[\underline{x}] \leq \Phi[\underline{x}]$). Generally, we should stress that:

- conditional expectations like $\mathrm{E}[\underline{x}|y]$ are deterministic functions of the conditioning value y;
- conditional expectations like $\mathrm{E}[\underline{x}|\underline{y}]$ are stochastic variables, depending on $\underline{y}$.

These remarks have to be added to the notes of Digression 2.A about the importance of notation.

Digression 2.I: Does information decrease entropy?

It is intuitive to say that, if a stochastic variable $\underline{x}$ has some relationship with another stochastic variable $\underline{y}$ then, if we observe the value of $\underline{y}$, our uncertainty on $\underline{x}$ would decrease. As entropy is a formal measure of uncertainty, this can be formally stated as follows: the conditional entropy of $\underline{x}$ given information on $\underline{y}$ is smaller than the unconditional entropy of $\underline{x}$. However, this simple truth is sometimes contradicted in scientific texts. The reasons of the contradiction are the inattentive use of concepts and inattentive notation. We will illustrate them with the following example.

In Digression 2.C and Digression 2.D we studied the probabilities of the dry and wet (rain) state in an area. Continuing this study, we now introduce the stochastic variables and $\underline{x}$ and $\underline{y}$ for today's and yesterday's state, respectively, with $\{\underline{x} = 0\}$ and $\{\underline{x} = 1\}$ representing a dry and wet state of today, respectively, and likewise for yesterday. We assume for illustration the conditional probabilities:

$$P\{\underline{x} = 1|\underline{y} = 1\} = 0.4, \qquad P\{\underline{x} = 1|\underline{y} = 0\} = 0.15$$

from which it directly follows that

$$P\{\underline{x} = 0|\underline{y} = 1\} = 0.6, \qquad P\{\underline{x} = 0|\underline{y} = 0\} = 0.85$$

and after some simple calculations we also find the marginal probabilities to be

$$P\{\underline{x} = 0\} = 0.8, \qquad P\{\underline{x} = 1\} = 0.2$$

Hence the unconditional entropy is:

$$\Phi[\underline{x}] = \mathrm{E}[-\ln P(\underline{x})] = -0.8 \ln 0.8 - 0.2 \ln 0.2 = 0.500$$

while the entropy conditional on yesterday being dry is:

$$\Phi\left[\underline{x}|\underline{y} = 0\right] = \mathrm{E}\left[-\ln P\left(\underline{x}\middle|\underline{y} = 0\right)\right] = -0.85 \ln 0.85 - 0.15 \ln 0.15 = 0.423$$

and that conditional on yesterday being dry is:

$$\Phi\left[\underline{x}|\underline{y} = 1\right] = \mathrm{E}\left[-\ln P\left(\underline{x}\middle|\underline{y} = 1\right)\right] = -0.6 \ln 0.6 - 0.4 \ln 0.4 = 0.673$$

Now it is true that the information that yesterday was a wet day increased the entropy from 0.5 (without any information) to 0.673. This happened because the probabilities of the two states, which initially were 0.8 vs. 0.2, far from the equiprobability (0.5) in which the entropy is maximized, have now approached each other (0.6 vs. 0.4) and thus the entropy increased.

But this happens for that particular value, $\underline{y} = 1$. If we consider all values (in our case two), on the average the (global) conditional entropy is

$$\mathrm{E}\left[\Phi\left[\underline{x}|\underline{y}\right]\right] = 0.500 \times 0.8 + 0.423 \times 0.2 = 0.473 < 0.500$$

In other words, the reply to the question in the above title is, Yes, information decreases entropy, but we must be attentive about the correct use of the probabilistic concepts.

2.14 Many variables

All above theoretical analyses can be easily extended to more than two stochastic variables. For instance, the distribution function of the n stochastic variables $\underline{x_1}, \dots, \underline{x_n}$ is

$$F_{x_1,\dots,x_n}(x_1, \dots, x_n) \coloneqq P\{\underline{x}_1 \le x_1, \dots, \underline{x}_n \le x_n\} \tag{2.94}$$

and is related to the n-dimensional probability density function by

$$F_{x_1,\ldots,x_n}(x_1,\ldots,x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_{x_1,\ldots,x_n}(\xi_1,\ldots\xi_n)\mathrm{d}\xi_n \cdots \mathrm{d}\xi_1 \tag{2.95}$$

The variables $\underline{x}_1, \ldots, \underline{x}_n$ are independent if:

$$F_{x_1,\ldots,x_n}(x_1,\ldots,x_n) = F_{x_1}(x_1)\cdots\ F_{x_n}(x_n) \tag{2.96}$$

A useful rule to mention is the so-called chain rule, which allows expressing joint densities as products of conditional densities:

$$f(x_1,\ldots,x_n) = f(x_n|x_{n-1},\ldots,x_1)\cdots\ f(x_2|x_1)f(x_1) \tag{2.97}$$

where for notational brevity we have omitted the subscripts of functions $f(\,)$, as these are identical to the arguments of the functions. A direct consequence for the evaluation of entropy is

$$\Phi\left[\underline{x}_1,\ldots,\underline{x}_n\right] = \mathrm{E}[\Phi\left[\underline{x}_n|\underline{x}_{n-1},\ldots,\underline{x}_1\right] + \cdots + \ \mathrm{E}[\Phi\left(\underline{x}_2|\underline{x}_1\right)] + \mathrm{E}[\Phi\left(\underline{x}_1\right)] \tag{2.98}$$

The expected values and moments are defined in a similar manner as in the case of two variables, and all properties discussed in section 2.12 are likewise generalized for functions of many variables.

2.15 Linear combinations of stochastic variables

A consequence of the definition of the expected value is the relationship

$$\mathrm{E}\left[c_1 g_1\left(\underline{x}_1,\underline{x}_2\right) + c_2 g_2\left(\underline{x}_1,\underline{x}_2\right)\right] = c_1\mathrm{E}\left[g_1\left(\underline{x}_1,\underline{x}_2\right)\right] + c_2\mathrm{E}\left[g_2\left(\underline{x}_1,\underline{x}_2\right)\right] \tag{2.99}$$

where c_1 and c_2 are any constants, and g_1 and g_2 are any functions. Apparently, this property can be extended to any number of functions g_i. Applying it for the weighted sum of two variables we obtain

$$\mathrm{E}\left[a_1\underline{x}_1 + a_2\underline{x}_2\right] = a_1\mathrm{E}\left[\underline{x}_1\right] + a_2\mathrm{E}\left[\underline{x}_2\right] \tag{2.100}$$

Likewise, we can calculate the variance of the weighted sum. After some algebraic operations we get

$$\mathrm{var}\left[a_1\underline{x}_1 + a_2\underline{x}_2\right] = a_1^2\mathrm{var}\left[\underline{x}_1\right] + a_2^2\mathrm{var}\left[\underline{x}_2\right] + 2a_1a_2\mathrm{cov}\left[\underline{x}_1,\underline{x}_2\right] \tag{2.101}$$

It is much more difficult to calculate the probability distribution function of such combinations. As an example, for the simplest case, the sum $\underline{z} = \underline{x}_1 + \underline{x}_2$ of two independent variables $\underline{x}_1$ and $\underline{x}_2$ has density:

$$f_z(z) = \int_{-\infty}^{\infty} f_{x_1}(z - x_2)\, f_{x_2}(x_2)\mathrm{d}x_2 \tag{2.102}$$

The latter integral is known as the *convolution integral* of $f_{x_1}(x)$ and $f_{x_2}(x)$. For nonnegative variables it takes the form:

$$f_z(z) = \int_0^z f_{x_1}(z - x_2)\, f_{x_2}(x_2)\mathrm{d}x_2, \qquad z > 0 \tag{2.103}$$

2.16 Variance-based correlation and the climacogram

While covariance and its equivalent standardized form, i.e., correlation, have been the most customary tools to characterize dependence, they are neither the only nor the most effective ones. Assuming two stochastic variables $\underline{x}_1$ and $\underline{x}_2$ (possibly representing different physical quantities) with means μ_i (i =1, 2), standard deviations σ_i, covariance σ_{12} and correlation coefficient r_{12}, we may form a different type of a correlation coefficient and covariance by examining a weighted sum of the two variables. Namely, we examine the average of the variables $\underline{x}_i$ after standardizing them with their standard deviations σ_i, which is necessary if they represent different physical quantities, in order to make them compatible for addition. From (2.101) we obtain that the variance of this average is:

$$\mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1} + \frac{\underline{x}_2}{\sigma_2}\right)\right] = \frac{1}{4}\mathrm{E}\left[\left(\frac{\underline{x}_1 - \mu_1}{\sigma_1} + \frac{\underline{x}_2 - \mu_2}{\sigma_2}\right)^2\right] = \frac{1}{2} + \frac{1}{2}\mathrm{cov}\left[\frac{\underline{x}_1}{\sigma_1}, \frac{\underline{x}_2}{\sigma_2}\right] \tag{2.104}$$

where we can recognize in the rightmost term the correlation coefficient r_{12}. Defining

$$\rho_{12} \coloneqq \mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1} + \frac{\underline{x}_2}{\sigma_2}\right)\right], \qquad \gamma_{12} \coloneqq \sigma_1\sigma_2\rho_{12} = \mathrm{var}\left[\frac{1}{2}\left(\sqrt{\frac{\sigma_2}{\sigma_1}}\underline{x}_1 + \sqrt{\frac{\sigma_1}{\sigma_2}}\underline{x}_2\right)\right] \tag{2.105}$$

we find from (2.104) that

$$\rho_{12} \coloneqq \frac{1 + r_{12}}{2}, \qquad \gamma_{12} \coloneqq \frac{\sigma_1\sigma_2 + \sigma_{12}}{2} \tag{2.106}$$

Obviously, the same information as in r_{12} is contained ρ_{12}, which lies in the interval [0, 1] with the values 0, 1/2, 1 representing fully anti-correlated, uncorrelated and fully correlated variables, respectively. Consequently, γ_{12} lies in the interval $[0, \sigma_1\sigma_2]$ with the values $0, \sigma_1\sigma_2/2, \sigma_1\sigma_2$ representing fully anti-correlated, uncorrelated and fully correlated variables, respectively.

The power of the notion of ρ_{12} and γ_{12} is the fact that, unlike r_{12}, they can readily be expanded to many variables to provide a macroscopic (or bulk) measure of correlation among all of them. Considering a number $\kappa > 0$ of stochastic variables, in the customary case where all have identical variances $\gamma_1 = \sigma^2$, we write:

$$\rho_\kappa \coloneqq \mathrm{var}\left[\frac{X_\kappa}{\kappa\sigma}\right] = \frac{\gamma_\kappa}{\gamma_1}, \qquad \gamma_\kappa \coloneqq \mathrm{var}\left[\frac{X_\kappa}{\kappa}\right] = \gamma_1\rho_\kappa, \qquad \underline{X}_\kappa \coloneqq \underline{x}_1 + \cdots + \underline{x}_\kappa \tag{2.107}$$

Clearly, γ_κ is the climacogram, already defined in Chapter 1, and ρ_κ is a dimensionless (standardized) climacogram. They range in the intervals $(0, \gamma_1)$ and (0, 1), respectively, with the highest value representing full correlation ($\underline{x}_1 + \cdots + \underline{x}_\kappa = \kappa\underline{x}_1 + c$, where c is a constant) and the lowest representing deterministic linear dependence, i.e. the condition that $\underline{x}_1 + \cdots + \underline{x}_\kappa = c$). In case of independence, $\gamma_\kappa \coloneqq \gamma_1/\kappa$ and $\rho_\kappa \coloneqq 1/\kappa$.

2.17 Limiting distributions and the central limit theorem

As we have seen in section 2.15, it is not very easy to calculate the distribution function of the sum of two stochastic variables from the distributions of the constituents. This becomes even less easy as the number of constituents increases. However, if this number becomes quite large, paradoxically the problem becomes easier—this is the ease of macroscopization. Central role in resolving this paradox plays the *central limit theorem*[*], one of the most important in probability theory. It concerns the limiting distribution function of a sum of stochastic variables – constituents, which, under some conditions but irrespectively of the distribution functions of the constituents, is always the same, the celebrated *normal distribution*. This is the most commonly used distribution in probability theory as well as in all scientific disciplines and, as we have seen in section 2.10, it is also derived from the principle of maximum entropy for Lebesgue background measure and constrained variance.

Let $\underline{x}_i$ $(i = 1, \dots, n)$ be stochastic variables whose sum $\underline{z}_n \coloneqq \underline{x}_1 + \cdots + \underline{x}_n$ has mean μ_z and variance σ_z^2. The central limit theorem says that, under some general conditions (see below), as n tends to infinity, the distribution of $\underline{z}$ will tend to the normal distribution (also known as Gauss or Gaussian distribution and denoted as $\mathrm{N}(\mu_z, \sigma_z)$), i.e.,

$$\lim_{n\to\infty} F_{z_n}(z) = \int_{-\infty}^{z} \frac{1}{\sigma_z\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\mu_z}{\sigma_z}\right)^2} \tag{2.108}$$

and in addition, if $\underline{x}_i$ are continuous variables, the density function of $\underline{z}_n$ has also a limit,

$$\lim_{n\to\infty} f_{z_n}(z) = \frac{1}{\sigma_z\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\mu_z}{\sigma_z}\right)^2} \tag{2.109}$$

We observe in (2.108) and (2.109) that the limits of the functions $F_{z_n}(z)$ and $f_{z_n}(z)$ do not depend on the distribution functions of $\underline{x}_i$, that is, the result is always the same. Thus, provided that the conditions for the applicability of the theorem hold, (a) we can know the macroscopic behaviour (the distribution function of the sum) without knowing details of the constituents, and (b) precisely the same distribution describes any variable that is a sum of a large number of components. Here lies the great importance of the normal distribution in all sciences (mathematical, physical, social, economic, etc., as well as stochastics per se).

In practice, the convergence for $n \to \infty$ can be regarded as an approximation if n is sufficiently large. But how large should n be so that the approximation be satisfactory? Generally, the literature suggests that a value n = 30 is satisfactory. However, this varies depending on the (joint) distribution of the constituents $\underline{x}_i$. Figure 2.5 gives a graphical illustration of the convergence to the normal distribution of the sum of n independent

[*] The term was most likely introduced by Pólya in 1920. A first version of the theorem was formulated and proved by Laplace in 1810 while at about the same time Gauss studied the normal distribution in characterizing measurement or model errors. Earlier, in 1733 de Moivre had introduced the normal distribution as an approximation of the binomial distribution (Fischer, 2010).

variables. Clearly, if the distribution of $\underline{x}_i$ is symmetric (left panel, with uniform distribution of $\underline{x}_i$), the convergence is rapid (even for $n = 3$) but if it is asymmetric (right panel, with exponential distribution of $\underline{x}_i$) a value higher that 32 (the highest n shown in the plot) is needed for a satisfactory convergence. In case of dependent $\underline{x}_i$ with positive correlation, the convergence is slower and a much larger n is needed.

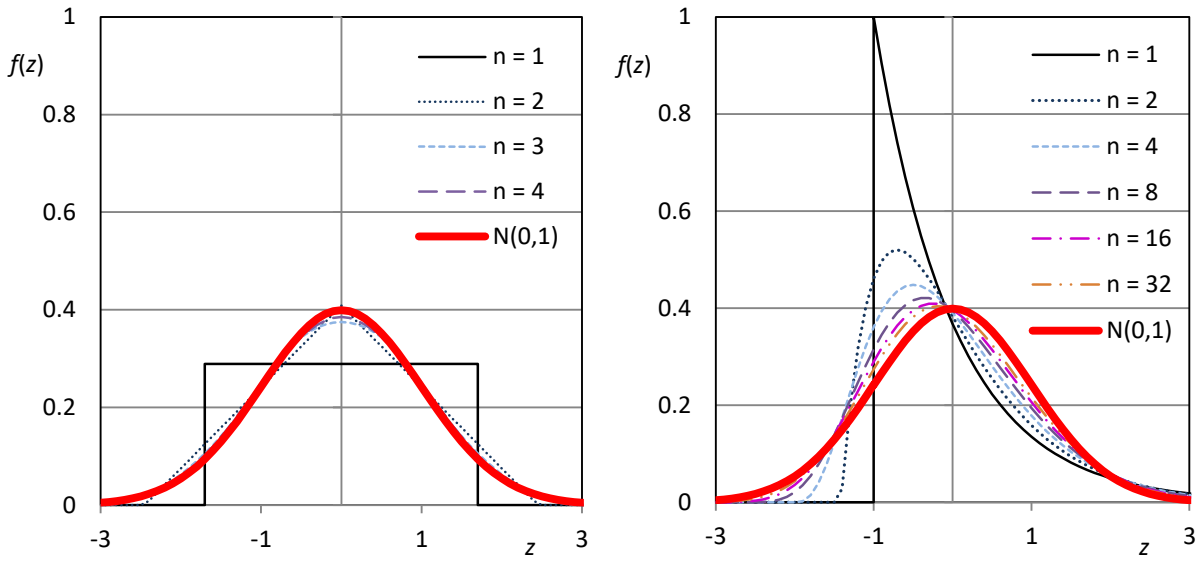

Figure 2.5 Convergence of the sum of independent identically distributed stochastic variables to the normal distribution (thick line). The thin continuous lines represent the probability density of the constituent variables $\underline{x}_i$, which have mean 0 and standard deviation 1. On the left panel the density is uniform on the interval $(-\sqrt{3}, \sqrt{3})$ with $f(x) = 1/(2\sqrt{3})$ and on the right panel exponential, $f(x) = \mathrm{e}^{-x-1}, x \geq -1$ (the parameters are chosen so as to have mean 0 and standard deviation 1). The dotted lines represent the densities of the sums $\underline{z}_n := (\underline{x}_1 + \cdots + \underline{x}_n)/\sqrt{n}$ for the values of n indicated in the legend. (The division of the sum by $\sqrt{n}$ helps for a better presentation of the curves, as all $\underline{z}_n$ have the same mean and variance, 0 and 1, respectively, and does not affect the essentials of the central limit theorem.)

The conditions for the validity of the central limit theorem are general enough, so that they are met in many practical situations. Some sets of conditions (e.g. Papoulis, 1990, p. 215) with particular interest are the following: (a) the variables $\underline{x}_i$ are independent identically distributed with finite third moment; (b) the variables $\underline{x}_i$ are bounded from above and below with variance greater than zero; (c) the variables $\underline{x}_i$ are independent with finite third moment and the variance of $\underline{z}_n$ tends to infinity as n tends to infinity. The theorem has been extended for variables $\underline{x}_i$ that are interdependent, but each one is effectively dependent on a finite number of other variables. Gnedenko and Kolmogorov (1949) proved an extended version of the theorem, according to which the sum of n stochastic variables with heavy tail distributions with tail index $\xi > 1/2$, therefore having infinite variance, will tend to the co-called Lévy alpha-stable distribution, as $n \to \infty$. If $\xi < 1/2$, the standard central limit theorem holds, i.e., the sum converges to the Gaussian distribution, which is a special case of Lévy alpha-stable distribution. In the field of hydroclimatic processes, the standard theorem suffices because we can justifiably assume

that those processes have finite variance: an infinite variance would presuppose infinite energy to materialize, which is absurd.

Most hydroclimatic processes (particularly rainfall and streamflow) have skewed distributions at fine time scales, and therefore the normal distribution is not a suitable model at these scales. However, the normal distribution describes with satisfactory accuracy variables that refer to longer time scales such as annual. Thus, the annual rainfall depth in an area with wet climate is the sum of many (e.g., 50-100) rainfall events during the year; this, however, is not valid for rainfall in dry areas. Likewise, the annual runoff volume passing through a river section can be regarded as the sum of 365 daily volumes. These are not independent, but the central limit theorem can be applicable again.

Nonetheless, it should be stressed that the convergence to the normal distribution concerns the body of the distribution. For example, what is depicted in Figure 2.5 is about the body of the distribution. What happens with the tail behaviour, i.e., the extremes? Apparently, once the theoretical conditions of validity are satisfied, the theoretical result should hold true. However, this may not be of any help in practice as the convergence in the tail is much slower. Figure 2.6 (left) shows that the convergence in the tail is indeed slow for the exponential distribution, much slower than that of the body of the same distribution shown in Figure 2.5 (right). The coefficient of skewness for the sum of 32 $\underline{x}_i$ is rather small (0.35) indicating a rather satisfactory approximation by the normal distribution. However, the Figure 2.6 (left) shows that for $\overline{F} = 0.001$ the distribution of the sum of 32 $\underline{x}_i$ is by an order of magnitude larger than that of the normal distribution.

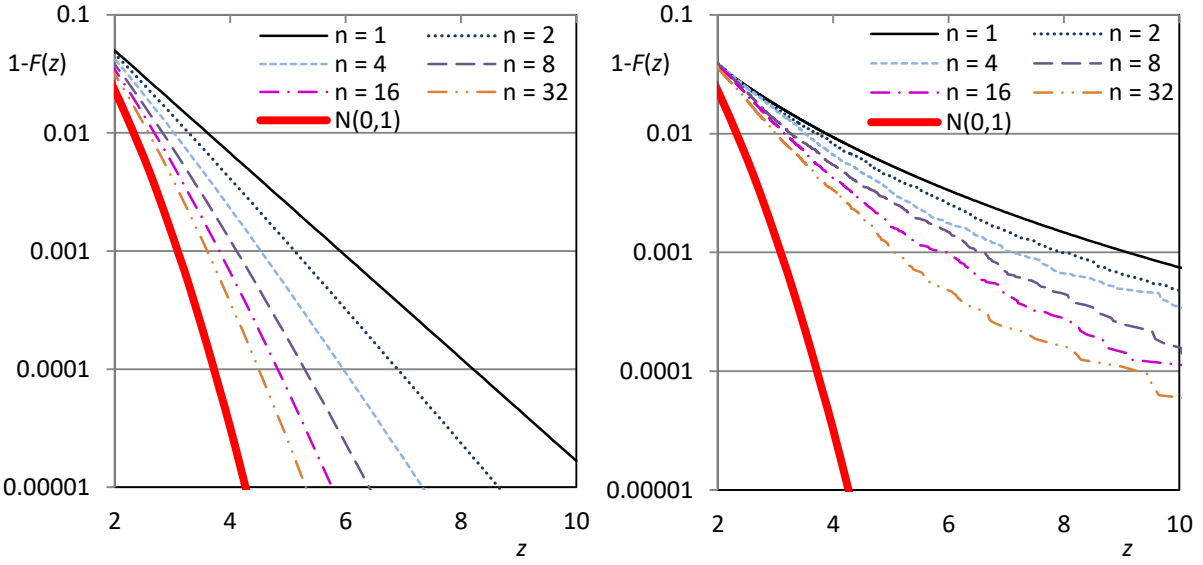

Figure 2.6 Convergence of the sum of independent identically distributed stochastic variables to the normal distribution (thick line) with focus on the right tail. The thin continuous lines represent the exceedance probability of the constituent variables $\underline{x}_i$, which have mean 0 and standard deviation 1. On the left panel the distribution is exponential with density $f(x) = \mathrm{e}^{-x-1}, x > -1$ as in the right panel of Figure 2.5) and on the right panel Pareto with tail index 1/4 and exceedance probability $\bar{F}(x) = \left(4/3 + x\sqrt{2}/3\right)^{-4}, x \geq -1/\sqrt{2}$ (the parameters are chosen so that the mean is 0 and standard deviation 1). The dotted lines represent the exceedance probability of the sums $\underline{z}_n := (\underline{x}_1 + \cdots + \underline{x}_n)/\sqrt{n}$ for the values of n indicated in the legend. Their curly shape in the right panel is due to the numerical (Monte Carlo) method used to construct them as the analytical integration is not possible beyond $n = 2$.

For a heavy tailed distribution there are differences of several orders of magnitude as shown for the Pareto distribution in Figure 2.6 (right). The tail index of this Pareto distribution is 1/4, which means that the moments below the fourth order exist and therefore the necessary conditions for the central limit theorem are satisfied. Despite that, the approximation of the distribution tail is unsatisfactory. Actually, it can be easily understood that, as the moments for order ≥ 4 of $\underline{x}_i$ are infinite, the same will hold for the sum of any finite number of $\underline{x}_i$, while the limiting normal distribution has all its moments finite. This conflict, along with the fact that the behaviour of extremes is closely connected to the high order moments of a distribution (see Chapter 6) suggests that we must be very attentive in the application of the theorem in hydroclimatic processes, particularly because these processes seem to exhibit heavy tails and long-range dependence.

2.18 Limiting extreme value distributions

By analogy with the central limit theorem referring to the sum or the average of many variables, limiting distributions may also arise, as $n \to \infty$, for the maximum of these variables, $\underline{y}_n \coloneqq \max(\underline{x}_1, \dots, \underline{x}_n)$, whose exact distribution function for independent and identically distributed $\underline{x}_n$ is:

$$F_{y_n}(y) = \big(F_x(y)\big)^n \tag{2.110}$$

The relevant theory has been developed in the 20th century. Historically, it was Fréchet (1927) the first to identify one of the asymptotic distributions of maxima, which bears his name. Fisher and Tippett (1928) showed that there are only three possible limiting distributions for extremes, while von Mises (1936) identified sufficient conditions for convergence to the three limiting laws. Gnedenko (1943) set the solid foundations of the asymptotic theory of extremes providing the precise conditions for the weak convergence to the limiting laws. It is worth noting in this respect the celebrated book by Gumbel (1958), who was one of the pioneers promoting and applying the formal theory into engineering practice. The theory is concisely presented in a review paper by Davison and Huser (2015). Assuming that $\underline{x}_i$ are independent and identically distributed, there exist a real number ξ and sequences of real numbers $\lambda_n > 0$ and ε_n such that the rescaled maximum $\underline{y}_n^* \coloneqq \max(\underline{x}_1, \dots, \underline{x}_n)/\lambda_n - \varepsilon_n$ has limiting distribution, as $n \to \infty$:

$$H(y) \coloneqq F_{y_\infty^*}(y) = \exp\left(-\left(1 + \xi\left(\frac{y}{\lambda} - \varepsilon\right)\right)^{-1/\xi}\right), \qquad \xi y \geq \xi\lambda\left(\varepsilon - \frac{1}{\xi}\right) \tag{2.111}$$

Here $\lambda > 0$ is a scale parameter, ε is a dimensionless location parameter and ξ is a shape parameter, identical to the tail index.

The parameter ξ has a unique value but the parameters λ and ε are not unique. They can be chosen as convenient (different choices will lead to appropriate modification of the sequences λ_n and ε_n). A natural choice is $\varepsilon = 0$, $\lambda = 1$. A more customary option is to choose a large n for which convergence has been achieved at a satisfactory degree, for that n set $\lambda_n = 1$ and $\varepsilon_n = 1$ (so that $\underline{y}_n^* \equiv \underline{y}_n = \max(\underline{x}_1, \dots, \underline{x}_n)$ without any rescaling) and calculate

λ and ε from equation (2.111). To this aim (and given that, for finite n, (2.111) is an approximation and not an exact relationship) we choose two points x_1 and x_2 and equate $F(x)^n$ with $H(x)$ at these points. For mathematical convenience we can choose the two points so that $-x_1/\lambda + \varepsilon = 0, -x_2/\lambda + \varepsilon = -1$, or $x_1 = \lambda\varepsilon, x_2 = \lambda\varepsilon + \lambda$. Hence, $F(\lambda\varepsilon)^n = \mathrm{e}^{-1}, F(\lambda\varepsilon + \lambda)^n = \mathrm{e}^{-(1+\xi)^{-1/\xi}}$. Solving for λ and ε we find:

$$\lambda = F^{-1}\left(\mathrm{e}^{-\frac{1}{(1+\xi)^{1/\xi}\,n}}\right) - F^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right), \qquad \varepsilon = \frac{F^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right)}{\lambda} \tag{2.112}$$

where for $\xi \to 0, (1+\xi)^{-1/\xi} \to 1/\mathrm{e}$. This is usually made unconsciously, for example when we study annual maxima of daily values and fit $H(y)$ of equation (2.111) on the annual maxima directly, without even deriving it from $F_x(x)$.

Depending on the value of ξ, the limiting distribution in equation (2.111), known as the generalized extreme value (GEV) distribution, is a compact expression including three cases with different behaviours:

- For $\xi = 0$, GEV takes the following form, known as the Gumbel distribution or extreme value type I (EV1) distribution:

$$H(y) = \exp(-\exp(-y/\lambda + \varepsilon)) \tag{2.113}$$

  This is a light-tailed distribution without an upper or lower bound.
- For $\xi > 0$, the distribution is known as Fréchet or extreme value type II (EV2) and has a lower bound at $\lambda\psi - \lambda/\xi$. This is a heavy-tailed distribution with tail index ξ.
- In case that $\xi < 0$ the distribution is known as the reverse Weibull or the extreme value type III (EV3) distribution. This has an upper bound for y at $\lambda\psi - \lambda/\xi$.

The GEV has the property to be max-stable, meaning that maxima from this distribution, after linear transformation, have the same distribution. More formally, Fréchet's necessary condition for max-stability is this: For any $n \in \mathbb{N}$ and $y \in \mathbb{R}$, there exist real numbers $a_n > 0$ and b_n such that:

$$\big(H(a_n y + b_n)\big)^n = H(y) \tag{2.114}$$

In fact, GEV is the only distribution satisfying this condition.

A specific parent distribution $F(x)$ belongs to the so-called (*max*-)*domain of attraction* of one of the three limiting laws, in the sense that distribution of rescaled maxima from this parent distribution is this particular limiting law. Formal mathematical conditions determining a parent distribution's domain of attraction were formulated by von Mises (1936) and Gnedenko (1943). The practical result is that heavy-tailed distributions with tail index $\xi > 0$ (e.g. Pareto, Pareto-Burr-Feller, Student and its extension, generalized log-gamma and generalized beta prime) belong to the domain of attraction of EV2. Light-tailed distributions (e.g. exponential, gamma, Weibull, normal and their generalizations) as well as heavy-tailed distributions with tail index $\xi = 0$ (e.g. lognormal) belong to the domain of attraction of EV1. In the domain of attraction of EV3 belong distributions bounded from above (e.g. uniform).

Because of its upper bound, EV3 is not an appropriate model for hydroclimatic extremes, for nature has no upper limits. The values of ξ which we expect to see in hydroclimatic processes are in the range $(0, 1/2)$ so that the variance be finite, as already discussed in section 2.17. The exact value of the tail index is important to specify in engineering design. The major question in this regard is how the value of an extreme quantity y grows as the probability of exceedance $\overline{H}(y)$ decreases tending to zero. To put it the reverse way, at which rate does y tend to infinity as the probability of exceedance tends to zero? The Gumbel distribution represents the mathematically proven lower limit to the rate of this growth. The alternative is the Fréchet law which represents a higher rate of growth. The two options may lead to substantial differences in design quantities for high return periods. As already discussed, the Fréchet law which has a positive tail index is a more realistic option.

When we are interested about minima, we can follow the same procedure observing that $\underline{z}_n \coloneqq \min(\underline{x}_1, \dots, \underline{x}_n) = -\max(-\underline{x}_1, \dots, -\underline{x}_n)$. Consequently, $P\{\underline{z}_n \le z\} = 1 - P\{\max(-\underline{x}_1, \dots, -\underline{x}_n) \ge -z\}$ and hence the limiting distribution is

$$G(z) \coloneqq F_{z_\infty^*}(y) = 1 - \exp\left(-\left(1 + \xi\left(-\frac{z}{\lambda} - \varepsilon\right)\right)^{-1/\xi}\right), \qquad \xi z \le \xi\lambda\left(\frac{1}{\xi} - \varepsilon\right) \tag{2.115}$$

Again, we have three cases: (a) $\xi = 0$, corresponding to the Gumbel (EV1) distribution of minima, i.e.,

$$G(z) = 1 - \exp(-\exp(z/\lambda + \varepsilon)) \tag{2.116}$$

(b) $\xi > 0$, corresponding to the reverse Fréchet distribution, which has upper bound $\lambda/\xi - \lambda\varepsilon$ and a heavy lower tail, and (c), $\xi > 0$, corresponding to the Weibull distribution, which has lower bound $\lambda/\xi - \lambda\varepsilon$ and a light upper tail.

While most of the above mathematical developments have assumed independent stochastic variables, the results can be approximately valid even in case of variables dependent in time. Specifically, Leadbetter (1983) demonstrated that, under mild conditions, maxima of dependent series follow the same form of distributional limit laws as those of independent series. However, the dependence changes the location and scale parameters (Davison and Huser, 2015) in such a manner as if $H(y)$ was replaced by $\big(H(y)\big)^\theta$, where $\theta \in (0,1]$ is the so-called *extremal index*. It can be seen that this replacement is equivalent with a change of the parameters λ and ε, while ξ remains the same. Also, the rate of convergence to the limit becomes slower in case of dependence. Phenomenologically, time dependence of a process causes clustering or grouping of extreme events. The unfortunate fact that dependence in time is quite often misinterpreted as nonstationarity, may explain the lately growing body of publications detecting nonstationarity in extremes (cf. Koutsoyiannis and Montanari, 2015).

Here it should be stressed that, if compared to the central limit theorem, which is characterized by a fast convergence to the limit (except in the extremes, as seen in Figure 2.6), the convergence to the max-stable distribution may be much slower at cases. The rate of convergence to the limit of distributions belonging to the domain of attractions of

EV2 is generally satisfactory. However, for those belonging to the domain of attraction of EV1, such as the normal and lognormal distribution, the rate is desperately slow. The meaning of a slow convergence in real-world applications, where n is finite and often small, is that the approximation of EV1 to the actual distribution of maxima is not satisfactory. Thus, it may be preferable to approximate the actual distribution of maxima of variables with distributions belonging to the domain of attraction of EV1 by the EV2 distribution, as illustrated in Digression 2.J.

Digression 2.J: How well do limiting distributions approximate the exact ones?

For independent identically distributed variables, the exact distribution of maxima is $F(x)^n$ (equation (2.110)). To approximate the exact distribution by the GEV we use equation (2.112). As an example, for the maxima from the standard normal distribution approximated by the EV1 we get:

$$\lambda = F_{\mathrm{N}}^{-1}\left(\mathrm{e}^{-\frac{1}{\mathrm{e}n}}\right) - F_{\mathrm{N}}^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right), \qquad \varepsilon = \frac{F_{\mathrm{N}}^{-1}\left(\mathrm{e}^{-\frac{1}{n}}\right)}{\lambda}$$

As a second example, for the Pareto distribution, $F(x) = 1 - (1+x)^{-1/\xi}$, approximated by the EV2 we get:

$$\lambda = \left(1 - \mathrm{e}^{-\frac{1}{n(1+\xi)^{1/\xi}}}\right)^{-\xi} - \left(1 - \mathrm{e}^{-\frac{1}{n}}\right)^{-\xi}, \qquad \varepsilon = \frac{\left(1 - \mathrm{e}^{-\frac{1}{n}}\right)^{-\xi} - 1}{\lambda}$$

We have applied this approximation for n = 10, 100 and 1000 for the normal and the Pareto distributions which belong to the domain of attraction of EV1 and EV2, respectively. The results are shown graphically in Figure 2.7, along with the case n = 1, i.e. the parent distribution per se for the sake of comparison.

The results are very good for the Pareto distribution and very bad for the normal distribution, even for n = 1000, where the EV1 overestimates the actual distribution severely. One may think of using the EV3 instead of EV1 for the approximation of the normal distribution. However, this is not advisable because the EV3, despite giving a better approximation, it entails an upper bound to extremes which distorts a fundamental characteristic of the phenomenon that is modelled.

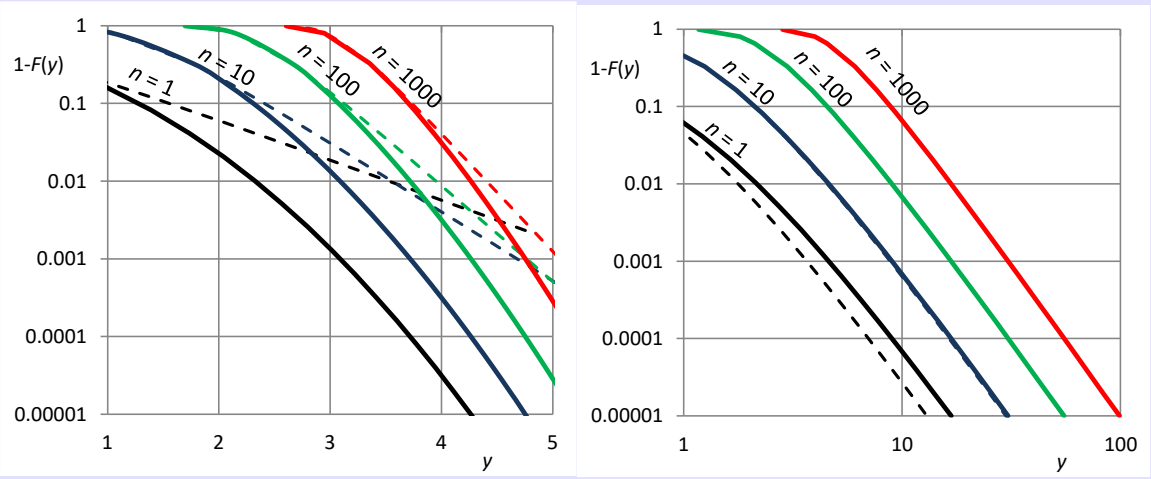

Figure 2.7 Approximation the true distribution of the maximum of n independent identically distributed variables (continuous lines) by the limiting extreme value distribution (dashed lines). **(left)** The parent distribution is the standard normal, N(0,1), and the approximating distribution is the EV1. **(right)** The parent distribution is the Pareto, $F(x) = 1 - (1+x)^{-1/\xi}$, with ξ = 0.25 and the approximating distribution is the EV2 with the same ξ.

Likewise, Figure 2.8 provides similar information for the lognormal distribution with mean and standard deviation of $\ln x$ equal to 0 and 1, respectively (denoted LN(0,1)). Like the normal, it belongs to the domain of attraction of EV1. Here the approximation is even worse than in the normal case but now the EV1 underestimates the exact distribution. For that reason, we could use EV2 as a better approximation (without having the problem of artificially inducing an upper bound). As seen in the right panel of Figure 2.8 this latter approximation is quite satisfactory.

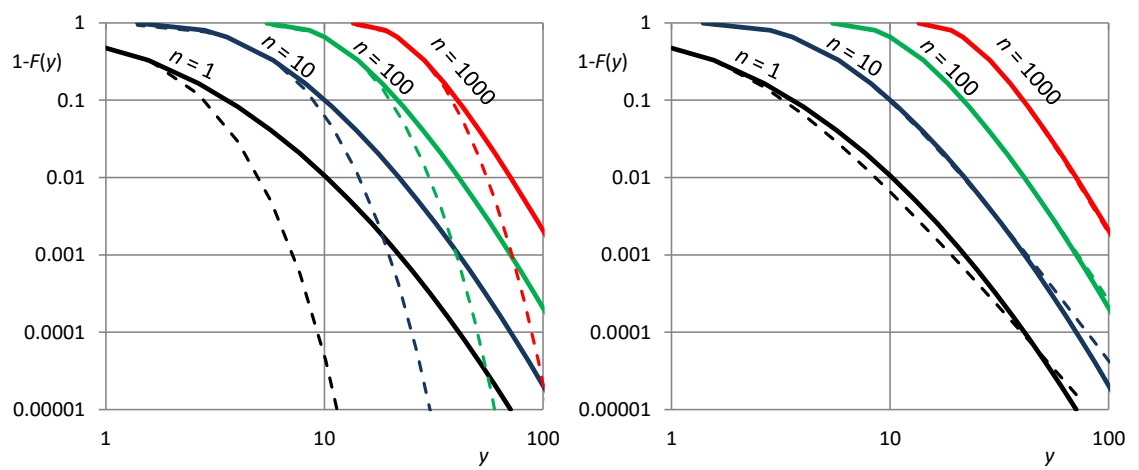

Figure 2.8 Approximation the true distribution of the maximum of n independent variables with lognormal distribution LN(0,1) (continuous lines) by the limiting extreme value distribution (dashed lines), which is **(left)** EV1 and **(right)** EV2 with $\xi = 0.3/n^{0.07}$.

2.19 Relationship of parent and extreme value distribution

Because of problems originating from the slow rate of convergence of the actual distribution to GEV (particularly EV1), it may be a good idea not to use the limiting distributions in practical applications but, instead, to model the tail of the parent distribution or even the entire parent distribution. Yet the theory of max-stable distributions retains its usefulness to infer the tail behaviour of the parent distribution.

Specifically, the tail behaviour of the parent distribution is described by the conditional distribution function:

$$F\left(x\middle|\underline{x} > u\right) = P\left\{\underline{x} \le x\middle|\underline{x} > u\right\} = \frac{F(x) - F(u)}{1 - F(u)} \tag{2.117}$$

for a value of the threshold u that is sufficiently large. Now, assuming that the parameterization of $H(x)$ with regard to λ and ε has been made with reference to a specific large n, as described for the derivation of (2.112), we choose u so that the exceedance probability $1 - F(u)$ equals $1/n$. (This is a very common choice as will be discussed in Digression 2.K). Thus, $F\left(x\middle|\underline{x} > u\right) = n(F(x) - 1) + 1$ or:

$$1 - F\left(x\middle|\underline{x} > u\right) = n\left(1 - F(x)\right) \tag{2.118}$$

On the other hand, we can write for $F(x)$ approaching 1,

$$-\ln H(x) \approx -\ln\left(F(x)\right)^n = -n \ln F(x) \approx n\left(1 - F(x)\right) \tag{2.119}$$

because $\ln F(x) = \ln(1 - \left(1 - F(x)\right) = -\left(1 - F(x)\right) - \left(1 - F(x)\right)^2 - \cdots \approx F(x)$. Hence, combining (2.118) and (2.119) we find:

$$F\left(x\middle|\underline{x} > \lambda\varepsilon\right) = 1 + \ln H(x) = 1 - \left(1 + \xi\left(\frac{x}{\lambda} - \varepsilon\right)\right)^{-1/\xi}, \qquad x \geq \lambda\varepsilon \tag{2.120}$$

where we equated u with $\lambda\varepsilon$ for consistency (i.e. to make $F\left(u\middle|\underline{x} > u\right) = 0$). This is the Pareto distribution for $\xi > 0$ while for $\xi = 0$ we get the exponential form:

$$F\left(x\middle|\underline{x} > \lambda\varepsilon\right) = 1 + \ln H(x) = 1 - \exp\left(\frac{x}{\lambda} - \varepsilon\right), \qquad x \geq \lambda\varepsilon \tag{2.121}$$

Furthermore, for values of x large enough to make $H(x)$ approach 1, we can use the same logic to get $\ln H(x) \approx H(x) - 1$ and hence

$$F\left(x\middle|\underline{x} > \lambda\varepsilon\right) \approx H(x) \tag{2.122}$$

This approximation error does not exceed ~1% for $H(x) > 0.99$ and ~5% for $H(x) > 0.9$.

We can generalize the above analysis for different values of the threshold ξ. In this case the resulting functional form remains the same, with the same tail index, but the location and scale parameters differ, i.e. (Davison and Huser, 2015):

$$F\left(x\middle|\underline{x} > u\right) = 1 - \left(1 + \xi\left(\frac{x}{\lambda_u} - \varepsilon_u\right)\right)^{-\frac{1}{\xi}} \tag{2.123}$$

where

$$\lambda_u := \lambda\left(1 + \xi(u/\lambda - \varepsilon)\right), \qquad \varepsilon_u := u/\lambda_u \tag{2.124}$$

It is readily confirmed that if we set $u = \lambda\varepsilon$ in (2.123) and (2.124) we recover (2.120). However, this equation is valid only for large values of u (unless the unconditional $F(x)$ is Pareto, in which case it is valid for any u.

A final note that may be relevant in some analyses is this. If the value of n in $\underline{y}_n := \max\left(\underline{x}_1, \ldots, \underline{x}_n\right)$, is not constant but a stochastic variable with Poisson distribution with mean ν, while $\underline{x}_i$ are independent, then the conditional distribution of $\underline{y}_n$ on specified n remains $F_{y_n}(y|n) = \left(F_x(y)\right)^n$ but for unspecified n the unconditional distribution becomes (Todorovic and Zelenhasic, 1970):

$$F_y(y) = \exp\left(-\nu\left(1 - F(y)\right)\right) \tag{2.125}$$

This resembles (2.119) with the difference that it is now exact rather than approximate.

As already discussed above and in section 2.18, because of the problems of the limiting extreme value distributions, it is preferable to focus the studies of extremes on the parent distributions and primarily their tails. From the above theoretical discussion, we have reasons to expect a parent distribution tail of Pareto type, or at least exponential, but this should be verified each time based on observations. Today there is abundance of hydrometeorological data on daily and subdaily scales and there is no need to extract annual or seasonal maxima. Instead, we should use the entire observational record or at least the values over some threshold. Only if the available observations are originally given in terms of time-block (e.g., annual) maxima, it is pertinent to refer to extreme value

distribution. Advantages of studying the distribution of the parent variable rather than the distribution of maxima are discussed in Digression 2.K

Digression 2.K: Block maxima vs. values over threshold vs. complete record

Traditionally, hydrometeorological records are analysed in two ways. The most frequent is to choose the highest of all recorded values for a given time period or "block" (typically year) and form a statistical sample (commonly referred to as "block maxima") with size equal to the number of blocks (typically years) of the record. The other is to form a sample of values over a threshold (here abbreviated as VOT but sometimes referred to as "peaks-over-threshold"—POT) with all recorded values over a certain threshold irrespective of the time they occurred. Usually the threshold is chosen high enough, so that the sample size is again equal to the number of years of the record. This however is not necessary: it can well be set equal to zero, so that all recorded values are included in the sample (the complete sample). However, a high threshold simplifies the study and helps focus the attention on the distribution tail. In addition, this choice simplifies the mathematical expression (compare equations (2.120) and (2.123)), leading to identity of distributional parameters of the distributions of block-maxima and values over threshold.

Additionally, studying the complete series of observations has the advantage of respecting the motto "Save hydrological observations!" (Volpi et al., 2019). Indeed, extracting maxima over some period results in waste of information because other extreme observations should also be informative about extremes. Such information (e.g., the second-largest value of a period, which can be higher than another period's largest value) is retained even if we use the values over threshold instead of the entire series of observations.

Furthermore, the design quantities should naturally correspond to the parent distribution, rather than the artificially induced maxima over an arbitrarily defined time period. This favours the use of the parent distribution. As we have seen (equation (2.122) the two are almost equal for very large design values, but for lower ones there are differences. Thus, even if our analyses are based in time-block extremes ($H(y)$), the results should eventually be converted to the parent distribution ($F(x)$) before they are used for design. The above discourse provides all the necessary mathematical support for such conversion.

The most important reason favouring the study of the complete record over that of block maxima and values over threshold is that only the former provided faithful information about time dependence of the underlying process. As we have already seen in Chapter 1, such dependence may be marked and possibly of long-range type. As we will see in Chapter 6, neglecting dependence results in underestimation of extremes. On the other hand, the procedure of extracting block maxima leads to severe distortion of the dependence structure (Iliopoulou and Koutsoyiannis, 2019), whereas the concept of taking values over threshold relies on a tacit assumption of time independence, which may be inappropriate, particularly for the streamflow process (Lombardo et al., 2019).

Chapter 3. Stochastic processes and quantification of change

3.1 Definitions

A deterministic world view is founded on a concept of sharp exactness. A deterministic mathematical description of a system uses regular variables (e.g. x) which are represented as numbers. The change of the system state is represented as a *trajectory* $x(t)$, which is the sequence of a system's states x as time t changes. Changes in time are studied using the concept of a dynamical system with certain system dynamics. The latter term denotes a transformation S_t which maps its initial state $x(0)$ in the trajectory of a dynamical system (at time 0) to its current state $x(t)$ (at time t), that is, $x(t) = S_t\big(x(0)\big)$ (Lasota and Mackey 1994).

In an indeterministic world view there is uncertainty or randomness, where the latter term does not mean anything more than unpredictability or intrinsic uncertainty. In turn, to study the change according to this approach we use the notion of a *stochastic process*. This is defined to be an arbitrarily (usually infinitely) large family of stochastic variables $\underline{x}(t)$ (Papoulis, 1991). To each one of them there corresponds an index t, which takes values from an *index set* T, most often referring to time. The time t can be either *discrete* (when T is the set of integers, $\mathbb{Z}$) or continuous (when T is the set of real numbers, $\mathbb{R}$); thus, we have respectively a *discrete-time* or a *continuous-time* stochastic process. As natural time runs continuously, the faithful representation of a natural process needs a model formulated for continuous time to avoid the risk of making artificial constructs. Nonetheless, the discrete-time representation is certainly necessary in simulation. This can be derived from the continuous time one. Typically, the discrete time representation $\underline{x}_\tau$ is derived from the continuous time one from $\underline{x}(t)$ as the temporal average:

$$\underline{x}_\tau \coloneqq \frac{1}{D} \int_{(\tau-1)D}^{\tau D} \underline{x}(u)\mathrm{d}u \tag{3.1}$$

where $\tau \in \mathbb{Z}$ represents the continuous-time interval $[(\tau-1)D, \tau D]$ and D is the time step; notice that we use different notation in the continuous and discrete time representation, in the latter case denoting time as a subscript. Each of the stochastic variables $\underline{x}(t)$ or $\underline{x}_\tau$ can be either discrete (e.g. the wet or dry state of a day) or continuous (e.g. the rainfall depth); thus, we have respectively a *discrete-state* or a *continuous-state* stochastic process.

The index set can also be a vector space, rather than the real line or the set of integers; this is the case for instance when we assign a stochastic variable (e.g. rainfall depth) to each geographical location (a two dimensional vector space) or to each location and time instance (a three-dimensional vector space). Stochastic processes with multidimensional index set are also known as stochastic (or random) fields.

A realization $x(t)$ of a stochastic process $\underline{x}(t)$, which is a regular (numerical) function of the time t, is known as a *sample function*. Typically, a realization can be known (simulated) at countable time instances, i.e. in discrete time (not in continuous time, even in a continuous-time process). Likewise, observation of a natural process is also made in

discrete time. The sequences of simulated or observed values are called a *time series*. Clearly then, a time series is a *finite* sequence of *numbers*, whereas a stochastic process is a family of *stochastic variables*, infinitely many for discrete time processes and uncountably infinitely many for continuous time processes. A large body in literature does not make this distinction and confuses stochastic processes with time series (see Digression 3.E).

3.2 Distribution function and moments

The distribution function of the stochastic variable $\underline{x}(t)$ i.e.,

$$F(x;t) \coloneqq P\{x(t) \leq x\} \tag{3.2}$$

is called *first order distribution function* of the process. Likewise, the *second order distribution function* is:

$$F(x_1, x_2; t_1, t_2) \coloneqq P\{\underline{x}(t_1) \leq x_1, \underline{x}(t_2) \leq x_2\} \tag{3.3}$$

and the n*th order distribution function* is:

$$F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) \coloneqq P\{\underline{x}(t_1) \leq x_1, \underline{x}(t_2) \leq x_2, \dots, \underline{x}(t_n) \leq x_n\} \tag{3.4}$$

A stochastic process is completely determined if we know the *n*th order distribution function for any *n*. The *n*th order probability density function of the process is derived by taking the derivatives of the distribution function with respect to all x_i.

The moments are defined in the same manner as in sections 2.12 and 2.14. Of particular interest are the following:

1. The *process mean*, i.e. the expected value of the variable $\underline{x}(t)$:

$$\mu(t) \coloneqq \mathrm{E}[\underline{x}(t)] = \int_{-\infty}^{\infty} x f(x;t)\mathrm{d}t \tag{3.5}$$

2. The *process variance*, i.e. the variance of the variable $\underline{x}(t)$:

$$\gamma_0(t) \coloneqq \mathrm{var}[\underline{x}(t)] = \int_{-\infty}^{\infty} \big(x - \mu(t)\big)^2 f(x;t)\mathrm{d}t \tag{3.6}$$

3. The process *autocovariance*, i.e. the covariance of the stochastic variables $\underline{x}(t)$ and $\underline{x}(t+h)$:

$$c(t;h) \coloneqq \mathrm{cov}[\underline{x}(t), \underline{x}(t+h)] = \mathrm{E}\left[\Big(\underline{x}(t) - \mu(t)\Big)\Big(\underline{x}(t+h) - \mu(t+h)\Big)\right] \tag{3.7}$$

   where $c(t;0) \equiv \gamma_0(t)$.

4. The process autocorrelation, i.e., the correlation coefficient of the stochastic variables $\underline{x}(t)$ and $\underline{x}(t+h)$):

$$r(t;h) \coloneqq \mathrm{corr}[\underline{x}(t), \underline{x}(t+h)] = \frac{c(t;h)}{\sqrt{\gamma_0(t)\gamma_0(t+h)}} \tag{3.8}$$

Additional characteristics will be given in section 3.5.

3.3 Stationarity

The term *process* has been introduced in the scientific vocabulary as synonymous to change, as evident in Kolmogorov's (1931) pioneering paper, in which he introduced the term *stochastic process*. This paper starts as "*A physical process* [is] *a change of a certain physical system*".

It is very common in science to try to identify invariant properties within change (Koutsoyiannis 2011a). For example, in the absence of an external force, the position of a body in motion changes in time but the velocity is unchanged (Newton's first law). If a constant force is present, then the velocity changes but the acceleration is constant (Newton's second law). If the force changes, e.g. the gravitational force with changing distance in planetary motion, the acceleration is no longer constant, but other invariant properties emerge, e.g. the angular momentum (Newton's law of gravitation; see also Koutsoyiannis 2011a).

Evidently, the notion of a stochastic process was invented to describe the irregular changes in natural systems more complex than the above, which are impossible to model deterministically in full detail or predict their future evolution in detail and with precision. Here, the great scientific achievement is the invention of macroscopic descriptions instead of modelling the details. This is essentially done using stochastics. Here lies the essence and usefulness of the stationarity concept, which seeks invariant properties in complex systems (Koutsoyiannis, 2011a, 2014a; Koutsoyiannis and Montanari, 2015). Following Kolmogorov (1931, 1938) and Khintchine (1934), a process is stationary if its statistical properties are invariant to a shift of time origin, i.e. $\underline{x}(t)$ and $\underline{x}(t')$ have the same (multivariate) distribution for any t and t'. Furthermore, following Kolmogorov (1947), a process is called *wide-sense stationary* if its mean is constant and its autocovariance depends only on time differences, i.e.:

$$\mathrm{E}\left[\underline{x}(t)\right] = \mu \text{ constant}, \qquad \mathrm{cov}\left[\underline{x}(t), \underline{x}(t+h)\right] = c(h) \tag{3.9}$$

A strict-sense stationary process is also wide-sense stationary but the inverse is not true.

A process that is not stationary is called nonstationary. In a nonstationary process one or more statistical properties depend on time, that is, they are *deterministic functions of time*. A typical case of a nonstationary process is a cumulative process whose mean is proportional to time. For instance, let us assume that the rainfall intensity at a geographical location and time of the year is modelled as a stationary process $\underline{x}(t)$, with mean μ. Let us further denote $\underline{X}(t)$ the rainfall depth collected in a large container (a cumulative raingauge) at time t and assume that at the time origin, $t = 0$, the container is empty, so that $\underline{X}(t) = \int_0^t x(s)\mathrm{d}s$. It is easy then to understand that $\mathrm{E}\left[\underline{X}(t)\right] = \mu t$. This is a deterministic (linear) function of time t and thus $\underline{X}(t)$ is a nonstationary process.

We should stress that stationarity and nonstationarity are properties of a process, not of a sample function or time series. There is some confusion in the literature about this, as a lot of studies assume that a time series is stationary or not, or can reveal whether the

process is stationary or not. As a general rule, to characterise a process nonstationary, it suffices to show that some statistical property is a *deterministic* function of time (as in the above example of the raingauge), but this cannot be straightforwardly inferred merely from a time series. A time series formed from observations of a natural process cannot be stationary, nor nonstationary.

Stochastic processes describing periodic phenomena, such as those affected by the annual cycle of Earth, are nonstationary. For instance, the daily temperature at a mid-latitude location could not be regarded as a stationary process. It could be modelled as a special kind of a nonstationary process with characteristics depending on time in a periodical manner (are periodic functions of time). Such processes are called *cyclostationary* processes.

3.4 Ergodicity

Stationarity is also related to another important stochastic concept, *ergodicity*.* Its importance derives from the fact that ergodicity is a prerequisite to make inference from data, that is, *induction*—the Aristotelian *επαγωγή* (*epagoge*). This is a type of inference weaker than *deduction*—the Aristotelian *συλλογισμός* (*syllogism*) —albeit very useful when deduction is not possible.

In dynamical systems, by definition (e.g. Mackey, 1992, p. 48), ergodicity is the property of a system whose all invariant sets under the dynamic transformation are trivial (have zero probability). In other words, in an ergodic transformation starting from any point, the trajectory of the system state will visit all other points, without being trapped to a certain subset. The ergodic theorem (Birkhoff, 1931; Khintchine, 1933; see also Mackey, 1992, p. 54), allows redefining ergodicity within the stochastic processes domain (Papoulis, 1991, p. 427; Koutsoyiannis 2010) in the following manner: A stochastic process $\underline{x}(t)$ is ergodic if the time average of any (integrable) function $g\left(\underline{x}(t)\right)$, as time tends to infinity, equals the true (ensemble) expectation, i.e.:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T g\left(\underline{x}(t)\right)\mathrm{d}t = \mathrm{E}[g(\underline{x}(t))] \tag{3.10}$$

for a continuous-time process or:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T} g\left(\underline{x}_\tau\right) = \mathrm{E}[g(\underline{x}(t))] \tag{3.11}$$

for a discrete-time process. The right-hand side in the above equations represents the true average, also known as *ensemble average*, whereas the left-hand side represents the time

* The concept of ergodicity was first conceived by Boltzmann (1884/85) who coined the terms *ergode* and *isodic*, both of which are etymologized from Greek words but which ones exactly is uncertain. Most probably, *ergodic* comes from the Greek *ἔργον* (ergon = work) and *ὁδός* (hodos = pathway). According to another interpretation, the second noun is *εἶδος* (eidos = form, kind, nature), or the whole word is a transliteration of the Greek adjective *ἐργώδης* (ergodes = laborious, troublesome; see Mathieu 1988).

average, for the limiting case of infinite time. The left-hand side of equations (3.10) and (3.11) is a stochastic variable (as a sum or integral of stochastic variables) and is not a function of the time *t*. Hence, the right-hand side should not be a function of the time *t*, i.e. the process should be stationary. Furthermore, the right-hand side is a number, not a stochastic variable. Equating a number with a stochastic variable implies that the stochastic variable has zero variance. This is precisely the condition that makes a process ergodic. And this allows the estimation (i.e. approximate calculation) of the true but unknown property $\mathrm{E}[g(\underline{x}(t))]$ from its time average, that is, from the available data. Without ergodicity inference from data would not be possible.

A stochastic process for which it can be shown that the property (3.10) or (3.11) holds true for the particular case that $g\left(\underline{x}(t)\right) = \underline{x}(t)$, whose expectation is the mean ($\mathrm{E}\left[\underline{x}(t)\right] = \mu$), is called *mean-ergodic*. The property could be extended for the multivariate functions $g(\,)$ and thus we can speak about *covariance-ergodic* processes. Further information, including conditions that should hold for ergodicity can be found in Papoulis (1991).

Now, if the system that is modelled in a stochastic framework has deterministic dynamics (meaning that a system input will give a single system response, as happens for example in most hydrological models), then a theorem applies (Mackey 1992, theorem 4.5 p. 52), according to which a dynamical system with dynamics $S_t(x)$ has a stationary probability density *if and only if* it is ergodic. Therefore, a stationary system is also ergodic and vice versa, and a nonstationary system is also non-ergodic and vice versa. Here we note that even if a system has deterministic dynamics, again it is legitimate to use a stochastic description, replacing the study of the evolution of system states $S_t(x)$ with the evolution of probability densities of states $f(x;t)$; one reason to prefer the stochastic description over the pure deterministic description is that the former incudes quantification of uncertainty, whereas the deterministic dynamics does not eliminate uncertainty (Koutsoyiannis 2010). Furthermore, we clarify that the deterministic description through the transformation $S_t(x)$ is fully compatible with a stochastic description that is stationary and ergodic, according to theorem stated above: while the system state is changing in time t according to the transformation $S_t(x)$, its statistical properties (and the probability density $f(x;t)$) can be constant in time (i.e. $f(x)$).

If the system dynamics is stochastic (a single input could result in multiple outputs), then ergodicity and stationarity do not necessarily coincide. However, recalling that a stochastic process is a model and not part of the real world, we can always conveniently device a stochastic process that is ergodic, provided that we have excluded nonstationarity. In conclusion, from a practical point of view ergodicity can always be assumed when there is stationarity, while this assumption if fully justified by the theory if the system dynamics is deterministic. Conversely, if nonstationarity is assumed, then ergodicity cannot hold, which forbids inference from data. This contradicts the basic premise in geosciences, including hydrology and climatology, where data are the only reliable information in building models and making inference and prediction.

Digression 3.A: Misuses of stationarity and ergodicity

The concepts of stationarity and ergodicity, despite having a central role in stochastics, have been widely misunderstood and broadly misused (Montanari and Koutsoyiannis, 2014; Koutsoyiannis and Montanari, 2015). In an attempt to find trends everywhere, according to the popular motto "stationarity is dead" (Milly et al. 2008), trend analysis of hydroclimatic processes is more fashionable today than ever before (Iliopoulou and Koutsoyiannis, 2020). The notion of a trend, as a fundamental constituent of time series, is very old, but it is fundamentally problematic (Koutsoyiannis, 2020a), despite being popular.

Ironically, most of these studies use time series data to estimate statistical properties, as if the process were ergodic, while at the same time what they cursorily estimate falsifies the ergodicity hypothesis. The correct tactic, even when dealing with provably nonstationary and nonergodic processes and our study is based on data, is to convert the process to a stationary and ergodic one before trying to make any inference from the data.

As an example, assuming that we deal with the cumulative rainfall process $\underline{X}(t)$, used as an example in section 3.3, we convert the process into a stationary one in discrete time by $\underline{x}_\tau \coloneqq \underline{X}(\tau D) - \underline{X}((\tau-1)D)$, where D is a time step, and perform the same transformation to the time series data. Then we can use the x_τ data to make inferences.

As a second example related to trends, let us examine a statement such as: "By analysing the time series x_τ (where τ denotes discrete time), we concluded that it is nonstationary and we identified an increasing trend with slope b." This is an incorrect statement and can be corrected in the following manner: "We analysed the time series x_τ based on the hypothesis that the stochastic process $\underline{x}_\tau - b\tau$ is stationary and ergodic, which enabled the estimation of the slope b." The latter statement respects the fact that we always need stationarity and ergodicity to make inference from data.

3.5 Second order characteristics of stochastic processes

We have already provided, along with the definition of a stochastic process (section 3.1), the definition of the autocovariance function, an important characteristic of the second-order distribution function of a stochastic process. However, there are other second-order characteristics that are useful to study, as they have good properties which help understand and simulate stochastic processes.

Before defining them, starting from the process of interest $\underline{x}(t)$ we will better explain the concepts of the cumulative process $\underline{X}(t)$ and the discrete-time process $\underline{x}_\tau$, which have already been introduced. As graphically shown in Figure 3.1, the cumulative process is defined as:

$$\underline{X}(t) \coloneqq \int_0^t \underline{x}(u)du \tag{3.12}$$

where obviously $\underline{X}(0) \equiv 0$. If $\underline{x}(t)$ aims to represent a natural processes, then $\underline{X}(t)$ should necessarily be nonstationary. However, the averaged process, $\underline{X}(t)/t$, will be stationary, provided that $\underline{x}(t)$ is stationary. With the help of the cumulative process, the discrete-time representation of the process (equation (3.1)) can also be written as:

$$\underline{x}_\tau^{(D)} \coloneqq \frac{1}{D}\int_{(\tau-1)D}^{\tau D} \underline{x}(u)\mathrm{d}u = \frac{\underline{X}(\tau D) - \underline{X}\big((\tau-1)D\big)}{D} \tag{3.13}$$

The superscript (D) in $\underline{x}_\tau^{(D)}$ denotes the time step of discretization; in cases that we use a single discretization step and there is no ambiguity we will omit it, writing $\underline{x}_\tau$.

The variance of $\underline{X}(t)$ at time t, i.e.:

$$\Gamma(t) \coloneqq \mathrm{var}\left[\underline{X}(t)\right] = t^2\gamma(t) \tag{3.14}$$

is known as *cumulative climacogram*. The variance of the averaged process $\underline{X}(k)/k$ at a time scale k, as a function of time scale k, is the continuous-time variant of the *climacogram*, already discussed in sections 1.3 and 2.16:

$$\gamma(k) \coloneqq \mathrm{var}\left[\frac{\underline{X}(k)}{k}\right] = \frac{\Gamma(k)}{k^2} \tag{3.15}$$

The autocovariance function $c(h)$ of the continuous-time process $\underline{x}(t)$ for time lag h, already defined in equation (3.7), is related to the climacogram by (Koutsoyiannis 2016):

$$c(h) \coloneqq \mathrm{cov}[x(t), x(t\ +\ h)] = \frac{1}{2}\frac{\mathrm{d}^2\Gamma(h)}{\mathrm{d}h^2} \tag{3.16}$$

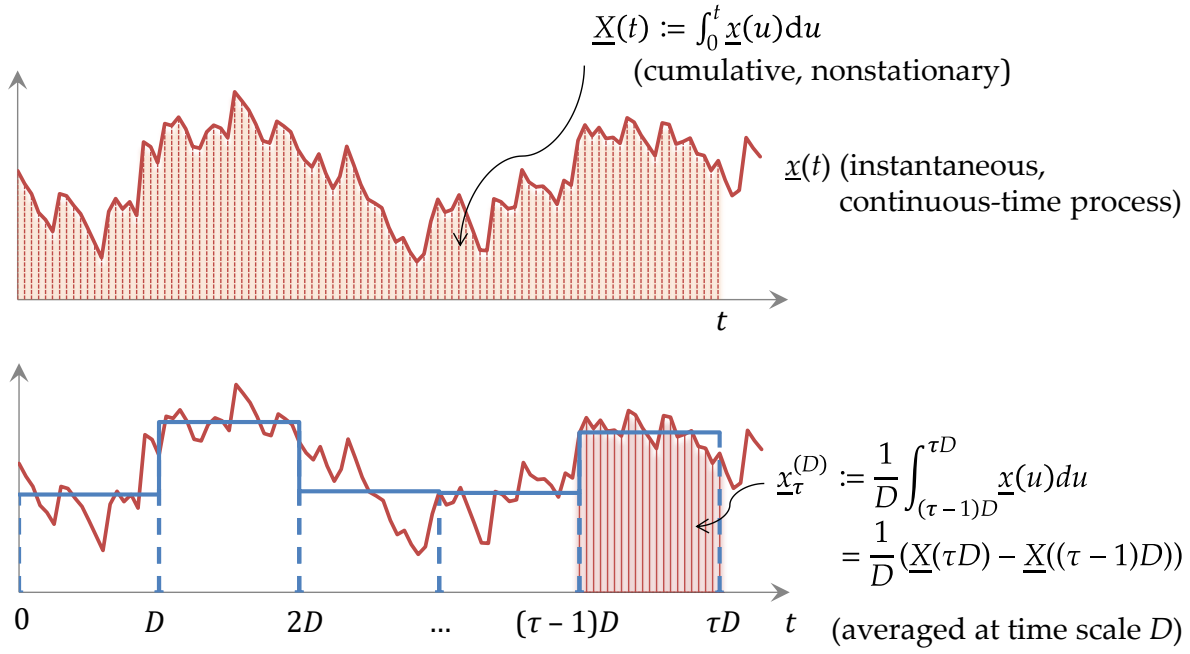

Figure 3.1 Explanatory sketch for a stochastic process in continuous time and its representation in discrete time. Note that the graphs display a realization of the process (it is impossible to display the process as such) while the notation is for the process per se.

If we deal with two processes $\underline{x}(t)$ and $\underline{y}(t)$ we can define the cross-covariance:

$$c_{xy}(h) \coloneqq \mathrm{cov}\left[\underline{x}(t), \underline{y}(t+h)\right] \tag{3.17}$$

This is a continuous-time metric. If we wish to involve also the time scale k of the averaged process, we can define the *cross-climacogram* (Koutsoyiannis, 2019b):

$$\gamma_{xy}(k;\eta) \coloneqq \sigma_x\sigma_y\,\mathrm{var}\left[\frac{\underline{X}(k)}{k\sigma_x} + \frac{\underline{Y}((\eta+1)k) - \underline{Y}(\eta k)}{k\sigma_y}\right] \tag{3.18}$$

where $\underline{Y}(k) \coloneqq \int_0^k \underline{y}(t)\mathrm{d}t$ and η is lag.

The *structure function* (also known as *semivariogram* or *variogram*), $v(h)$, is another second-order tool, defined as:

$$v(h) \coloneqq \frac{1}{2}\mathrm{Var}\left[\underline{x}(t) - \underline{x}(t+h)\right] = c(0) - c(h) \tag{3.19}$$

The *power spectrum* (also known as *spectral density*), $s(w)$, where w denotes frequency is defined as the Fourier transform of the autocovariance function, i.e.:

$$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h \tag{3.20}$$

The power spectrum should necessarily be nonnegative at all w ($s(w) \geq 0$), and this entails that the autocovariance $c(h)$ should be a positive definite function. Also, the climacogram $\gamma(k)$ should be a positive definite function (Koutsoyiannis, 2017).

The power spectrum has some analogies with another stochastic tool, the so-called climacospectrum (Koutsoyiannis, 2017), which is directly given in terms of the climacogram. Specifically, it is proportional to the difference of the variances of the averaged process at time scales k and $2k$:

$$\psi(k) \coloneqq \frac{k\big(\gamma(k) - \gamma(2k)\big)}{\ln 2} \tag{3.21}$$

The climacospectrum is also written in an alternative manner in terms of frequency $w = 1/k$:

$$\tilde{\psi}(w) \coloneqq \psi(1/w) = \frac{\gamma(1/w) - \gamma(2/w)}{(\ln 2)\,w} \tag{3.22}$$

It is useful to note that the entire area under the power spectrum $s(w)$, as well as that under the curve $\tilde{\psi}(w)$, are precisely equal to each other and to the variance γ_0.

All definitions of second-order characteristics in continuous time are gathered together in Table 3.1. Once one of these characteristics is known in the continuous-time representation, we can also calculate any other of them in continuous time as well as those in discrete time, as shown in Table 3.2. The reverse is not true, i.e., from a model formulated in discrete time we cannot infer precisely the characteristics of the continuous-time representation. It may be seen in Table 3.2 that the expressions of the discrete time characteristics may differ substantially from those in continuous time, and thus attention is needed to avoid confusion and misuse. The climacogram and the climacospectrum are exceptions, as they are not affected by discretization (they admit the same expressions for both continuous and discrete time), and have some additional advantages, such as simplicity, close relationship to entropy (see below), and more stable behaviour (Dimitriadis and Koutsoyiannis, 2015a; Koutsoyiannis, 2016; 2017) which make them the preferable tool in stochastic modelling—even though they are less popular than other tools.

All these tools are transformations of one another, so that if we are given one, we can calculate precisely any other of them. The transformations are listed in Table 3.3.

Table 3.1 Summary of notation and second order characteristics of a stationary stochastic process in continuous time.

Name	Symbol and definition	Remarks	Eqn. no.
Stochastic process of interest	$\underline{x}(t)$	Assumed stationary	
Time, continuous	t	Dimensional	
Cumulative process	$\underline{X}(t) \coloneqq \int_0^t \underline{x}(\xi)\mathrm{d}\xi$	Nonstationary	(3.12)
Variance, instantaneous	$\gamma_0 \coloneqq \mathrm{var}[\underline{x}(t)]$	Constant (not a function of t)	(3.6)
Cumulative climacogram	$\Gamma(t) \coloneqq \mathrm{var}\left[\underline{X}(t)\right]$	$\Gamma(0) \equiv 0$	(3.14)
Climacogram	$\gamma(k) \coloneqq \mathrm{var}\left[\dfrac{\underline{X}(k)}{k}\right] = \dfrac{\Gamma(k)}{k^2}$	A function of time scale, $\gamma(0) = \gamma_0$	(3.15)
Time scale, continuous	k	Units of time	
Autocovariance function	$c(h) \coloneqq \mathrm{cov}[x(t), x(t+h)]$	$c(0) = \gamma_0$	(3.16)
Time lag, continuous	h	Units of time	
Structure function (semivariogram, variogram)	$v(h) \coloneqq \frac{1}{2}\mathrm{Var}\left[\underline{x}(t) - \underline{x}(t+h)\right]$	$v(h = \gamma_0 - c(h)$	(3.16)
Power spectrum (spectral density)	$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h$	$\int_0^\infty s(w)\mathrm{d}w = \gamma_0$	(3.20)
Frequency, continuous	$w = 1/k$	Units of inverse time	

Table 3.2 Summary of notation and second-order characteristics of a stationary stochastic process in discrete time.

Name	Symbol and definition	Remarks	Eqn. no.
Stochastic process, discrete time	$\underline{x}_\tau \coloneqq \dfrac{1}{D}\int_{(\tau-1)D}^{\tau D} \underline{x}(u)\mathrm{d}u = \dfrac{\underline{X}(\tau D) - \underline{X}\big((\tau-1)D\big)}{D}$		(3.13)
Discretization time step	D	Length of time window of averaging	
Time, discrete	$\tau \coloneqq t/D$	Dimensionless	
Climacogram	$\gamma_\kappa = \gamma(\kappa D) = \dfrac{\Gamma(\kappa D)}{(\kappa D)^2}$	$\gamma_1 = \mathrm{var}\left[\underline{x}_\tau\right] = \gamma(D)$	(3.23)
Time scale, discrete	$\kappa = k/D$	Dimensionless	(3.24)
Autocovariance function	$c_\eta \coloneqq \mathrm{cov}\left[\underline{x}_\tau, \underline{x}_{\tau+\eta}\right]$	$c_0 = \gamma(D) = \gamma_1$	
Time lag, discrete	$\eta = h/D$	Dimensionless	(3.25)
Structure function	$v_\eta = \gamma_1 - c_\eta$		(3.26)
Power spectrum	$s_\mathrm{d}(\omega) = \dfrac{1}{D}\sum_{j=-\infty}^{\infty} s\left(\dfrac{\omega + j}{D}\right)\mathrm{sinc}^2\big(\pi(\omega + j)\big)$		(3.27)
Frequency, discrete	$\omega = wD = 1/\kappa$	Dimensionless	(3.28)

Note: In time-related quantities, Latin letters denote dimensional quantities and Greek letters dimensionless ones, as specified above.

Table 3.3 Relationships between second-order characteristics of a stochastic process.

Related characteristics	Symbol and definition	Inverse relationship	Eqn. no.
$\gamma(k) \leftrightarrow c(h)$	$\gamma(k) = 2\int_0^1 (1-\chi)c(\chi k)\mathrm{d}\chi$	$c(h) = \frac{1}{2}\frac{\mathrm{d}^2\left(h^2\gamma(h)\right)}{\mathrm{d}h^2}$	(3.29)
$s(w) \leftrightarrow c(h)$	$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h$	$c(h) = \int_0^\infty (w)\cos(2\pi w h)\,\mathrm{d}w$	(3.30)
$\gamma(k) \leftrightarrow s(w)$	$\gamma(k) = \int_0^\infty s(w)\ \mathrm{sinc}^2(\pi w k)\,\mathrm{d}w$	$s(w) \coloneqq 2\int_0^\infty \frac{\mathrm{d}^2\left(h^2\gamma(h)\right)}{\mathrm{d}h^2}\cos(2\pi w h)\,\mathrm{d}h$	(3.31)
$v(h) \leftrightarrow c(h)$	$v(h) = \gamma_0 - c(h)$	$c(h) = v(\infty) - v(h)$ with $v(\infty) = \gamma_0$	(3.32)
$\psi(k) \leftrightarrow \gamma(k)$	$\psi(k) \coloneqq \frac{k\left(\gamma(k) - \gamma(2k)\right)}{\ln 2}$	$\gamma(k) = \ln 2\sum_{i=0}^{\infty}\frac{\psi(2^i k)}{2^i k} = \gamma(0) - \ln 2\sum_{i=1}^{\infty}\frac{\psi(2^{-i}k)}{2^{-i}k}$	(3.33)
$\gamma_\kappa \equiv \gamma(\kappa D) \leftrightarrow c_\eta$	$\gamma_\kappa = \frac{1}{\kappa}\left(c_0 + 2\sum_{\eta=1}^{\kappa-1}\left(1-\frac{\eta}{\kappa}\right)c_\eta\right) = \frac{\Gamma(\kappa D)}{(\kappa D)^2}$ where $\Gamma(0)=0, \Gamma(D)=c_0 D^2$ and, recursively, $\Gamma(\kappa D) = 2\Gamma\left((\kappa-1)D\right) - \Gamma\left((\kappa-2)D\right) + 2c_{j-1}D^2$	$c_\eta = \frac{1}{D^2}\left(\frac{\Gamma(\lvert\eta+1\rvert D) + \Gamma(\lvert\eta-1\rvert D)}{2} - \Gamma(\lvert\eta\rvert D)\right)$	(3.34)
$c_\eta \leftrightarrow s_\mathrm{d}(\omega)$	$s_\mathrm{d}(\omega) = 2c_0 + 4\sum_{\eta=1}^{\infty} c_\eta\cos(2\pi\eta\omega)$	$c_\eta = \int_0^{1/2} s_\mathrm{d}(\omega)\cos(2\pi\omega\eta)\,\mathrm{d}\omega$	(3.35)
$v_\eta \leftrightarrow c_\eta$	$v_\eta = \gamma(D) - c_\eta$	$c_\eta \coloneqq \gamma(D) - v_\eta$	(3.36)

Digression 3.B: What is dependence in time?

Dependence of a stochastic process in time (also known as intertemporal dependence or simply time dependence) is typically expressed by the autocovariance or the autocorrelation function. In turn, its typical interpretation is *memory*. This has been so common than in many texts the term *memory* has replaced the term *dependence*—even in the titles of several publications, papers and books. Perhaps the scientist who was most influential in establishing this interpretation was Mandelbrot (for example, Mandelbrot and Wallis, 1968, speak about short and long memory, both of which they contrast to independence), even though other scientists had used the term before (e.g. Krumbein, 1968). Clearly, in stochastics the term memory is metaphorical, while in other disciplines (neuropsychology, computer science) it is literal. In science there is no reason to use a metaphorical term when we have a literal term, particularly when the metaphorical term has another scientific meaning.

Perhaps the metaphorical term *memory* distracts, rather than helps, intuition and understanding of time dependence in a stochastic process. In particular, its variant *long memory* is totally inappropriate as it stimulates people to imagine a mechanism inducing long memory (e.g. hundreds of years) and of course it is difficult to conceptualize such a mechanism. A better interpretation is a mechanism producing change, rather that recalling information (as is the meaning of memory). And indeed, changes produce dependence—not the other way round. Furthermore, dependence and change need not be interpreted as nonstationarity as many think.

But before discussing how change produces time dependence in a process that is stationary, we will discuss how dependence manifests itself into a time series. In one word, this manifestation is through *patterns*. In pure randomness, without time dependence (like in a sequence of dice outcomes or in the sequence of digits of π) no patterns appear. To better illustrate such patterns, we examine several time series with a small length, $n = 16$. For convenience we make these time series two-valued, with values –1 and 1 and with average of the 16 values equal to zero, which means that eight values will be –1 and eight 1. The *estimates* of the variance, the lag-one autocovariance and the lag-one autocorrelation coefficient will thus be, respectively:

$$\hat{\gamma}_1 = \frac{1}{16}\sum_{\tau=1}^{16} x_\tau^2 = 1, \qquad \hat{c}_1 = \frac{1}{16}\sum_{\tau=1}^{16} x_\tau x_{\tau+1}\,, \qquad \hat{r}_1 = \frac{\hat{c}_1}{\hat{\gamma}_1} = \hat{c}_1$$

where we set $x_{17} = x_1$ in order to have 16 terms in the sum for $\hat{c}_1$ and thus make possible values up to ±1 (even though this is not typically made in analyses of time series). The formal meaning of the term *estimate* is clarified in section 4.3.

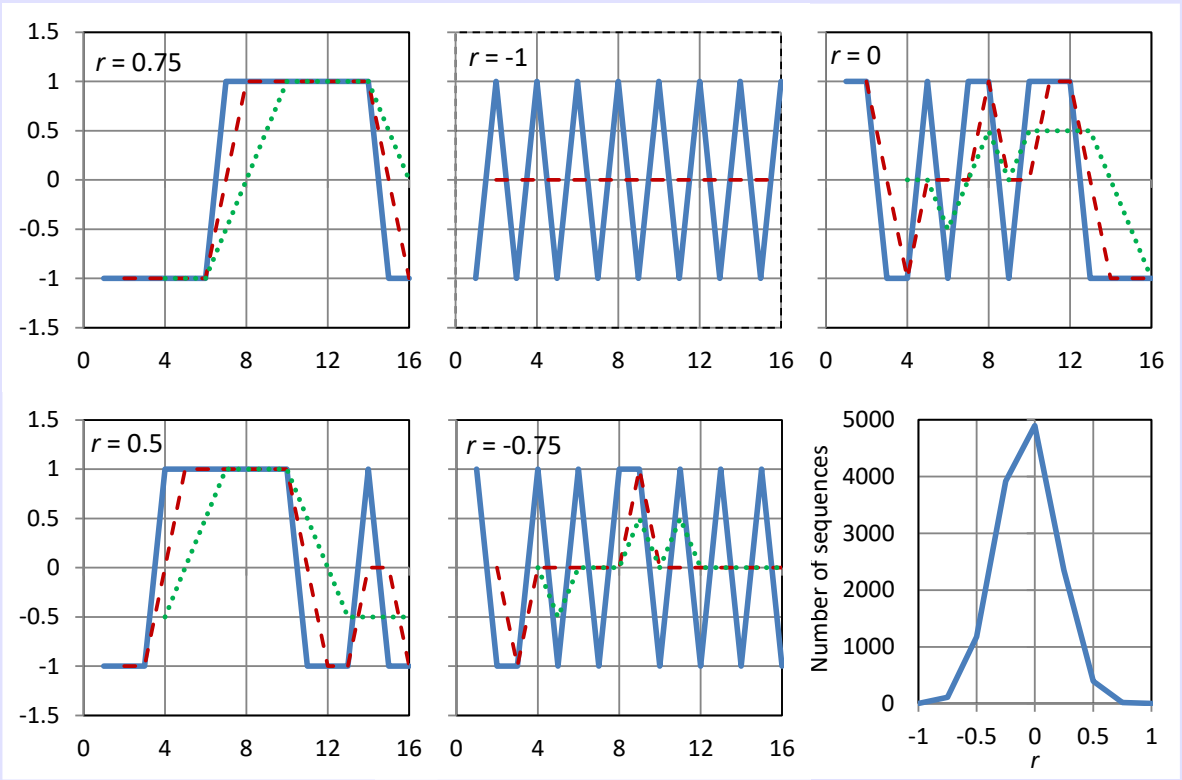

Figure 3.2 Examples of arrangements of eight ones and eight minus ones in the form of time series with length 16, mean zero and unit variance, along with the resulting estimate of the lag-one autocorrelation coefficients *r*. In addition to the original time series (scale 1; continuous line), time-averaged time series are also shown at scales 2 (dashed lines) and 4 (dotted lines). In the bottom right panel, the frequency distribution of *r* for all $16!/(8!)^2 = 12\,870$ possible cases (permutations) are shown.

Some instances of such time series are shown in Figure 3.2. In the upper left panel, all eight ones are grouped together so that $\sum_{\tau=1}^{16} x_\tau x_{\tau+1} = 7 + 7 - 2 = 12$ and $\hat{r}_1 = 075$. This is the highest possible value that a particular arrangement of 16 items, each being ±1, can give. Obviously, there are 16 possible arrangements that will give $\hat{r}_1 = 0.75$. If our time series had length of N, the highest $\hat{r}_1$ would be $(N-4)/N = 1 - 4/N$ and would approach the value +1 for large N. Consequently, a large autocorrelation is caused by grouping together of similar (in our example same) values, and this grouping has been termed *persistence*. If the grouping appears but is not that "perfect", such as in the lower left panel, then again, the autocorrelation will be positive but lower ($\hat{r}_1 = 0.5$ in this example).

In contrast, if the patterns appear to be of alternating, rather than grouping, type, then the autocorrelation coefficient is negative. Thus, in the "perfect" alternating shape of the upper middle panel of Figure 3.2 we have $\sum_{\tau=1}^{16} x_\tau x_{\tau+1} = -16$ and $\hat{r}_1 = -1$. In the lower middle panel alternation is not perfect and $\hat{r}_1 = -0.75$. Finally, the upper right panel is free of patterns and $\hat{r}_1 = 0$.

Now, the effect of change is illustrated in Figure 3.3, where we plot a time series generated from the normal distribution without time dependence. We now assume that the process is affected by a mechanism producing change, namely shifts up and down, at random points in time. As illustrated in Figure 3.3 and detailed in the figure caption, in this case patterns are produced and (positive) autocorrelation is induced.

Had such change been describable in deterministic terms, as a deterministic function of time, that is, had it been precisely predictable in terms of location of times where it occurs and in terms of magnitude of state shifts, we would speak about nonstationarity. But since, as we said, the points of change are random points in time, they resist a deterministic description and the entire process with the change producing mechanism is a *stationary stochastic process with dependence*. Unfortunately, this simple truth is not widely understood and therefore the inconsistent interpretations of change as nonstationarity abound in hydroclimatic literature.

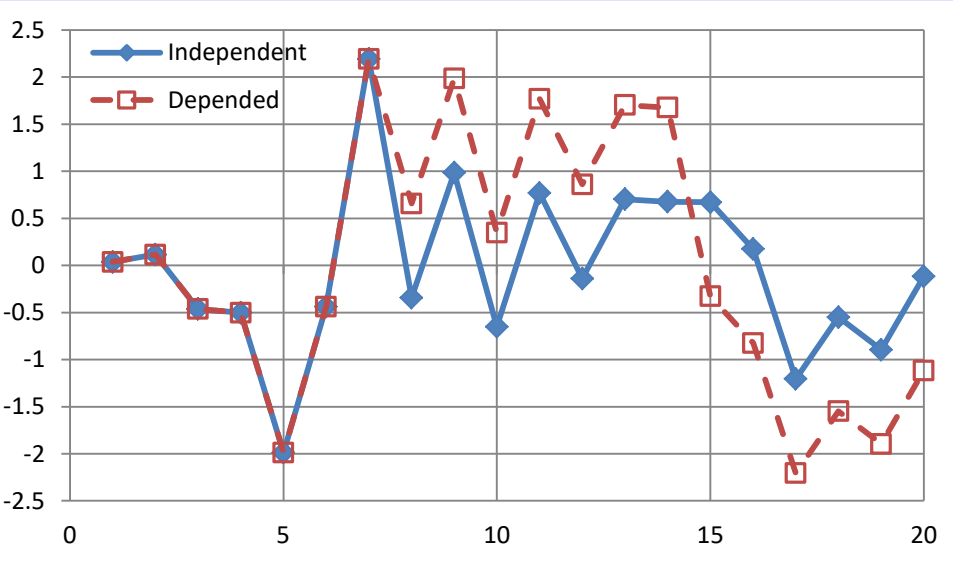

Figure 3.3 Illustration of the fact that change causes autocorrelation using a time series of length 20, generated from the normal distribution N(0,1) without time dependence; the estimates of the statistical characteristics from the time series, plotted as full points connected with continuous lines, are $\hat{\mu} = -0.05, \hat{\gamma}_0 = 0.9^2, \hat{r}_1 = 0.05$. By shifting a time segment up (by +1, items 8-14) and another segment down (by –1, items 15-20) we obtain a new time series (empty points connected with dashed lines) in which the autocorrelation has become $\hat{r}_1 = 0.59$.

3.6 Asymptotic power laws and the log-log derivative

It is quite common that nonnegative functions $f(t)$ defined in $[0, \infty)$, whose limits at 0 and ∞ exist, are associated with asymptotic *power laws* as $t \to 0$ and ∞ (Koutsoyiannis, 2014b, 2017). *Power laws* are functions of the form

$$f(t) \propto t^b \tag{3.37}$$

A power law is visualized on a graph of $f(t)$ plotted against t with logarithmic axes, so that the plot forms a straight line with slope b. Formally, the slope b is expressed by the *log-log derivative* (LLD):

$$f^{\#}(t) \coloneqq \frac{\mathrm{d}(\ln f(t))}{\mathrm{d}\,(\ln t)} = \frac{tf'(t)}{f(t)} \tag{3.38}$$

If the power law holds for the entire domain, then $f^{\#}(t) = b =$ constant. In this case we speak about a *simple scaling* behaviour. Most often, however, $f^{\#}(t)$ is not constant. Of particular interest are the *asymptotic values* for $t \to 0$ and ∞, symbolically $f^{\#}(0)$ and $f^{\#}(\infty)$, which define *two asymptotic power laws*. We note that, if $0 < f(0) < \infty$, then $f^{\#}(0) = 0$, which means that $f(0)$ has to be 0 or ∞ in order for $f^{\#}(0) \neq 0$. Basic properties of LLD are given in Table 3.4.

Table 3.4 Basic properties of LLD (from Koutsoyiannis, 2017).

Description	Mathematical formula
Multiplication and addition by constants	$(\lambda\, f(t) + \mu)^{\#} = f^{\#}(t)$
Sum of two functions	$\big(f_1(t) + f_2(t)\big)^{\#} = \dfrac{f_1(t)f_1^{\#}(t) + f_2(t)f_2^{\#}(t)}{f_1(t) + f_2(t)}$
Product of two functions	$\big(f_1(t)f_2(t)\big)^{\#} = f_1^{\#}(t) + f_2^{\#}(t)$
Quotient of two functions	$\big(f_1(t)/f_2(t)\big)^{\#} = f_1^{\#}(t) - f_2^{\#}(t)$
Raise to a power	$\big(f(t)^{\lambda}\big)^{\#} = \lambda f^{\#}(t)$
Function composition	$\big((f \circ g)(t)\big)^{\#} = \big(f(g(t)\big)^{\#} = f^{\#}\big(g(t)\big)\, g^{\#}(t)$

In particular, the asymptotic properties of the second order characteristics of a stochastic process for $t \to 0$, where now t denotes time, characterize the *local behaviour* of a process, while those for $t \to \infty$ characterize the *global behaviour*. We will discuss these properties in section 3.8, after introducing the related concept of entropy production in section 3.7.

3.7 Entropy production in stochastic processes

In a stochastic process the change of uncertainty in time can be quantified by the *entropy production,* i.e. the time derivative of the entropy $\Phi\big[\underline{X}(t)\big]$ of the cumulative process $\underline{X}(t)$ (Koutsoyiannis, 2011b):

$$\Phi'\big[\underline{X}(t)\big] \coloneqq \frac{\mathrm{d}\Phi\big[\underline{X}(t)\big]}{\mathrm{d}t} \tag{3.39}$$

A more convenient (and dimensionless) measure is the *entropy production in logarithmic time* (EPLT):

$$\varphi(t) \equiv \varphi\big[\underline{X}(t)\big] \coloneqq \Phi'\big[\underline{X}(t)\big]t \equiv \frac{\mathrm{d}\Phi\big[\underline{X}(t)\big]}{\mathrm{d}(\ln t)} \tag{3.40}$$

For a Gaussian process, the entropy depends on its variance $\Gamma(t)$ only and is given as (cf. Papoulis, 1991):

$$\Phi\big[\underline{X}(t)\big] = \frac{1}{2}\ln(2\pi\mathrm{e}\Gamma(t)\beta^2) \tag{3.41}$$

where β is the background measure density, assumed to be constant (Lebesgue). The EPLT of a Gaussian process is thus easily shown to be:

$$\varphi(t) = \frac{\Gamma'(t)t}{2\Gamma(t)} = 1 + \frac{\gamma'(t)t}{2\gamma(t)} = \frac{\Gamma^{\#}(t)}{2} = 1 + \frac{\gamma^{\#}(t)}{2} \tag{3.42}$$

That is, EPLT is visualized and estimated by the slope of a log-log plot of the climacogram. We note that if, because of using the cumulative process, the background measure was taken βt insteaf of β, the result would be practically the same (plus a constant 1).

When the past and the present are observed, instead of the unconditional variance $\gamma(t)$ we should use a variance $\gamma_{\mathrm{C}}(t)$ conditional on the known past and present. This turns out to equal a differenced climacogram (Koutsoyiannis, 2017):

$$\gamma_{\mathrm{C}}(k) = \varepsilon\big(\gamma(k) - \gamma(2k)\big), \qquad \varepsilon = \frac{1}{1 - 2^{\gamma^{\#}(\infty)}} \tag{3.43}$$

We can subsequently define the *conditional entropy production in logarithmic time* (CEPLT) in a manner analogous to (3.42). By also considering the definition of the climacospectrum in (3.21) and (3.22), CEPLT can be written as:

$$\varphi_{\mathrm{C}}(t) = 1 + \frac{\gamma_{\mathrm{C}}^{\#}(t)}{2} = \frac{1 + \psi^{\#}(t)}{2} = \frac{1 - \tilde{\psi}^{\#}(1/t)}{2} \tag{3.44}$$

Thus, for a Gaussian process the conditional entropy production is given in terms of log-log slope of the process climacospectrum. We will use the same result as an approximation for non-Gaussian processes too, even though in a non-Gaussian process the entropy expression becomes more complicated than (3.41) with other terms additional to variance.

3.8 Asymptotic scaling of second order properties

EPLT and the CEPLT are related to LLDs of second order tools such as climacogram, climacospectrum, power spectrum, etc. With a few exceptions, these slopes are nonzero asymptotically, hence entailing asymptotic scaling or asymptotic power laws with the LLDs being the scaling exponents. It is intuitive to expect that an emerging asymptotic scaling law would provide a good approximation of the true law for a range of scales.

If the scaling law was appropriate for the entire range of scales, then we would have a simple scaling law. Such simple scaling sounds attractive from a mathematical point of view, but it turns out to be impossible in physical processes (Koutsoyiannis, 2017; Koutsoyiannis et al., 2018; see also below). It is thus physically more realistic to expect two different types of asymptotic scaling laws, one in each of the ends of the continuum of scales. The respective scaling exponents are given in terms of two parameters, *M* (to give credit to Mandelbrot) and *H* (to give credit to Hurst) according to the following relationships:

- The parameter *M* characterizes the *local scaling* or *smoothness* or *fractal behaviour*, when $k \to 0$ or $w \to \infty$:

$$M \coloneqq \varphi_{\mathrm{C}}(0) - 1 = \frac{\gamma_{\mathrm{C}}^{\#}(0)}{2} = \frac{v^{\#}(0)}{2} = \frac{\psi^{\#}(0) - 1}{2} = \frac{-s^{\#}(\infty) - 1}{2} \tag{3.45}$$

- The parameter H characterizes the *global scaling* or *persistence* or *Hurst-Kolmogorov* behaviour, when $k \to \infty$ or $w \to 0$:

$$H \coloneqq \varphi_{\mathrm{C}}(\infty) = 1 + \frac{\gamma_{\mathrm{C}}^{\#}(\infty)}{2} = 1 + \frac{\gamma^{\#}(\infty)}{2} = 1 + \frac{c^{\#}(\infty)}{2} = \frac{\psi^{\#}(\infty) + 1}{2} = \frac{-s^{\#}(0) + 1}{2} \tag{3.46}$$

These scaling behaviours have emerged from maximum entropy considerations, and this may provide the theoretical background in modelling complex natural processes by such scaling laws. Generally, scaling laws are a mathematical necessity and could be constructed for virtually any continuous function defined in [0, ∞). In other words, there is no magic in power laws, except that they are, logically and mathematically, a necessity.

3.9 Bounds of scaling

Both parameters M and H take on values in the interval (0,1) (with the limiting cases $M = 1$ and $H = 0$ being possible). This fact, combined with equations (3.45) and (3.46), defines limits of the possible scaling laws in natural processes. The limits are not quite well known and several studies have reported values out of the limits (see Digression 3.C for an example about how to avoid such a mistaken result).

For the global behaviour, it has been shown (Koutsoyiannis et al., 2018) that a process with $-s^{\#}(0) > 1$ is nonergodic. As already explained, inference from data is only possible when the process is ergodic and thus, claiming that $-s^{\#}(0) > 1$ based on data is self-contradictory. Steep slopes ($-s^{\#}(w) > 1$) are mathematically and physically possible for medium and large w and indeed they are quite frequent in geophysical and other processes. Because of the equality of slopes of power spectrum and climacospectrum, the ergodicity limitation holds also for the slope of the climacospectrum, i.e., $\psi^{\#}(\infty) = -\tilde{\psi}^{\#}(0) < 1$. On the other hand, too steep negative asymptotic slopes of the climacospectrum are also impossible. Indeed (because of (52)), $\psi^{\#}(k) = -\tilde{\psi}^{\#}(1/k) < -1$ would entail $\varphi_{\mathrm{C}}(k) < 0$ and $\Gamma_{\mathrm{C}}'(k) < 0$ (Koutsoyiannis, 2017). This means that the variance of the cumulative process would be a decreasing function of time, which is absurd. This holds both for the global case ($k \to \infty$, in which the conditional variance $\Gamma_C(\infty)$ equals the unconditional $\Gamma(\infty)$) and the local case ($k \to 0$, for the conditional variance $\Gamma_{\mathrm{C}}(0)$).

For the local behaviour, there is another severe limitation imposed by physical reasoning. The case $\psi^{\#}(0) = -s^{\#}(\infty) < 1$ would entail infinite variance. Infinite variance would require infinite energy to emerge, which is physically inconsistent (see also section 2.17). Therefore, the physical lower limit for $\psi^{\#}(0) = -s^{\#}(\infty)$ is 1. A final—and quite severe—limitation is an upper bound of the local scaling exponent, which is 3 for $\psi^{\#}(0) = -s^{\#}(\infty)$ (Koutsoyiannis, 2017). The problem if this limitation is violated is that the resulting autocovariance function is not positive definite or, equivalently, that the

resulting power spectrum is not always (for any frequency w) positive but takes on negative values for some w. Likewise, the Fourier transform of the climacogram takes on negative values for some w. Proof is provided in Koutsoyiannis (2017).

The above limits define the "green square" of admissible values of φ_C, M and H in Figure 3.4, which is also depicted in terms of admissible values of slopes $\psi^{\#}$ and $s^{\#}$ (noting that $s^{\#}$ can, by exception, take on values out of the square when $\varphi_C(0) = 2$ or $\varphi_C(\infty) = 0$). The reasons why a process out of the square would be impossible or inconsistent, as discussed above, are also marked in the figure.

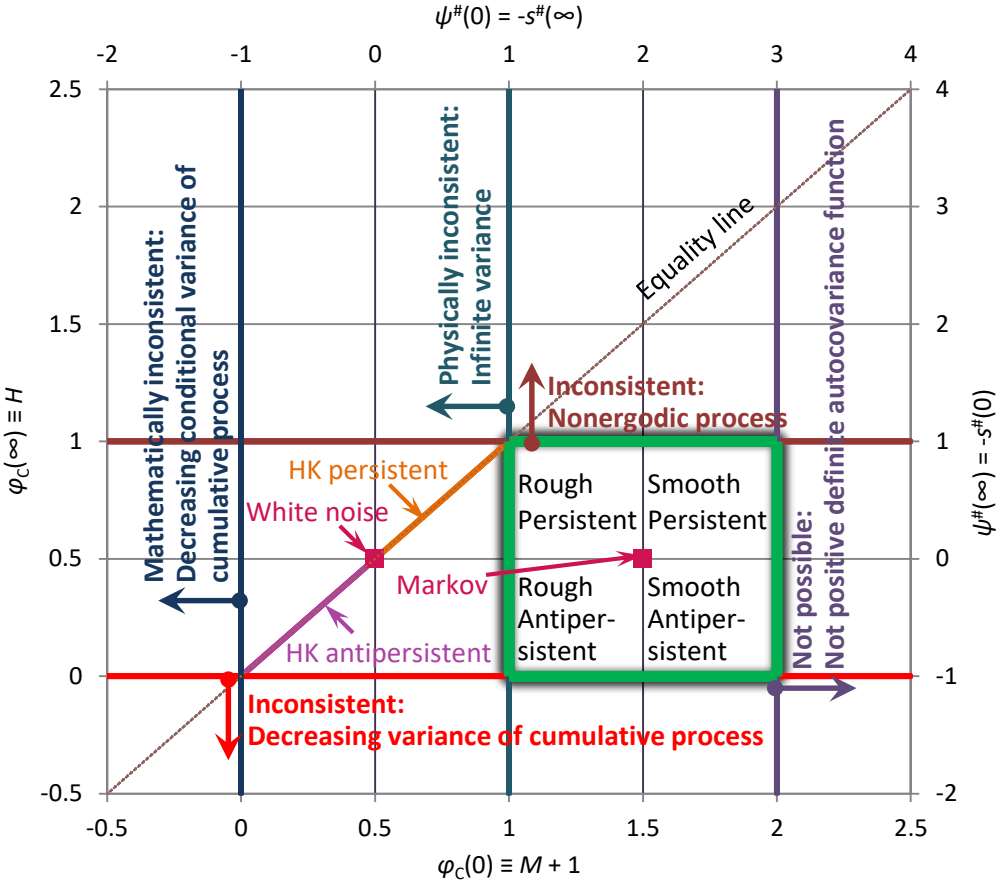

Figure 3.4 Bounds of asymptotic values of CEPLT, $\varphi_C(0)$ and $\varphi_C(\infty)$, and corresponding bounds of the log-log slopes of power spectrum and climacospectrum. The "green square" represents the admissible region; note that $s^{\#}$ can, by exception, take on values out of the square when $\varphi_C(0) = 2$ $(M = 1)$ or $\varphi_C(\infty) = 0$ $(H = 0)$. The reasons why a process out of the square would be impossible or inconsistent are also marked. The lines $\varphi_C(0) = 3/2$ $(M = 1/2)$ and $\varphi_C(\infty) = 1/2$ $(H = 1/2)$ define *neutrality* (which is represented by a Markov process) and support the classification of stochastic processes into the indicated four categories (smaller squares within the "green square").

The centre of the square, with coordinates $\varphi_C(0) = 3/2$, $\varphi_C(\infty) = 1/2$ represents a neutral process, whose typical representative is the Markov process (to be examined in section 3.11). Larger values of $\varphi_C(0)$ (where $M > 1/2$) indicate a smooth process and smaller ones (where $M < 1/2$) a rough process. Also, larger values of $\varphi_C(\infty)$ (where $H > 1/2$) indicate a persistent process and smaller ones (where $H < 1/2$) an antipersistent process.

A useful observation in Figure 2 is that the entire "green square" lies below the equality line, which means that the same scaling exponent is not possible for both local and global behaviour, or else, it is impossible to have a physically realistic simple scaling process. There is one exception, the upper-left corner of the "green square", which corresponds to the so-called "pink noise" or "$1/f$ noise" and will be discussed further in Section 3.3.

On the left of the "green square" in Figure 3.4 another square is formed, which represents processes that are mathematically feasible but physically unrealistic, because they entail infinite variance. In particular, the centre of this square represents the *white noise*, characterized by independence in time, which is discussed in section 3.10. One of the diagonals of this square represents the Hurst-Kolmogorov process, discussed in section 3.12.

Digression 3.C: Misuses of stationarity and ergodicity (2)

Continuing the examples on misuse of the concepts of stationarity and ergodicity in Digression 3.A, we refer here to another example, whose standard formulation could be: "From the time series x_τ, we calculated the power spectrum and found that its slope for low frequencies is steeper than –1, which means that the process is nonstationary." We note that a large number of studies exploring several data sets have reported steep constant slopes of power spectrum, i.e. $\beta < -1$, which are thought to confirm the nonstationarity of the process. The fact is, however, that the entire line of thought is theoretically inconsistent and such reported numerical results are artefacts due to insufficient data or inadequate estimation algorithms. Once we make the power spectrum of a process as a function of frequency, we have tacitly assumed a stationary process. In a nonstationary process, both the autocovariance and the spectral density, i.e. the Fourier transform of the autocovariance, are functions of two variables, one being related to "absolute" time (see e.g. Dechant and Lutz, 2015). Thus, there is no meaning in using a stationary representation (setting the power spectrum as a function of frequency only) and, at the same time, claiming nonstationarity. Furthermore, once we use the power spectrum of a process for inference, as we always do, we should be aware that inference from data is only possible when the process is ergodic. As shown in Koutsoyiannis et al. (2018), in an ergodic process, the asymptotic slope on the left tail of the power spectrum cannot be steeper than –1. Thus, there is no meaning in reporting slopes in empirical power spectra < -1 and at the same time making any claim about the process properties (e.g. of nonstationarity) based on the power spectrum. Actually, such a steep slope, when emerging from processing of data, does not suggest that a process is non-ergodic, it rather signifies inconsistent estimation. Nonetheless, we should be aware, that steep slopes (< -1) are mathematically and physically possible for medium and large frequencies—actually they are quite frequent in geophysical processes.

Consequently, possible reformulations of the above inconsistent statement could be the following:

- We cursorily interpreted a slope steeper than –1 in the power spectrum as evidence of nonstationary, while a simple explanation would be that the frequencies on which our data enable calculation of the power spectrum values are too high.
- We cursorily applied the concept of the power spectrum of a stationary stochastic process, forgetting that the empirical power spectrum of a stationary stochastic process is a (nonstationary) stochastic process per se (see section 4.10). The high variability of the latter (or the inconsistent numerical algorithm we used) resulted in a slope for low frequencies steeper than –1, which is absurd. Such a slope would suggest a non-ergodic process while our calculations were based on the hypothesis of a stationary and ergodic process.
- We cursorily applied the concept of the power spectrum of a stationary stochastic process using a time series which is realization of a nonstationary stochastic process and we found an inconsistent result; therefore, we will repeat the calculations recognizing that the power

spectrum of a nonstationary stochastic process is a function of two variables, frequency and "absolute" time.

3.10 White noise: how natural is it?

We are all familiar with the notion of independent events at discrete time, such as coin, dice and roulette wheel experiments. If such an experiment is made sequentially in time, we can model it as a stochastic process $\underline{v}'_\tau, \tau = 1,2 \ldots$ with mean μ and variance γ_1. For convenience we subtract its mean, defining the process $\underline{v}_\tau := \underline{v}'_\tau - \mu$ for which:

$$\mathrm{E}[\underline{v}_\tau] = 0, \qquad \mathrm{var}[\underline{v}_\tau] = \mathrm{E}[\underline{v}_\tau^2] = \sigma^2, \qquad c_\eta = \mathrm{cov}[\underline{v}_\tau, \underline{v}_{\tau+\eta}] = \begin{cases} \sigma^2, & \eta = 0 \\ 0, & \eta \neq 0 \end{cases} \tag{3.47}$$

It is easy to show that the time-averaged process:

$$\underline{v}_\tau^{(\kappa)} := \frac{1}{\kappa} \sum_{i=(\tau-1)\kappa}^{\tau\kappa} \underline{v}_i \tag{3.48}$$

has the following properties:

$$\mathrm{E}\left[\underline{v}_\tau^{(\kappa)}\right] = 0, \qquad \gamma_1^{(\kappa)} = \mathrm{var}\left[\underline{v}_\tau^{(\kappa)}\right] = \frac{\sigma^2}{\kappa},$$
$$c_\eta^{(\kappa)} = \mathrm{cov}\left[\underline{v}_\tau^{(\kappa)}, \underline{v}_{\tau+\eta}^{(\kappa)}\right] = \begin{cases} \dfrac{\sigma^2}{\kappa}, & \eta = 0 \\ 0, & \eta \neq 0 \end{cases} \tag{3.49}$$

Is it legitimate to say that the discrete-time process $\underline{v}_\tau$ originates from a continuous time process $\underline{v}(t)$? And if yes, what are the properties of the latter? The mathematical reply to the former question is positive. To materialize the continuous-time variant it suffices to generalize the climacogram in (3.49) changing the time scale from an integer κ to a real number $k := \kappa D$:

$$\gamma(k) = \mathrm{var}[\underline{v}(t)] = \frac{\sigma^2 D}{k} \tag{3.50}$$

It is easily seen that if $k \to 0$, the process variance tends to infinity. Thus, to express the properties of the continuous-time process, we need to involve the Dirac delta function $\delta(t)$, whose properties are:

$$\delta(t) = \begin{cases} \infty, & t = 0 \\ 0, & t \neq 0 \end{cases}, \qquad \int_a^b \delta(t)\mathrm{d}t = 1 \tag{3.51}$$

where $[a, b]$ is any interval that contains the 0. To connect the discrete-time process $\underline{v}_\tau$ to the continuous-time process $\underline{v}(t)$, we assume that the former is the time-average of the latter on the time interval of length D, as in equation (3.13). If we define $\underline{v}(t)$ as a stationary stochastic process which has the following properties:

$$\mathrm{E}[\underline{v}(t)] = 0, \qquad \mathrm{cov}[\underline{v}(t), \underline{v}(t')] = \mathrm{E}[\underline{v}(t)\underline{v}(t')] = \sigma^2 D\, \delta(t - t') \tag{3.52}$$

then it results in a discrete-time process with the properties of equation (3.49). Indeed, the variance of $\underline{x}_\tau$ will be:

$$\begin{aligned}\mathrm{var}[\underline{x}_\tau] = \mathrm{var}[\underline{x}_1] &= \mathrm{E}\left[\frac{1}{D}\int_0^D \underline{x}(t)\mathrm{d}t\right] = \frac{1}{D^2}\mathrm{E}\left[\int_0^D \underline{x}(t)\mathrm{d}u \int_0^D \underline{x}(s)\mathrm{d}s\right] \\ &= \frac{1}{D^2}\int_0^D\int_0^D \mathrm{E}[\underline{x}(t)\underline{x}(s)]\,\mathrm{d}t\mathrm{d}s = \frac{1}{D^2}\int_0^D\int_0^D \sigma^2 D\delta(t-s)\,\mathrm{d}t\mathrm{d}s \\ &= \frac{\sigma^2 D}{D^2}\int_0^D 1\mathrm{d}s = \sigma^2\end{aligned} \tag{3.53}$$

The power spectrum of process $\underline{v}(t)$ is found (from equation (3.20)) to be constant:

$$s(w) = \sigma^2 D \tag{3.54}$$

Because all frequencies w are present in the power spectrum with equal density, the process $\underline{v}(t)$ is called *white noise*. This name has been given by analogy to white light, which is a mixture of all visible frequencies. We note though that this is a misnomer as the power spectrum of the white light is far different from flat.

While mathematically the white noise is a well-founded concept and useful for many theoretical analyses, it may not be physically realistic for several reasons, such as the following:

- Its variance is infinite: $\mathrm{var}[\underline{v}(t)] = \mathrm{E}\left[\left(\underline{v}(t)\right)^2\right] = \sigma^2 D\ \delta(0) = \infty$. If this represented a natural process, this process would have infinity energy.
- Its autocorrelation for lags however small is zero. In a natural process, the autocorrelation should be close to 1 for lags close to zero.
- Its spectral density is finite as frequency tends to infinity.

These problems are remedied by applying some kind of filtering to the process $\underline{v}(t)$. An example is to set an upper limit w_c to the frequency, beyond which the spectral density becomes zero (a so-called *low-pass* or *high-cut* filter). The second-order characteristics of the thus obtained stochastic process $\underline{\tilde{v}}(t)$ are:

$$\tilde{\gamma}_0 = \sigma^2 D\ w_\mathrm{c}, \qquad \tilde{c}(h) = \sigma^2 D\ w_\mathrm{c}\ \mathrm{sinc}(2\pi w_\mathrm{c} h), \qquad \tilde{s}(w) = \begin{cases}\sigma^2 D\,, & w \le w_\mathrm{c} \\ 0, & w > w_\mathrm{c}\end{cases} \tag{3.55}$$

It may be readily seen that the above three inconsistencies have been remedied. On the other hand, the process $\underline{\tilde{v}}(t)$ does not yield precisely the process $\underline{v}_\tau$ in discrete time. However, if we choose $w_\mathrm{c} \gg 1/D$, we can obtain good approximations.

Digression 3.D: Random walk, Wiener process and Brownian motion

Assuming that the discrete-time white noise process $\underline{v}_\tau$ is two-valued, e.g. taking on the values +1 and –1 with equal probabilities $p = 0.5$ (so that $\mathrm{E}[\underline{v}_\tau] = 0$), the cumulative process $\underline{V}_\tau := \sum_{i=1}^{\tau} \underline{v}_i$, which can take on values in the interval $[-\tau, \tau]$, is called a random walk. This is a nonstationary

process with its variance being proportional (actually equal in this simple case) to the time τ that has passed from the beginning of the walk, i.e. $\mathrm{var}[\underline{V}_\tau] = \tau$. Its mean is zero at all times.

If both the time t and the state $\underline{v}(t)$ of the white noise are continuous, then the resulting cumulative process $\underline{V}(t) \coloneqq \int_0^t \underline{v}(s)\mathrm{d}s$ is called the *Wiener process*. This is again a nonstationary process with mean zero and variance proportional to the time *t*, i.e. $\mathrm{var}[\underline{V}(t)] = \sigma^2 t$, where σ^2 has been defined above; note that the quantity $\sigma^2/2$ is known as the *diffusion constant*.

The Wiener process is used to model diffusion phenomena and the Brownian motion under free conditions, i.e., when there are no bounds in the motion, nor a restoring force (e.g. gravity in atmospheric motion). However, in real world systems the motion is not free (these conditions do not exist) and the Brownian motion is bound. In such systems the resulting process is not Wiener but a stationary process. More information on these processes can be found in Papoulis (1991).

3.11 The linear Markov process

We will now discuss a more interesting case of filtering of the white noise by means of a stochastic version of a linear differential equation. To establish such an equation, we use a simple hydrological system, a linear reservoir with inflow $v(t)$ and outfow $x(t)$. The reservoir state is characterised by its storage $S(t)$ and the change in outflow (reservoir spill) is assumed (as an approximation) to be proportional to the change in storage, $\mathrm{d}x = \mathrm{d}S/\alpha$, where $\alpha > 0$ is a constant with units of time. The continuity equation is $dS/dt = v - x$ and if we make the substitution $\mathrm{d}S = \alpha \mathrm{d}x$ we find that the system dynamics is the first-order linear differential equation (for a nonlinear version see Digression 9.A):

$$\alpha \frac{\mathrm{d}x(t)}{dt} + x(t) = v(t) \tag{3.56}$$

Now, let us assume that the inflow is a stochastic process and specifically a white noise process. For convenience we subtract its mean so that $\underline{v}(t)$ has the characteristics given in equation (3.52). The output $\underline{x}(t)$ will be a stochastic process as well. Thus, we can write the stochastic version of equation (3.56) as:

$$\alpha \frac{\mathrm{d}\underline{x}(t)}{dt} + \underline{x}(t) = \underline{v}(t) \tag{3.57}$$

As simple as may it seem, the transition from the deterministic version in equation (3.52) to the stochastic version in equation (3.57) involves mathematical troubles. In fact, the process $\underline{x}(t)$ is hardly differentiable and the derivative $\mathrm{d}\underline{x}(t)/dt$ does not generally exist. Thus, stochastic differential equations require their own rules of calculus. Here we use the following simple rule: We solve the differential equation as if it were deterministic with well-defined derivative. Naturally, the mathematical expression of the solution will not contain derivatives. In that expression we replace the deterministic functions with stochastic processes. Thus, the differentiability problem is bypassed.

In this manner, the linear differential equation (3.57) is easily solved to give:

$$\underline{x}(t) = \underline{x}(0)\mathrm{e}^{-t/\alpha} + \frac{\mathrm{e}^{-t/\alpha}}{\alpha} \int_0^t \underline{v}(u)\mathrm{e}^{u/\alpha}\mathrm{d}u \tag{3.58}$$

We observe in equation (3.58) that:

1. The two additive terms on the right-hand side are independent as the outflow of the present, $\underline{x}(0)$, cannot depend on the future inflows $\underline{v}(u), 0 < u \leq t$.
2. The outflow does depend on the outflow of the present, $\underline{x}(0)$, but not on other $\underline{x}(t)$ of the past ($t < 0$).

A stochastic process that has the latter property is called a Markov process. More generally, a Markov process is one in which the future does not depend on the past once the present is known; symbolically:

$$P\{\underline{x}(t)|\underline{x}(s) = x(s), s \leq 0 < t\} = P\{\underline{x}(t)|\underline{x}(0) = x(0)\} \tag{3.59}$$

The particular Markov process $\underline{x}(t)$ of equation (3.58) can be called the linear Markov process and it is also known as *Ornstein–Uhlenbeck* process, while the stochastic differential equation (3.57) is known as the *Langevin equation* (Papoulis, 1991). The mean of the process is:

$$\mathrm{E}[\underline{x}(t)] = \mathrm{E}[\underline{x}(0)]\mathrm{e}^{-t/\alpha} \tag{3.60}$$

Subtracting the latter equation from (3.58), squaring and taking expected values we get:

$$\begin{aligned} \mathrm{var}[\underline{x}(t)] &= \mathrm{var}[\underline{x}(0)]\,\mathrm{e}^{-2t/\alpha} + \frac{\sigma^2 D}{\alpha^2}\mathrm{e}^{-2t/\alpha}\int_0^t \mathrm{e}^{2u/\alpha}\mathrm{d}u \\ &= \frac{\sigma^2 D}{2\alpha} + \left(\mathrm{var}[\underline{x}(0)] - \frac{\sigma^2 D}{2\alpha}\right)\mathrm{e}^{-2t/\alpha} \end{aligned} \tag{3.61}$$

From (3.60) and (3.61) we conclude that $\mathrm{E}[\underline{x}(t)]$ and $\mathrm{var}[\underline{x}(t)]$ tend fast (exponentially) to 0 and $\lambda \coloneqq \sigma^2 D/2\alpha$, respectively, regardless of the values $\mathrm{E}[\underline{x}(0)]$ and $\mathrm{var}[\underline{x}(0)]$. In particular, if $\mathrm{E}[\underline{x}(0)] = 0$ and $\mathrm{var}[\underline{x}(0)] = \gamma_0 = \lambda$, then the process has constant mean (0) and variance (λ) at all times.

It is easily seen that the following equation is a consequence of (3.58):

$$\underline{x}(t+h) = \underline{x}(t)\mathrm{e}^{-h/\alpha} + \frac{\mathrm{e}^{-h/\alpha}}{\alpha}\int_t^{t+h} \underline{v}(u)\mathrm{e}^{u/\alpha}\mathrm{d}u \tag{3.62}$$

Multiplying this latter by $\underline{x}(t)$ and taking expected values we get:

$$c(t,h) = \mathrm{E}[\underline{x}(t+h)\underline{x}(t)] = \mathrm{E}[\underline{x}(t)^2]\mathrm{e}^{-h/\alpha} \tag{3.63}$$

and in the case $\{\mathrm{E}[\underline{x}(0)] = 0, \mathrm{var}[\underline{x}(0)] = \lambda\}$ this becomes:

$$c(h) = \lambda\mathrm{e}^{-h/\alpha} \tag{3.64}$$

In other words, the autocovariance is a function of the lag *h* only and the process is wide-sense stationary. The other second-order characteristics of the process in continuous and discrete time, derived through the generic equations contained in Table 3.3, are summarized in Table 3.5.

The celebrated linear Markov process is nothing more than filtered white noise through a linear differential equation. The filtering eliminates the problems related to the

appearance of infinities discussed in section 3.10 and, thus, it is physically consistent. Furthermore, the simplicity of the equations of its second-order properties makes it attractive and easy to use. On the other hand, its Markovian property, i.e. the independence of the future from the past, once the present is known, may contradict our perception that history does always influence the future developments. We may thus regard it as too simplistic a model of natural reality. Furthermore, the fact that it minimizes entropy production for large times ($t \to \infty$) (Koutsoyiannis, 2011b; see also Digression 3.G) may be another obstacle in accepting it as a good model to represent natural processes.

Table 3.5 Second-order characteristics of the Markov process at continuous and discrete time.

Property	Formula	Eqn. no.
Variance		
Continuous-time process (Instantaneous)	$\gamma_0 = \gamma(0) = c(0) = \lambda$	(3.65)
Averaged process at scale k (Climacogram)	$\gamma(k) = \frac{2\lambda}{k/\alpha}\left(1 - \frac{1-\mathrm{e}^{-k/\alpha}}{k/\alpha}\right)$	(3.66)
Autocovariance function		
Continuous-time, lag h	$c(h) = \lambda \mathrm{e}^{-\lvert h\rvert/\alpha}$	(3.64)
Discrete time, lag $\eta = h/D$	$c_0 = \gamma(D), c_\eta = \frac{\lambda\left(1-\mathrm{e}^{-D/\alpha}\right)^2}{(D/\alpha)^2}\,\mathrm{e}^{-(\eta-1)D/\alpha}, \quad \eta \geq 1$	(3.67)
Power spectrum		
Continuous-time, frequency w	$s(w) = \frac{4\alpha\lambda}{1+(2\pi\alpha w)^2}$	(3.68)
Discrete time, frequency $\omega = wD$	$s_\mathrm{d}(\omega) = 4\alpha\lambda\left(1 - \frac{\sinh(D/\alpha)}{D/\alpha}\frac{1-\cos(2\pi\omega)}{\cosh(D/\alpha)-\cos(2\pi\omega)}\right)$	(3.69)

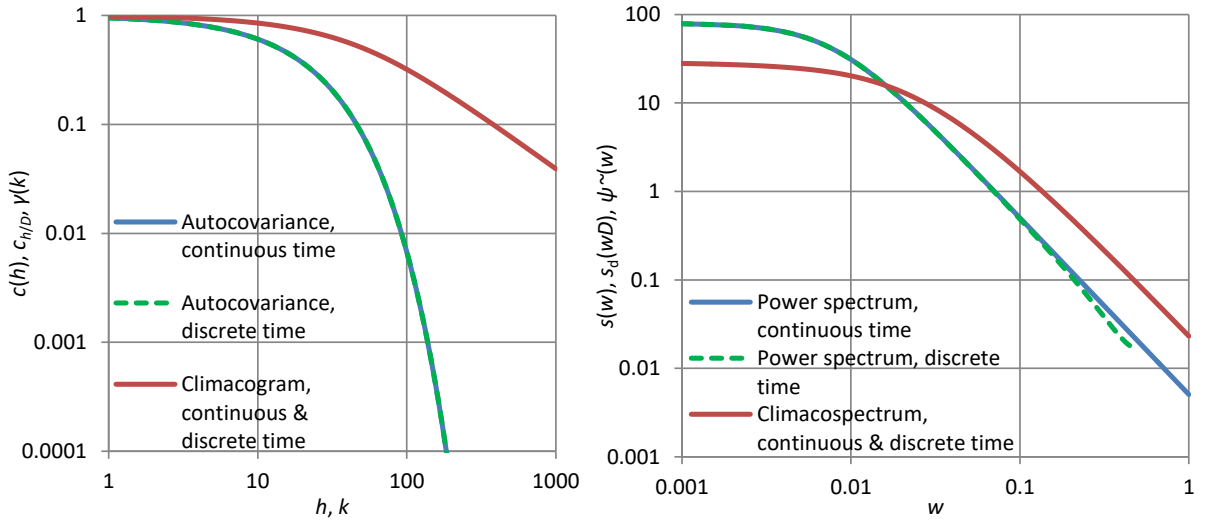

Figure 3.5 Second-order characteristics of a linear Markov process with parameters $\lambda = 1, \alpha = 20$ and discretization time step $D = 1$. The climacogram and climacospectrum are precisely the same for the continuous- and discrete-time representations. The autocovariance and the power spectrum have some differences between the two representations, which are invisible in the former case and visible in the latter.

A discretized Markov process at time step D tends to be uncorrelated in time as D increases. Therefore, at large time scales the Markov model is indistinguishable from white noise: indeed, from equation (3.66) we conclude that for large k (or small α) the variance is inversely proportional to the time scale, as in the white noise. Thus, even though sometimes it is said that the Markov model reflects short-term persistence, it is better not to use the term persistence in this case. Certainly, it entails short-range dependence in time. However, its asymptotic properties (cf. equations (3.45) and (3.46)) are (Koutsoyiannis, 2011b):

$$\begin{aligned} M &= \frac{1}{2},\ \ \varphi_{\mathrm{C}}(0) = \frac{3}{2},\ \ \gamma^{\#}(0) = \ c^{\#}(0) = \ 0,\ \ \psi^{\#}(0) = \ 2,\ \ s^{\#}(\infty) = -2 \\ H &= \frac{1}{2},\ \ \varphi_{\mathrm{C}}(\infty) = \frac{1}{2},\ \ \gamma^{\#}(\infty) = -1,\ \ c^{\#}(\infty) = \ -\infty,\ \ \psi^{\#}(\infty) = \ s^{\#}(0) = 0 \end{aligned} \tag{3.70}$$

Thus, according to the classification of section 3.9, the process is neutral: neither antipersistent nor persistent and neither rough nor smooth.

While the linear differential equation, on which the introduction of the Markov model has been based, has some physical basis, the assumption that the inflow is white noise is physically problematic, as we clarified in section 3.10. This is another reason making the simple Markov model not quite good for natural systems. This problem, even though rarely noticed, is also met in most of the cases of stochastic differential equations, which are deterministic equations perturbed by white noise.

Related to the Markov process in continuous time is the discrete-time process:

$$\underline{x}_\tau = a\underline{x}_{\tau-1} + \underline{v}_\tau + b\underline{v}_{\tau-1} \tag{3.71}$$

commonly known as ARMA(1,1), which stands for autoregressive – moving-average process of orders (1,1). Here $\underline{v}_\tau$ is discrete-time white noise with variance σ_v^2, and a and b are parameters. It can be easily shown (homework) that its second order characteristics are interrelated by:

$$c_0 = ac_1 + (1 + ab + b^2)\sigma_v^2, \qquad c_1 = ac_0 + b\sigma_v^2, \qquad c_\eta = a^{\eta-1}c_1,\ \ \eta \geq 1 \tag{3.72}$$

By comparison with equation (3.67) we see that the ARMA(1,1) process is identical to the discrete-time representation of the Markov process if we choose:

$$a = \mathrm{e}^{-D/\alpha}, \qquad c_0 = \gamma_1 = \frac{2\lambda}{D/\alpha}\left(1 - \frac{1 - \mathrm{e}^{-D/\alpha}}{D/\alpha}\right), \qquad c_1 = \frac{\lambda\left(1 - \mathrm{e}^{-D/\alpha}\right)^2}{(D/\alpha)^2} \tag{3.73}$$

Alternatively, if we know the first three terms of the autocovariance function in discrete time, then, without referring to the continuous time formulation, the parameter a can be found as the ratio

$$a = c_2/c_1 \tag{3.74}$$

The remaining parameters b and σ_v^2 can be found from the first two equations in (3.72) in terms of $c_0 \equiv \gamma_1$ and c_1.

The special case in which:

$$b = 0 \Leftrightarrow c_1/c_0 = a \tag{3.75}$$

is known as the AR(1) process, standing for autoregressive process of order 1. This is the limiting case as $D/\alpha \to 0$. It can also appear in a discrete-time representation of the Markov process for finite time step D, if we use instantaneous quantities, rather than time averages—the so-called *sampled process*, defined in discrete time as:

$$\underline{x}_\tau \coloneqq \underline{x}(\tau D) \tag{3.76}$$

(compare this with (3.13)). The AR(1) process is thus:

$$\underline{x}_\tau = a\underline{x}_{\tau-1} + \underline{v}_\tau \tag{3.77}$$

and its second order characteristics are interrelated by:

$$c_0(1-a^2) = \sigma_v^2, \qquad c_\eta = a^{|\eta|}c_0 \tag{3.78}$$

Additional information about discrete time processes of this type is given in Digression 3.E.

Digression 3.E: The Time Series School and its processes

The AR(1) and ARMA(1,1) processes discussed in section 3.11 are representatives of bigger families of models developed within the *Time Series School*. It is worth mentioning one more process from these families, the AR(2) process, which is:

$$\underline{x}_\tau = a_1\underline{x}_{\tau-1} + a_2\underline{x}_{\tau-2} + \underline{v}_\tau \tag{3.79}$$

It can be easily shown (homework) that its second order characteristics are interrelated by:

$$c_0 = a_1c_1 + a_2c_2 + \sigma_v^2, \qquad c_1 = a_1c_0 + a_2c_1, \qquad c_\eta = a_1c_{\eta-1} + a_2c_{\eta-2}, \ \eta \geq 1 \tag{3.80}$$

Once the covariances c_0, c_1, c_2 are known (estimated from data or derived theoretically) the three parameters a_1, a_2, σ_u^2 can be easily found as the system of equations is linear. These equations are called *Yule – Walker equations* as they were introduced by Yule (1927) and Walker (1931), both British statisticians who, starting from an analysis of sunspot numbers, studied autoregressive processes and in particular their periodogram and autocorrelation properties.

Obviously, higher order AR and ARMA models can be formulated, and actually are in common use, along with additional families such as ARIMA(p,d,q) (standing for autoregressive integrated moving average models) and ARFIMA(p,d,q) (with the additional 'F' standing for fractional). However, we will not refer to them, preferring to base our analyses on the *Stochastic School*, pioneered by A. Kolmogorov, which offers more solid grounds, both for foundation and application, than the Time Series School.

We should note, however, that the latter School and its models are way more popular than the former in many disciplines, including hydrology and climatology. It appears that the Time Series School was initiated by the American economist W.M. Persons. In studying the problem "*When to buy or sell*", Persons (1919) introduced the study of time series, which he called *statistical series*, and asserted that they "*result from the combination of four elements: secular trend, seasonal variation, cyclical fluctuation, and a residual factor.*" He also proposed methods for "*Eliminating secular trends*" and "*Eliminating seasonal variation*". Interestingly, the Ukrainian/Russian/Soviet mathematical statistician and economist Slutsky (1927) demonstrated that what Persons (and other economists) regarded as cyclical component is nothing but a statistical artefact with no essential meaning (see e.g. Kyun and Kim 2006; Barnett, 2006). Subsequently, the notion of a cyclical component was abandoned but the decomposition of a time series to the remaining three components, trends, seasonal variation and residuals is popular even today. Perhaps the first definition of a time series was given by the American statistician Bailey (1929):

A time series is a series of observations taken at different times and recorded with the time at which they were taken.

The biggest progress in the Time Series School was made in Uppsala by the Norwegian-born (with career in Sweden) econometrician and statistician H.O.A. Wold and the New-Zealand-born mathematician and statistician P. Whittle, who in their doctoral theses provided the stochastic foundation of time series analysis. Wold (1938, 1948) proved that a stochastic process (even though he referred to it as a time series) can be decomposed into a *regular process* (i.e., a process linearly equivalent to a white noise process) and a *predictable process* (i.e., a process that can be expressed in terms of its past values). This has been known as *Wold's decomposition*. Whittle (1951, 1952, 1953) laid the mathematical foundation of autoregressive and moving average models in univariate and multivariate setting. Later, in their influential book, Box and Jenkins (1970) named these models with the above acronyms and they became popular with these names and also with the name *Box – Jenkins models* (cf. Stigler's law of eponymy; Stigler, 2002).

Despite the wider influence of the Time Series School over the Stochastic School, there are several problems with the former. First, the term *time series* is ambiguous, sometimes denoting a series of observations as in the original definition of Bailey (1929) (or, equivalently, a realization of a stochastic process), and other times denoting the stochastic process *per se* (as in the aforementioned use by Wold). As we have already emphasized, here the term *time series* is used with the first meaning, a series of numbers, while for a series of stochastic variables we use the term stochastic process. Second, with the exception of the simplest models of these families, such as the AR(1) and ARMA(1,1), time series models are too artificial because, being complicated discrete-time models, they do not necessarily correspond to a continuous time process, while natural processes typically evolve in continuous time. Furthermore, their identification, typically based on the estimation of the autocorrelation function from data, usually neglects estimation bias and uncertainly, which in stochastic processes (as opposed to purely random processes) are often tremendous (Lombardo et al., 2014).

Indeed, from their onset (Whittle, 1952), time series models have been tightly associated with a large number of parameters and they usually become over-parameterized and thus not parsimonious. These parameters are estimated from data, which usually are too few to support a reliable estimation. The decomposition of a time series to components, trends, seasonal variation and residuals, is fundamentally problematic, despite being popular. Remarkably, a meaningful definition of a trend has never been given. Also, it may be hard to conceive how *time per se* could be regarded as an explanatory variable in a complex process and what the logical basis is in expressing the statistics of a process as a deterministic function of time. Accumulation of data series with long time spans (cf. Chapter 1) has shown that, what have been regarded as trends, are mostly parts of long term fluctuations (and in accord to Slutsky's work, they could also be regarded as statistical artefacts). Finally, "deseasonalization"—in Persons's original terminology "*Eliminating seasonal variation*"—is a delusion; we can hardly remove seasonality in the multivariate distribution of a stochastic process; what we typically do is in the marginal distribution—and thus there is no elimination.

3.12 The Hurst-Kolmogorov process

The Hurst-Kolmogorov (HK) process has been already introduced in section 1.3 and its discrete-time version was given in equation (1.6). Its continuous-time version is quite similar:

$$\gamma(k) = \lambda \left(\frac{\alpha}{k}\right)^{2-2H} \tag{3.81}$$

This equation can serve as the definition of the HK process. By setting $H = 1/2$ we recover equation (3.50), which means that the HK process is a generalization of the white noise.

Its other second-order characteristics are given in Table 3.6. Their LLDs are constant for all time lags and scales and all frequencies:

$$\varphi(k) = \varphi_C(k) = H, \quad \gamma^{\#}(k) = c^{\#}(h) = 2H - 2, \quad \psi^{\#}(k) = -s^{\#}(w) = 2H - 1 \qquad (3.82)$$

including their asymptotic values at 0 and ∞. Accordingly, $M = H - 1$.

Table 3.6 Second-order characteristics of the Hurst-Kolmogorov process at continuous and discrete time.

Property	Formula	Eqn. no.
Variance		
Continuous-time process (Instantaneous)	$\gamma_0 = \gamma(0) = c(0) = +\infty$	(3.83)
Averaged process at scale k (Climacogram)	$\gamma(k) = \lambda(\alpha/k)^{2-2H}$	(3.81)
Autocovariance function		
Continuous-time, lag h	$c(h) = \begin{cases} \lambda H(2H-1)\left(\frac{\alpha}{h}\right)^{2-2H}, & H > 1/2 \\ \lambda\,\delta\left(\frac{h}{\alpha}\right), & H = 1/2 \\ \lambda H(2H-1)\left(\frac{\alpha}{h}\right)^{2-2H} + \delta''\left(\frac{h}{\alpha}\right), & H < 1/2 \end{cases}$	(3.84)
Discrete time, lag $\eta = h/D$	$c_\eta = \lambda(\alpha/D)^{2-2H}\left(\frac{\lvert j-1\rvert^{2H} + \lvert j+1\rvert^{2H}}{2} - \lvert j\rvert^{2H}\right)$	(3.85)
Power spectrum[1]		
Continuous-time, frequency w	$s(w) = \frac{2\alpha\lambda\,\Gamma(2H+1)\sin(\pi H)}{(2\pi\alpha w)^{2H-1}}$	(3.86)

[1] The power spectrum of the discrete-time (averaged) process exists (it is finite for $w > 0$) but it does not have a closed expression. However, for small frequencies ($\omega = wD < 0.1$), the continuous-time expression is a very good approximation for the discrete-time process, i.e. $s_d(\omega) \approx s(\omega/D)$.

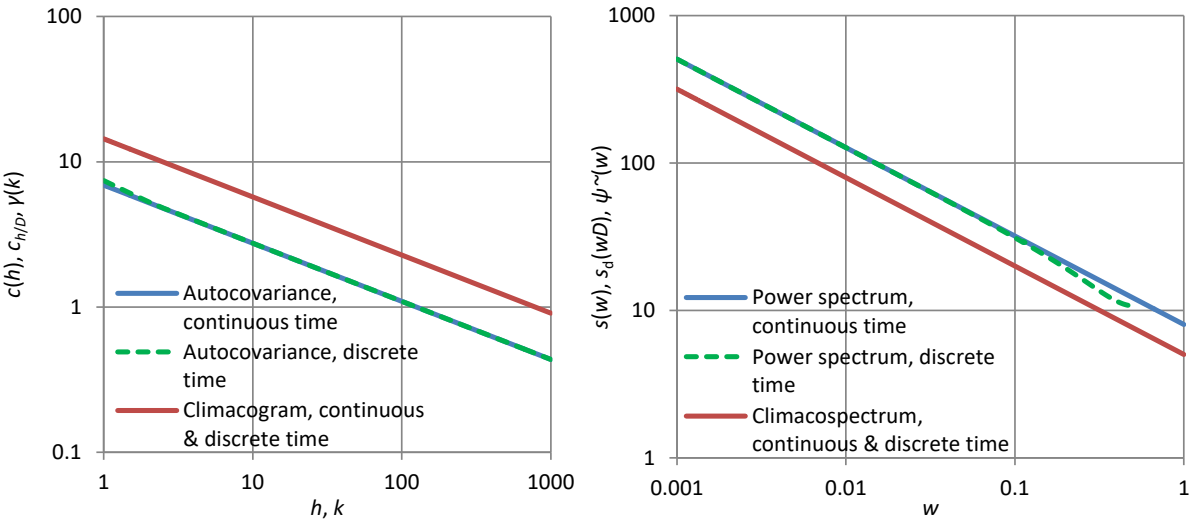

Figure 3.6 Second-order characteristics of a Hurst-Kolmogorov process with parameters $\lambda = 1, \alpha = 20, H = 0.8$ and discretization time step $D = 1$. The climacogram and climacospectrum are precisely the same for the continuous- and discrete-time representations. The autocovariance and the power spectrum have some differences between the two representations, which are visible in both cases.

The process is also known as *fractional Gaussian noise* (FGN) due to Mandelbrot and van Ness (1968), although these authors used a more complicated approach to define it. Here we do not use the term FGN as the adjective *fractional* is not quite informative (there cannot be a non-fractional process; note that the white noise, in which $H = 0.5$ is fractional too), the adjective *Gaussian* is too restrictive (we will implement non-Gaussian HK) and the noun *noise* is too negative and perhaps misleading when we try to describe Nature's processes. As already mentioned, the mathematical process had been earlier proposed by Kolmogorov (1940), while Hurst (1951) pioneered the detection in geophysical time series of the behaviour described by this process; hence the name HK we use for this process.

Because this process has infinite instantaneous variance, the sampled process in discrete time is not meaningful (many characteristics take infinite values). However, the averaged process is well behaving with all of its characteristics (including its variance) finite, which makes it quite useful in applications.

The HK process is almost equally simple and parsimonious with the Markov process; again, it contains only one parameter, H, in addition to those describing its marginal distribution. Notice that the process variance is controlled by the product $\lambda\,\alpha^{2-2H}$, so that λ and α, are not in fact separate parameters. Despite that, we prefer the formulation shown in Table 3.6 with three nominal parameters for dimensional consistency: α and λ are scale parameters with dimensions $[t]$ and $[x^2]$, respectively, while H, the Hurst coefficient, is dimensionless in the interval (0, 1).

For $H = 1/2$ the process reduces to pure white noise. For $1/2 < H < 1$ the process is persistent and for $0 < H < 1/2$ antipersistent. Most of the expressions shown in Table 3.6 hold in all three cases. However, the autocovariance $c(h)$ has different expressions in the three cases, as shown in Table 3.6. Specifically, for $H < 1/2$, the autocovariance $c(h)$ is negative for any lag $h > 0$, tending to $-\infty$ as $h \to 0$. However, at $h = 0, c(0) = +\infty$, because this is the variance of process which cannot be negative; thus, there is a discontinuity at $h = 0$. Consequently, the averaged process has positive variance and all covariances negative. Such a process is not physically realistic because real-world events at near times are always positively correlated, which means that for small h, $c(h)$ should be positive. Also, the infinite variance cannot appear in nature. Thus, the HK process can describe natural phenomena only for $1/2 < H < 1$ and for time scales not too small. Furthermore, values $H > 1$ that sometimes are being reported in the literature are mathematically invalid (Koutsoyiannis, 2014b, 2017; Koutsoyiannis et al. 2018; see also Figure 3.4) and are results of inconsistent algorithms. In terms of entropy production, the process maximizes it for large times ($t \to \infty$) but minimizes it for small times ($t \to 0$).

Digression 3.F: Developments in stochastic modelling in hydrology before and after Hurst

Hurst's (1951) discovery of the natural behaviour named after him was triggered by a real-world problem of engineering hydrology, the design or reservoirs. This gave hydrology a central role in understanding this behaviour and subsequently in the dissemination process to other disciplines. It is a further mark of distinction that the large-scale "export" from hydrology to other fields has

characterized Hurst's research, as hydrology is most often an importer of stochastic methods from other fields (O'Connell et al., 2016).

The understanding that hydrological processes cannot be effectively modelled by deterministic techniques preceded Hurst's research. Techniques that could be classified as applications of the Monte Carlo method had appeared in the hydrological literature much earlier than the "official start" of the Monte Carlo method in 1949 and of Hurst's (1951) paper. Hazen (1914), made a pioneering study in which he introduced the reservoir storage-yield-reliability relationship, a concept that would remain unexploited in the western hydrological literature yet constituting the scientific basis of modern reservoir design (Klemes, 1987). In that study he proposed an empirical simulation technique and formed a synthetic time series by combining historical flow records of different rivers 'spliced' sequentially together. Sudler (1927) extended the work of Hazen by resampling from a sequence of historical river flows using cards, which he shuffled to form new sequences of data. Obviously, this method heavily distorts the time dependence of river flows whose importance was not known at that time.

For it was Hurst (1951) who understood that importance along with the omnipresence in natural processes of a clustering behaviour of similar events in time, a behaviour that is now understood as (long-term) persistence, long-range dependence (LRD) or Hurst-Kolmogorov dynamics. In his attempt to compare natural and random events, Hurst performed physical experiments to generate random numbers. Specifically, he tossed 10 coins (sixpences) simultaneously and repeated this 1025 times (note that 10 binary digits are equivalent to about 3 decimal digits). As he notes, his rate was 100 random numbers per 35 min (while that would be of the order of a microsecond in modern computer environments, even slow ones). He also used another method, shuffling and cutting a pack of 52 cards, in which he improved the rate to 100 random numbers per 20 min.

The behaviour discovered by Hurst is now known to many disciplines, most prominently in information sciences, biological and medical sciences, economics and finance, and geophysical sciences—excepting climate science where it is rather unknown. Even within the hydrological community it took decades before assimilating Hurst's discovery of persistence (O'Connell et al. 2016). Thus, the initial studies implementing primitive variants of stochastic simulation did not reproduce LRD. Barnes (1954), in designing a reservoir in Australia, used a table of random numbers from normal distribution to generate a 1000-year sequence of synthetic annual data. Thomas and Fiering (1962) generated flows correlated in time, but using only the lag-one autocorrelation, obviously neglecting LRD. Beard (1965) and Matalas (1967) generated concurrent flows at several sites. Chow (1969), and Chow and Kareliotis (1970) systematized the use of time series models (in particular—and using their terminology—moving average models, sum of harmonics models and autoregression models) and highlighted their value in the economic planning of water supply and irrigation projects. It is evident from the above pioneering studies, as well as of subsequent myriads of studies, that hydrologists have followed (and today still do) the Time Series School rather than the more rigorous Stochastic School.

3.13 The Filtered Hurst-Kolmogorov process

The HK process should not be regarded as a model of general validity, but one that it is valid for large scales—and we will indeed use it as more physically plausible than processes with exponential decrease of autocovariance (e.g. the Markov process). To this aim, we can appropriately filter HK to make it a physically consistent process for all scales. This is the same with what we did to the white noise to make it physically consistent by removing infinities.

Similar to the white noise process, if we filter an input $\underline{v}(t)$ that is now an HK process, either by a moving average filter or by a linear differential equation system, then it is easy to see that the filtered output is a physically realistic process with finite variance $\gamma(0)$, practically unaffected climacogram $\gamma(k)$ at large scales, with $\gamma^{\#}(\infty) = 2H - 2$ (as in the

original HK process) but highly modified climacogram at small scales, thus having a valid structure with $M = \varphi_C(0) - 1 = (\psi^{\#}(0) - 1)/2 = H$.

However, to enrich the process we can make the parameter M independent of H, thus making it more flexible to model real world data. For the model application it is not necessary to specify the linear filter needed to convert the HK process into a filtered Hurst-Kolmogorov (FHK) process (in some cases this would be too involved). It suffices to specify a convenient expression of the climacogram. Below we provide three such expressions (from Koutsoyiannis, 2017). All expressions contain the dimensionless parameters M and H with the meaning and values discussed in section 3.8.

1. The generalized Cauchy-type (FHK-C) climacogram:

$$\gamma(k) = \lambda(1 + (k/\alpha)^{2M})^{\frac{H-1}{M}} \tag{3.87}$$

2. The generalized Dagum-type (FHK-D) climacogram:

$$\gamma(k) = \lambda\left(1 - \left(1 + (k/\alpha)^{2(H-1)}\right)^{\frac{M}{H-1}}\right) \tag{3.88}$$

3. The composite Cauchy-Dagum-type (FHK-CD) climacogram, derived by summing an FHK-C with $M = 1$ and an FHK-D with $H = 0$:

$$\gamma(k) = \lambda_1(1 + (k/\alpha_1)^2)^{H-1} + \lambda_2(1 - (1 + (k/\alpha_2)^{-2})^{-M}) \tag{3.89}$$

4. A second form of FHK-CD (FHK-CD2), derived by summing an FHK-C with $M = 1/2$ and an FHK-D with $H = 1/2$:

$$\gamma(k) = \lambda_1(1 + k/\alpha_1)^{2H-2} + \lambda_2(1 - (1 + \alpha_2/k)^{-2M}) \tag{3.90}$$

FHK-CD in ether of the variants (3.89) and (3.90), is most convenient, as the first additive term determines merely the persistence of the process and the second one the smoothness of the process. In addition, it is more flexible and richer than its constituents, as it contains two couples of scale parameters; however, if parsimony is sought, then it can take the same number of parameters as each of the constituents by setting $\alpha_1 = \alpha_2 = \alpha$ and $\lambda_1 = \lambda_2 = \lambda$ (note that, for dimensional consistency, λ and α are minimal parameter requirements).

In the special case $M = 1 - H$ both FHK-C and FHK-D result in precisely the same expression:

$$\gamma(k) = \frac{\lambda}{1 + (k/\alpha)^{2(1-H)}} \tag{3.91}$$

For large k/a the FHK process tends to the HK one. This is illustrated in Figure 3.7, where in addition the linear Markov model (for the same value of the lag-one autocovariance) is plotted for comparison. We notice that, as time tends to zero, the Markov and the FHK models have the same entropy production while the HK model is associated with minimal entropy production. For intermediate times the Markov model gives higher entropy production than the other two models, but this is done at the "expense" of giving too low entropy production at large time scales, at which both the HK and the FHK give precisely the same high entropy production.

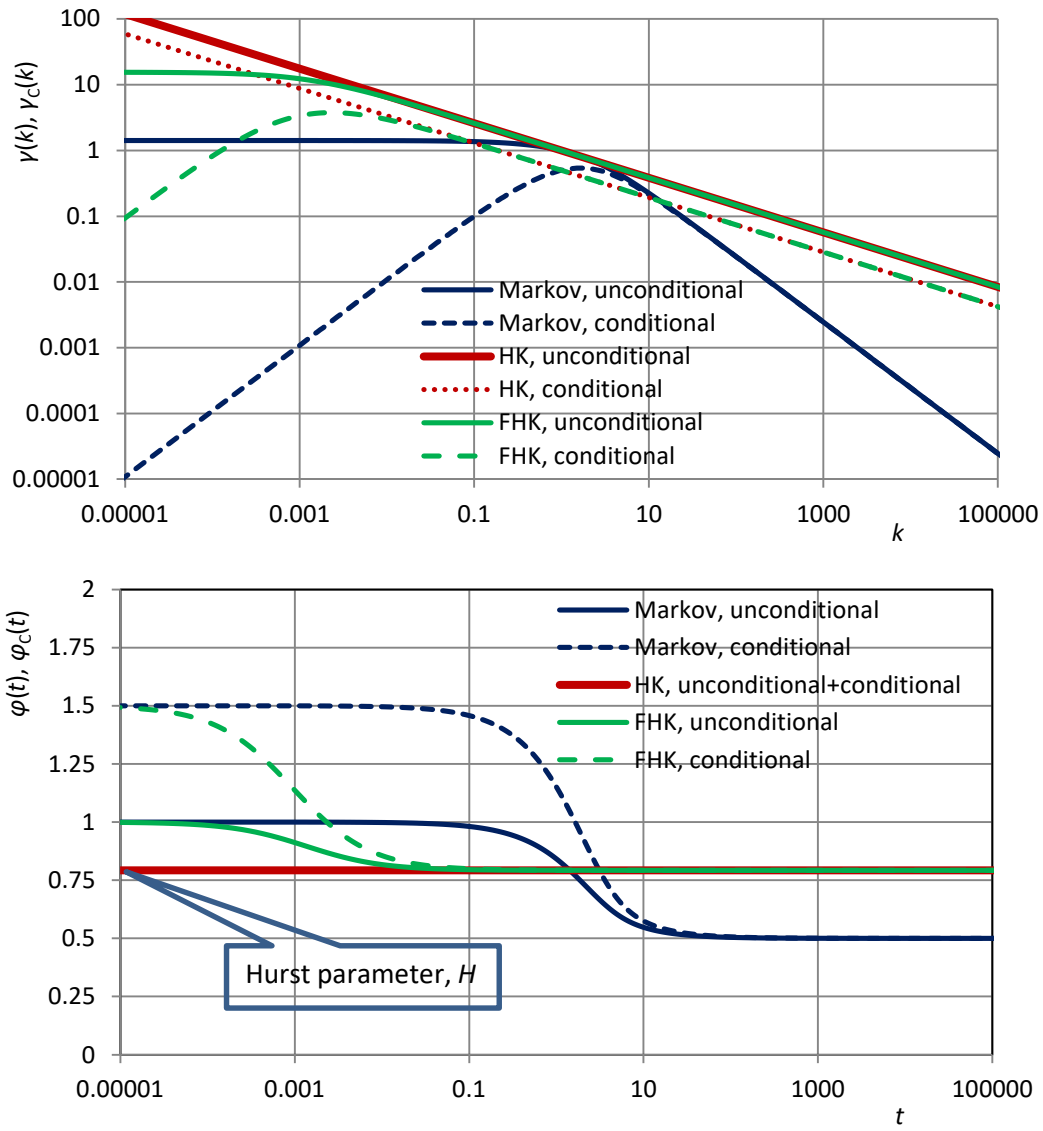

Figure 3.7 (upper) Climacograms and **(lower)** EPLT ($\varphi(t)$) and CEPLT ($\varphi_C(t)$) of the three indicated example processes for neutral smoothness (M = 0.5). At time scale D = 1 all three processes have the same variance $\gamma(1) = 1$ and the same autocovariance for lag 1, $c_1^{(1)} = 0.5$. Their parameters are: for the linear Markov process a = 0.8686, λ = 1.4176; for the HK process a = 0.0013539, λ = 15.5032, H = 0.7925; for the FHK process a = 0.0013539, λ = 15.5093, κ = 0.5, H = 0.7925 (note that for the HK process the parameter set $a = \lambda = 1$ is equivalent to the above, but the former set was preferred in order to be comparable to the FHK).

Digression 3.G: Entropy production and time series patterns

The different patterns in time series generated by different M and H (specifically for the Cauchy-type climacogram) are illustrated in the plots of Figure 3.8, also in comparison with two other models, the white noise (panel (a)) and the linear Markov model (panel (b)). These two serve as good benchmark models for comparisons: the former is free of patterns as it reflects pure randomness, and the latter is fully neutral (neither rough nor smooth as $\varphi_C(0) = 3/2$, and neither antipersistent nor persistent as $\varphi_C(\infty) = 1/2$).

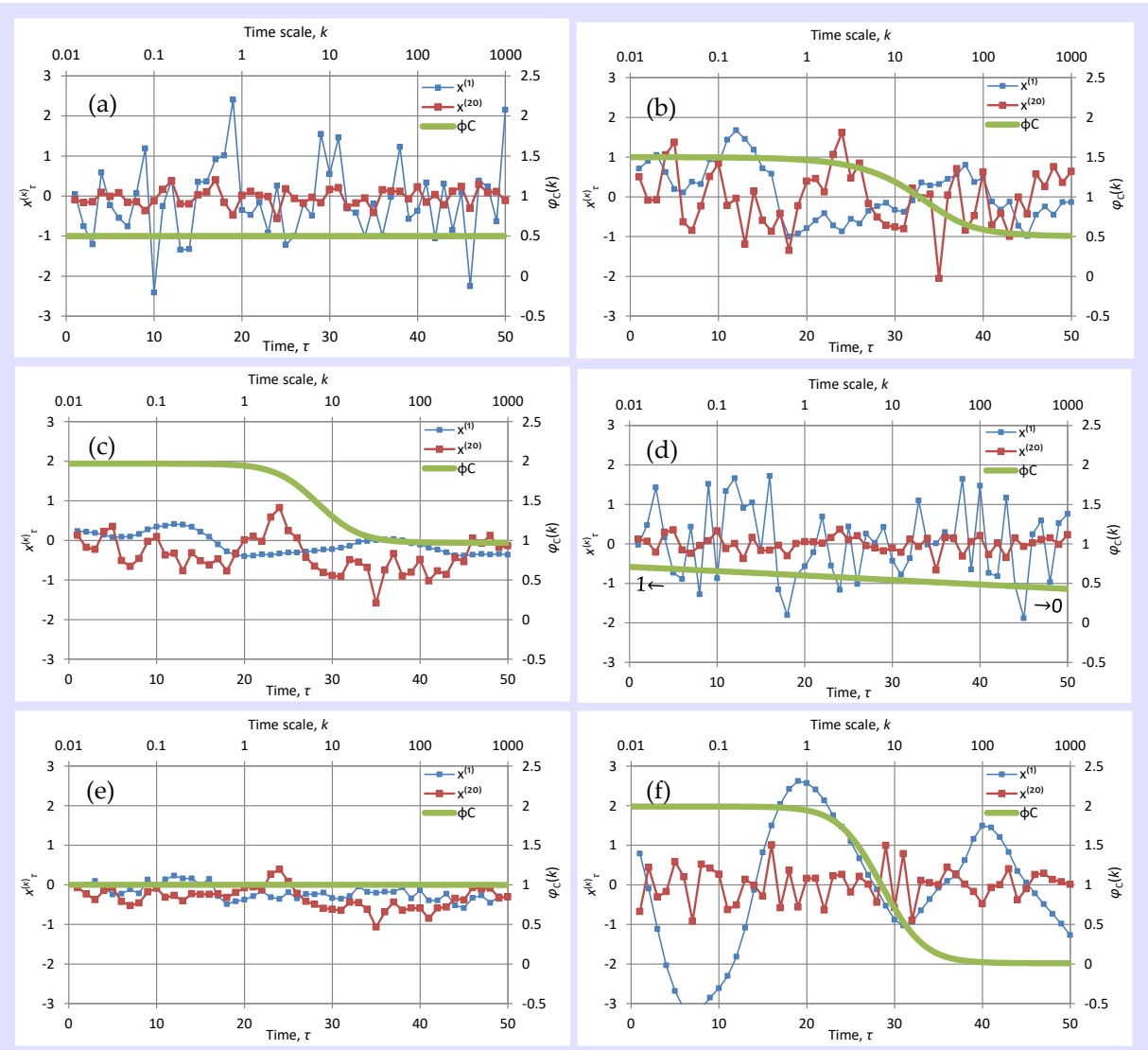

Figure 3.8 The first fifty terms at time scales $k = 1$ and 20 of time series produced by various models, along with "stamps" of the models (thick lines plotted with respect to the right vertical axes) represented by the CEPLT, $\varphi_C(k)$. The different models are **(a)** white noise; **(b)** Markov; **(c)** FHK, with CEPLT close to the absolute maximum ($H = M = 0.97$); **(d)** FHK, with CEPLT close to the absolute minimum ($H = M = 0.05$); **(e)** FHK, with CEPLT close to the absolute maximum for large scales ($H = 0.99$) and close to the absolute minimum for small scales ($M = 0.01$); **(f)** FHK with CEPLT close to the absolute minimum for large scales ($H = 0.01$) and to the absolute maximum ($M = 0.99$) for small scales.

The time series plotted in Figure 3.8 were generated by the symmetric moving average (SMA) scheme which will be described in Chapter 7, with 1024 coefficients (weights) a. In all cases the discretization time scale is $D = 1$, the characteristic time scale $\alpha = 10$, and the characteristic variance scale λ is chosen so that for time scale D, $\gamma(D) = 1$. The mean is 0 in all cases and the marginal distribution is normal. The FHK is implemented using the Cauchy-type climacogram. Each of the panels shows the first fifty terms of time series produced by each of the model implementations at time scales $k = 1$ and 20. In addition, each panel contains a "stamp" of the specific model represented by the plot of CEPLT, $\varphi_C(k)$. In this way the time series patterns can be connected to the entropy production of the generating mechanism.

In panel (c) the CEPLT is close to the absolute maximum both for small and large scales ($H = M = 0.97$ so as to obtain $\varphi_C(0) = 1.97 \approx 2$ and $\varphi_C(\infty) = 0.97 \approx 1$); notable is the very smooth shape at scale 1 and the large departures from the mean (which is 0) at scale 20. On the contrary, in panel (d) the CEPLT is close to the absolute minimum for all scales ($H = M = 0.05$, so as to obtain $\varphi_C(0) = 1.05 \approx 1$ and $\varphi_C(\infty) = 0.05 \approx 0$—for better visualization it was preferred not to use values of H

and $M < 0.05$). Furthermore, in panel (e) the CEPLT is close to the absolute maximum for large scales ($H = \varphi_C(\infty) = 0.99 \approx 1$) and close to the absolute minimum for small scales ($M = 0.01$ resulting in $\varphi_C(0) = 1.01 \approx 1$). Finally, in panel (f) the conditions are opposite to those in (e) i.e., the CEPLT is equal to the absolute minimum for large scales ($H = \varphi_C(\infty) = 0.01 \approx 0$) and to the absolute maximum for small scales ($M = 0.99$ resulting in $\varphi_C(0) = 1.99 \approx 2$).

The particular case of panel (e) is close to what is usually called "pink noise" or "$1/f$ noise", as the power spectrum has almost constant slope −1 for the entire frequency domain (which is the same in the climacospectrum). This means that using the FHK model we can theoretically represent and practically produce even "pink noise" in a consistent stationary setting without linking it to a nonstationary process (Keshner, 1982; Wornell, 1993), which involves several theoretical inconsistencies. Indeed, the small change of slope of from 0.99 to 1.01 is not actually visible, especially considering the very rough shape of the empirical periodogram, which certainly cannot support differentiation between 0.99 and 1. The FHK model can be used also in other ways to produce "pink noise", that is, by selecting a very large (small) parameter α so as to expel from our field of vision the asymptotic behaviour on large (small) scales. And we can imagine that in several cases of empirical explorations using observations of natural processes, the observation resolution and length, compared to characteristic scale(s) of the process, are such as to hide the asymptotic behaviour of the process. We can use this as a trick to obtain virtually constant power spectrum slopes much steeper than −1. Specifically, we can use a large α that does not allow viewing the asymptotic behaviour at low frequencies or large scales and the slope (see example in Koutsoyiannis, 2017). But this should not mislead us to interpret the steep slopes as indicators of nonstationarity.

3.14 Dependence and behaviour of extremes

When we study extremes, we are typically satisfied by specifying the marginal distribution. As analysed in Chapter 2, this is generally sufficient for design purposes, where the design is based upon the concept of return period. In this respect the dependence structure of the process of interest may not affect the design procedure per se. However, the dependence in a stochastic process modifies substantially the temporal distribution of extremes. In a process with dependence there are patterns, and specifically periods with clustered extremes and periods with absence or infrequent occurrence of extremes. We should thus adapt our perception of the behaviour of extremes to become consistent with this reality; without such adaptation our perception is typically guided by the "roulette-wheel" paradigm, in which there are no patterns.

There is an additional, more severe, consequence of the presence of dependence. Hydroclimatic studies necessarily rely on data to make inference. Data records are typically insufficient and actually become even more so in the presence of extremes. The latter problem affects also the specification of the marginal distribution. This is illustrated by a simulation experiment in Digression 3.H. Quantification of the consequences will be given in Chapter 4 and a way to take into account the dependence in specifying the marginal distribution from data will be discussed in Chapter 6.

Digression 3.H: Relationship of persistence and distribution tail

To illustrate whether or not (and how) the persistence (or long-range dependence or just change) affects the estimation of the marginal distribution of a discrete-time stationary process $\underline{x}_\tau$ we perform a simulation experiment. We assume that the marginal distribution of $\underline{x}_\tau$ is exponential: $f_x(x|\lambda) = \lambda\, e^{-\lambda x}$. Further, we make two alternative assumptions:

(a) that the parameter λ is constant, $\lambda = 5$, and
(b) that λ is slowly varying with mean $\mu_\lambda = 5$ and standard gamma distribution, $f_\lambda(\lambda) = \lambda^{\kappa-1}e^{-\lambda}/\Gamma(\kappa)$ with $\kappa = \mu_\lambda = 5$.

To simulate a slowly varying λ we initially generate a time series of a stochastic process $\underline{\lambda}'$ with same distribution as $\underline{\lambda}$ from the HK process with a high $H = 0.95$. Then we form a time series of $\underline{\lambda}$ with the rule $\lambda_i = \lambda_i'$ with probability 1/100, otherwise $\lambda_i = \lambda_{i-1}$. The latter rule assures that each value λ_i lasts for 100 time units on the average. The HK process used for λ_i' assures that there is change on all scales, not just at scale 100. Koutsoyiannis (2004a) has shown that the unconditional distribution of x in this case is Pareto rather than exponential, i.e. $f_x(x) = \kappa\ (1 + x)^{\kappa-1}$.

In either of the two alternatives, once λ is known at time step τ, we generate x_τ from the exponential distribution independently of previous and next x_τ. In alternative (a), the resulting process will be white noise. However, in alternative (b), the change of the parameter induces dependence, while the process $\underline{x_\tau}$ remains stationary (because the change is stochastic, resisting a deterministic description).

Figure 3.9 (upper row) depicts two time series x_τ, each with length 10 000, generated with alternative (a) (left panel) and (b) (right panel). Moving averages for a time scale of 500 are also plotted in the two panels indicate the absence of patterns (pure randomness, white noise) in alternative (a) and the long-range dependence (not nonstationarity) in alternative (b).

Now let us assume that this time series represents a hypothetical hydroclimatic process on annual scale. Let us further assume that a researcher has a record of 100 observations or fewer. Most probably all these refer to the same value of the parameter λ_i. Consequently, the researcher would diagnose that:

- the process behaves like white noise—and indeed, the slope of the climacogram (Figure 3.9, lower right) for scales < 10 (one tenth of the sample size) is –1;
- the marginal distribution is exponential—because indeed it is exponential conditionally on a single value of λ.

The two distributions for constant and varying λ (cases (a) and (b)) are shown in the bottom-left panel of Figure 3.9, along with the distribution of λ in case (b), as empirically derived from the simulations. The adoption of the former underestimates the design quantities for large return periods. Furthermore, the bottom-right panel shows the dramatic differences in climacograms of the two cases. The climacogram in case (b) starts with a slope –1 for scales < 10, but for large scales this becomes –0.33, suggesting $H = 0.84$. The varying slope is consistent with the findings of Markonis and Koutsoyiannis (2016) for the rainfall process. Overall, this simulation experiment shows two things.

- Long series are needed to diagnose natural behaviours and in particular the multi-scale change in natural processes.
- The mechanisms producing change may also lead to thickening of the distribution tail and thus enhancing the occurrence probability or the intensity of extremes.

The effect of (a) is particularly important when we study maxima, neglecting the small values (below a high threshold), a practice that tends to hide the existence of long-range dependence even in long records (see Iliopoulou and Koutsoyiannis, 2019).

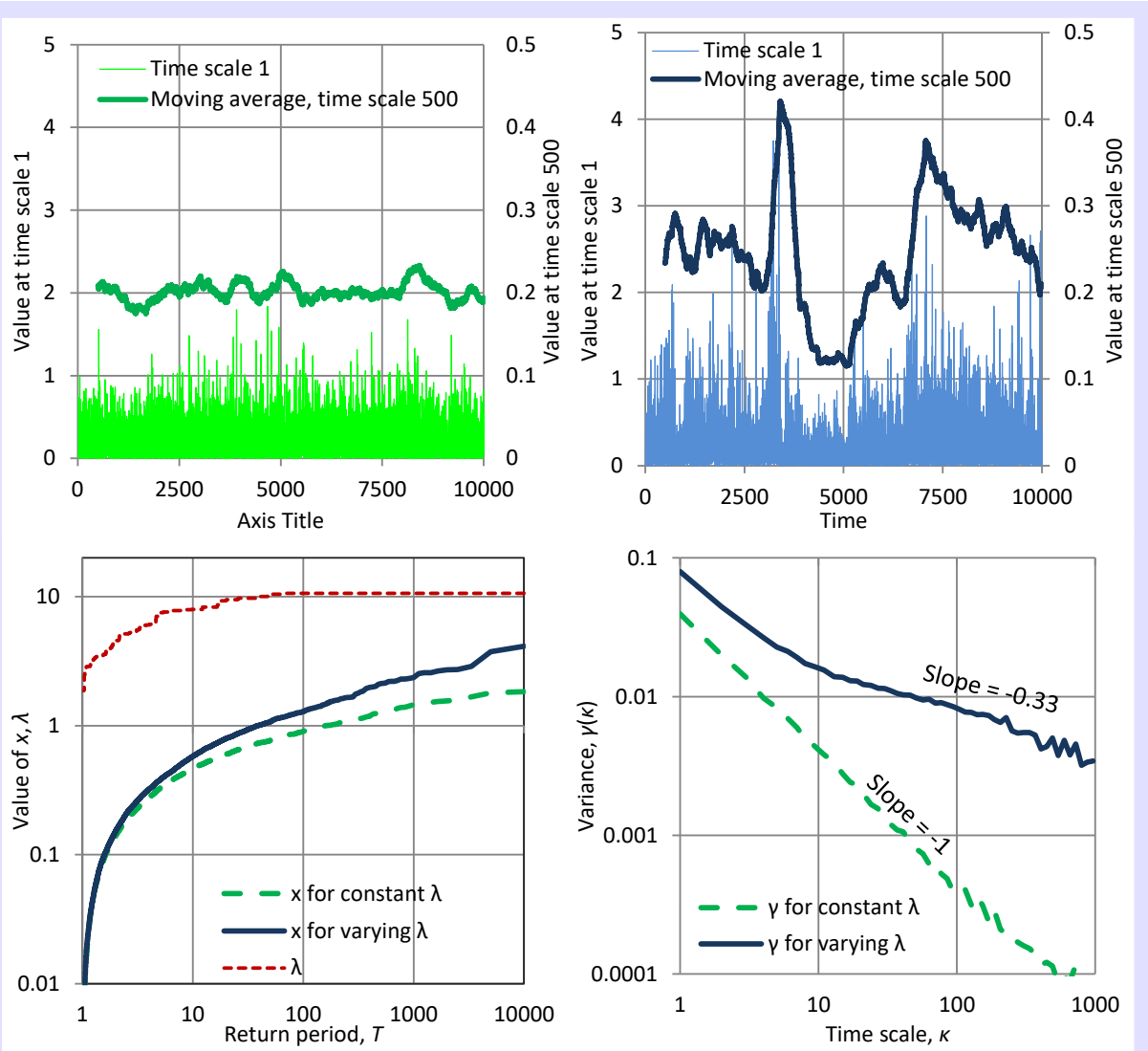

Figure 3.9 Graphs for the hypothetical example studied in Digression 3.H (see text for explanation).

Chapter 4. Fundamental concepts of statistics and their adaptation to stochastic processes

4.1 Introductory comments

The first aim of this chapter is to serve as a synopsis (rather than a systematic and complete presentation) of fundamental statistical concepts. It is well known that the aim statistics per se is to provide a methodology for drawing conclusions based on observations. The conclusions are only inferences based on induction, not deductive mathematical proofs (see Digression 4.A); however, if the associated probabilities approach 1, they become almost certainties.

The classical statistical theory is entirely based on the assumption that observations are from a *sample*, a concept (formally defined in section 4.2) whose very definition relies on independence of observations. However, when we deal with hydroclimatic processes there cannot be independence. Instead of samples we have time series and there is dependence in time. Even when we are interested on the spatial behaviour of processes, again we have to deal with dependence in space. Hence, the second aim of this chapter is to adapt and extend the classical statistical concepts and methodologies to make them applicable to a universe in which there is dependence.

Two major tasks in statistics are *estimation* and *hypothesis testing*. Statistical estimation can be distinguished in *parameter estimation* and *prediction* and can be performed either on a *point basis* (resulting in a single value, typically the expectation), or on an *interval basis* (resulting in an interval in which the quantity sought lies, associated with a certain probability or confidence). The results of the estimation procedures are called *estimates*. Uses of statistical estimation in hydroclimatic applications include the estimation of parameters of marginal probability distributions or of the stochastic model describing the dependence in time, and of distributions quantiles. All this concepts are briefly discussed both in a theoretical level, to clarify the concepts and avoid misuses, and a more practical level to illustrate the application of the concepts. Statistical hypothesis testing is also an important tool that constitutes the basis of the decision theory. In hydroclimatic studies, it is useful not only in decision making, but also in exploratory tasks, such as in detecting relationships among different processes. Hypothesis testing can be performed by the classical framework known as *statistical significance* (related to a *null hypothesis*) or within a *Bayesian framework*. These topics are not covered in this text. On the other hand, we put emphasis on the concept of *order statistics* (section 4.12), which is much more important when dealing with extremes.

Digression 4.A: Deduction and induction

The theory of probability has provided solid scientific grounds for philosophical concepts such as indeterminism and causality. In typical scientific and technological applications, probability has provided the tools to quantify uncertainty, rationalize decisions under uncertainty, and make predictions of future events under uncertainty, in lieu of unsuccessful deterministic predictions (see Koutsoyiannis, 2010).

Quite importantly, probability has also provided the basis for extending the typical mathematical logic, offering the mathematical foundation of induction. Thus, probability made it

possible to incorporate into mathematics the entire Aristotelian logic, which in addition to *deductive reasoning* or *deduction* (the Aristotelian *apodeixis*) also includes *induction* (the Aristotelian *epagoge*).

In classical mathematical logic, determinism can be paralleled to the premise that all truth can be revealed by deductive reasoning. This type of reasoning consists of repeated application of strong syllogisms such as:

If A is true, then B is true;	If A is true, then B is true;
A is true;	B is false;
Therefore, B is true.	Therefore, A is false.

Deduction uses a set of axioms to prove propositions known as theorems, which, given the axioms, are irrefutable, absolutely true statements (once we accepted the axioms as valid). It is also irrefutable that deduction is the preferred route to truth; the question is, however, whether or not it has any limits.

David Hilbert's famous aphorism (later inscribed in his tombstone at Göttingen) "*Wir müssen wissen, wir werden wissen*" ("We must know, we will know"), expressed his belief that there were no limits. According to this belief, more formally known as *completeness*, any mathematical statement could be proved or disproved by deduction from axioms. However, developments in mathematical logic, and particularly Gödel's *incompleteness theorem*, challenged the almightiness of deduction suggesting the usefulness and necessity of induction.

Induction uses weaker inference rules of the type:

If A is true, then B is true;	If A is true, then B is true;
B is true;	A is false;
Therefore, A becomes more plausible.	Therefore, B becomes less plausible.

It does not offer a proof that a proposition is true or false and may lead to errors. However, it is very useful in decision making, when deduction is not possible, which happens quite frequently in the real world and in everyday life (see Jaynes, 2003).

The important achievement of probability is that it quantifies (expresses in the form of a number between 0 and 1) the degree of plausibility of a certain proposition or statement. The formal probability framework uses both deduction, for proving theorems, and induction, for inference with incomplete information or data. For the latter we use the branch of stochastics called statistics.

4.2 Samples and time series

Loosely speaking, statistics draws conclusions for a *population*, based on a *sample*. Although the content of *population* is not strictly defined in the statistical literature, the term describes any collection of objects whose measurable attributes are of interest. The population can refer to the real world and be finite (e.g. the inhabitants of Europe, the mean annual flows of year 2000 for all river basins on Earth with size greater than 100 km^2). It can also be an abstraction of a real world entity referring to the possible (typically infinite) outcomes of a real or a hypothetical experiment (e.g. the population of all possible annual flows in a river cross-section). Here we deal with populations of the latter type and, because of this, it is not necessary to use the term population at all—and hence to define it. Rather, the notions of a stochastic variable and a stochastic process suffice. Therefore, we will not use terms like *population mean* to distinguish from the *sample mean*. Instead, we will refer to the former concept with the terms like *true mean*, *ensemble mean* or simply *mean*, where the term *ensemble* suggests all possible outcomes of repeated experiments.

On the contrary, the term *sample* has a clear definition. Specifically, a sample of size (or length) n of a stochastic variable $\underline{x}$, defined on a basic set Ω, with probability density function $f(x)$, is a sequence of n *independent identically distributed* (IID) stochastic variables $(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n)$ defined on the sample space $\Omega_n = \Omega \times \cdots \times \Omega$, each having density $f(x)$ (Papoulis, 1990, p. 238). After observation of the variables $\underline{x}_i$, to each variable there corresponds *one* numerical value. Consequently, we will have a numerical sequence $x_1, x_2, \dots, x_n$, called the *observed sample*. It is clear from this definition that a sample is not a subset of the population, as some may think, but a concept related to the Cartesian product of the population.

The concept of a sample is, thus, related to sequences of two types: an abstract sequence of stochastic variables and the corresponding sequence of their numerical values. It has been a common practice to use the term *sample* indistinguishably for both sequences, omitting the term *observed* from the latter. However, the two concepts are fundamentally different and we should be attentive to distinguish each time in which of the two cases the term sample refers to.

The above definition and in particular the IID specification suggests that the construction of a sample of size n, or the *sampling*, is done by performing n repetitions of an experiment. The repetitions should be independent to each other and be performed under virtually the same conditions. However, in dealing with natural phenomena (out of the laboratory) it is not possible to repeat the same experiment, and thus literally there cannot be sampling. Instead, what is actually done is measurement of the natural process at different times. As a consequence, it is not possible to ensure that independence and same conditions will hold. Actually, in most of the cases we can be sure about the opposite. Then the use of classical statistics may become dangerous as the estimates and inferences may be totally wrong.

Still, however, we can do our job in a reliable manner if, instead of using classical statistics, we rely on stochastics. Actually, there is the following correspondence between the classical statistical concepts and the stochastic ones:

Classical statistics (independence) →	Statistics within stochastics (dependence)
Sample →	Stochastic process (discrete or discretized)
Observed sample →	Time series

Typically, the use of stochastics assuming dependence makes the mathematical derivations and calculations more complicated, while the resulting uncertainty is greater when there is dependence.

4.3 Expectation and its estimation

As we have stressed in Chapter 2, functions of stochastic variables, e.g. $\underline{z} \coloneqq g(\underline{x})$ are stochastic variables and expected values of stochastic variables are regular variables; for example $\mathrm{E}[\underline{x}]$ and $\mathrm{E}[g(\underline{x})]$ are constants—neither functions of x nor of $\underline{x}$—i.e.:

$$\mathrm{E}[\underline{x}] \coloneqq \int_{-\infty}^{\infty} x f(x)\mathrm{d}x =: \mu, \qquad \mathrm{E}[g(\underline{x})] \coloneqq \int_{-\infty}^{\infty} g(x) f(x)\mathrm{d}x, \tag{4.1}$$

where $f(x)$ is the probability density function. It should be stressed that these expectations are not time averages. Sometimes to make it clearer we call them true or ensemble means, variances, covariances etc. For an ergodic process, true expectations are related to time averages through the following asymptotic relationship (section 3.4):

$$\underline{\hat{G}}^{(\infty)} \coloneqq \lim_{T\to\infty} \frac{1}{T} \int_0^T g\left(\underline{x}(t)\right) dt = \mathrm{E}[g(\underline{x}(t))] =: G \tag{4.2}$$

We notice that the left-hand side, $\underline{\hat{G}}^{(\infty)}$, is a random variable while the right-hand side, G, is a regular variable; their equality implies that the variance of $\underline{\hat{G}}^{(\infty)}$ is zero.

When dealing with data from a process $\underline{x}(t)$ with a joint distribution function that is unknown, neither the left- nor the right-hand side of (4.2) can be known a priori. Assuming that we have a time series, at a time step D, with observations $x_\tau \coloneqq (1/D)\int_{(\tau-1)D}^{\tau D} x(u)\mathrm{d}u$, τ = 1, ..., n (see equation (3.1)) we can approximate the left-hand side by:

$$\hat{G} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n} g(x_\tau) \tag{4.3}$$

The regular variable $\hat{G}$ is called an *estimate* of the true expectation G. Replacing in equation (4.3) the values x_τ with the stochastic variables $\underline{x}_\tau$ we define:

$$\underline{\hat{G}} \coloneqq \frac{1}{n}\sum_{\tau=1}^{n} g(\underline{x}_\tau) \tag{4.4}$$

The stochastic variable $\underline{\hat{G}}$ is called an *estimator* of the true expectation G. In classical statistics $\underline{\hat{G}}$ is also called a *statistic*, where the latter term means a (scalar) function of the *sample vector* $\underline{\boldsymbol{x}} \coloneqq [\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n]^{\mathrm{T}}$.

While the above procedure to form an estimator $\underline{\hat{G}}$ of the true expectation G is useful in many cases, we should have in mind that many different estimators can be formulated for a certain parameter G. An estimator is typically biased (with few exceptions, the most notable being the estimator of the mean; see below), meaning that:

$$\mathrm{E}[\underline{\hat{G}}\,] \neq G \tag{4.5}$$

A formal definition of bias is:

$$b \coloneqq \mathrm{E}[\underline{\hat{G}}\,] - G \tag{4.6}$$

An estimator is also characterized by its variance and its mean square error, i.e.

$$\gamma_G \coloneqq \mathrm{var}[\underline{\hat{G}}\,], \qquad e_G \coloneqq \mathrm{E}\left[\left(\underline{\hat{G}} - G\right)^2\right] = \gamma_G + b^2 \tag{4.7}$$

An estimator is called:

- *unbiased* if $b = 0$.
- *consistent* if, with probability 1, $\underline{\hat{G}} - G \to 0$ as $n \to \infty$;
- *best* if e_G is minimum.
- *most efficient* if it is unbiased and best.

What is most fundamental to keep as the central point from the above discussion is this. When dealing with quantification of uncertainty, for each parameter there are four different concepts, with slightly different names but very different meaning and content. These are often confused in the literature and the same symbol and name are used for all, which creates confusion and may result in wrong conclusions. Table 4.1 clarifies the four different concepts using the variance as an example.

Table 4.1 Different variants of the variance, as an example for clarifying the four different concepts.

Name	Symbol and definition	Type of variable	Type of determination
Variance (true)	$\gamma_0 := \int_{-\infty}^{\infty} \left(\underline{x} - \mu\right)^2 f(x)\mathrm{d}x$	Regular variable	Theoretical calculation from model (by integration)
Variance estimate	$\hat{\gamma}_0 := \frac{1}{n}\sum_{\tau=1}^{n}(x_\tau - \hat{\mu})^2$ where: $\hat{\mu} := \frac{1}{n}\sum_{\tau=1}^{n} x_\tau$	Regular variable	Estimation from data—but model is also necessary (e.g. to calculate the estimation bias and uncertainty)
Variance estimator	$\underline{\hat{\gamma}}_0 := \frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^2$ where: $\underline{\hat{\mu}} := \frac{1}{n}\sum_{\tau=1}^{n}\underline{x}_\tau$	Stochastic variable	Theoretical calculation from model
Variance estimator limit	$\underline{\hat{\gamma}}_0^{(\infty)} = \lim_{T\to\infty}\frac{1}{T}\int_0^T \left(\underline{x}(t) - \underline{\hat{\mu}}^{(\infty)}\right)^2 \mathrm{d}t$ where $\underline{\hat{\mu}}^{(\infty)} := \lim_{T\to\infty}\frac{1}{T}\int_0^T \underline{x}(t)\,\mathrm{d}t$	Stochastic variable, which for an ergodic process has zero variance and becomes a regular variable, equal to γ_0	Theoretical calculation from model

We may notice in Table 4.1 that the data can be used only with one of the variance variants, namely the variance estimate, while a theoretical model is necessary to determine any of them. Even for the variance estimate, a model is necessary to estimate the estimation bias and uncertainty. And before specifying that model, it is fundamentally necessary to ensure that the assumptions of stationarity and ergodicity are valid for the process and the data we are dealing with. If they are valid, then the four concepts become three because the variance estimator limit becomes identical to the true variance. But if stationarity and ergodicity do not hold, then one may again use the data, do the

calculations and find a result. However, this result is meaningless and cannot be called the variance estimate.

4.4 Moment estimators

The estimator of the noncentral moment (moment about the origin) of order p, μ'_p, of a stochastic variable $\underline{x}$, formed according to the method described in section 4.3, is:

$$\underline{\hat{\mu}}'_p := \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i^p \tag{4.8}$$

It can be proved (Kendall and Stewart, 1963, p. 229) that:

$$\mathrm{E}\left[\underline{\hat{\mu}}'_p\right] = \mu'_p \tag{4.9}$$

Consequently, the noncentral moment estimators are unbiased. If $\underline{x}_i$ is a (IID) sample of size n then the variance of the estimator is:

$$\mathrm{var}\left[\underline{\hat{\mu}}'_p\right] = \frac{1}{n}\left(\mu'_{2p} - \mu'^2_p\right) \tag{4.10}$$

It can be observed that if the moments are finite, then the variance tend to zero as $n \to \infty$; therefore, the estimator is consistent. However, if $\underline{x}_i$ is a stochastic process (with time dependence) then (4.10) does not hold, even for p as low as 1.

The estimator of the central moment μ_p, is:

$$\underline{\hat{\mu}}_p := \frac{1}{n}\sum_{\tau=1}^{n} \left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^p \tag{4.11}$$

where $\underline{\hat{\mu}} \equiv \underline{\hat{\mu}}'_1$ is the estimator of the mean. This is a biased estimator for any $p > 1$. Even for relatively low p (e.g. 2-4), the bias can be substantial, in the case that the process exhibits long-range dependence (see section 4.6 about the variance). In the case of (IID) samples and low p, the bias is much smaller and can be easily quantified (see e.g. Koutsoyiannis, 1997). For higher p the estimation of moments becomes almost impossible; this applies not only to the biased estimators of central moments, but also to the unbiased estimators of noncentral moments. The reasons are the high variance and the extraordinarily high skewness of the estimators, which means that their expectation can be different from the mode (the most probable value) by orders of magnitude. Because of that, classical moments have been called *unknowable* (see Digression 4.B) and its estimation from data is not recommended. In Chapter 6 we will study a new type of moments, the knowable moments (K-moments) which can be reliably estimated for high orders and are particularly useful in analyses of extremes.

In the framework developed and followed in this text, we avoid estimation of classical moments of order higher than 2. For this reason, in the following sections we will study the estimators of classical moments of orders 1 and 2 only.

Digression 4.B: Are classical moments knowable?

The estimators of the noncentral moments $\underline{\hat{\mu}}'_p$ (or even the central ones if μ is known a priori, which however is almost never the case) are in theory unbiased, but it is impractical to use them in estimation if $p > 2$ (cf. Lombardo *et al.* 2014). It is well known that for large p and positive x_i the following approximate relationship holds:

$$\left(\sum_{i=1}^{n} x_i^p\right)^{1/p} \approx \max_{1\le i\le n}(x_i)$$

This is related to the well-known mathematical fact that the maximum norm is the limit of the p-norm as $p \to \infty$. This result can be generalized for x_i that are not necessarily positive but satisfy the condition $\max_{1\le i\le n}(x_i) > |\min_{1\le i\le n}(x_i)|$. A numerical illustration of how fast the convergence of the left-hand side to the right-hand side of the above equation is provided in Table 4.2.

Table 4.2 Illustration of the fact that raising to a power and adding converges fast to the maximum value.

Linear, $p = 1$	Pythagorean, $p = 2$	Cubic, $p = 3$	High order, $p = 8$
$3 + 4 = 7$	$3^2 + 4^2 = 5^2$	$3^3 + 4^3 = 4.5^3$	$3^8 + \mathbf{4}^8 \approx \mathbf{4}^8$
$3 + 4 + 12 = 19$	$3^2 + 4^2 + 12^2 = 13^2$	$3^3 + 4^3 + 12^3 = 12.2^3$	$3^8 + 4^8 + \mathbf{12}^8 \approx \mathbf{12}^8$

Therefore, for relatively large p the estimate of μ'_p will be:

$$\hat{\mu}'_p = \frac{1}{n}\sum_{i=1}^{n} x_i^p \approx \frac{1}{n}\left(\max_{1\le i\le n}(x_i)\right)^p$$

(Note that for large p the term $(1/n)$ in the right-hand side can be omitted with a negligible error). Thus, for an unbounded variable $\underline{x}$ and for large p, we can conclude that $\underline{\hat{\mu}}'_p$, while theoretically is an unbiased estimator of μ'_p, in practice it is more an estimator of an extreme quantity than an estimator of μ'_p. (As we will see in section 4.12, the estimated quantity is the nth order statistic raised to power p). This happens because the convergence of $\underline{\hat{\mu}}'_p$ to μ'_p is very slow, while the convergence to the maximum value is fast.

This is further illustrated in Figure 4.1 for the eighth moment of a process specified in the figure caption. While even for n as large as 64 000 the sample moment estimate continues to be smaller, by several moments of magnitude, than the theoretical value, the proximity of the moment estimate to the maximum value is evident even for n as small as 10. The jagged shapes of the curves are a clear indication of the dominance of maxima in the moment estimation: the steps occur when a new higher maximum value enters the sample, while the gradual decreases before those are due to the increase of the sample size without a higher maximum value. The ensemble simulation results in the right panel show that the 99% prediction limits (see their definition in section 4.11) from 1000 simulations are not able even to envelop the true value.

As a result, unless p is very small, μ'_p is not a knowable quantity: we cannot infer its value from a sample. This is the case even if n is very large as in Figure 4.1. Also, the various $\underline{\hat{\mu}}'_p$ are not independent to each other as they only differ on the power to which the maximum value is raised.

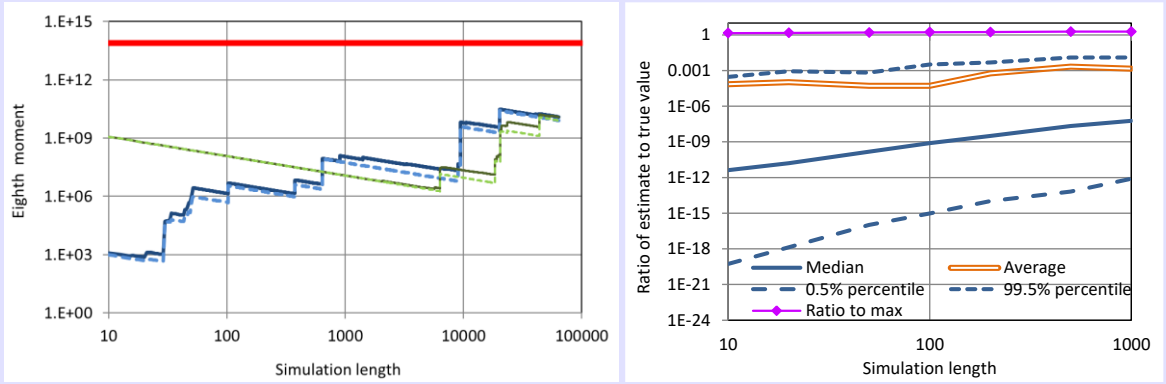

Figure 4.1 Illustration of the slow convergence of the sample estimate of the eighth noncentral moment to its true value, which is depicted as a thick horizontal line and corresponds to a lognormal distribution LN(0,1) where the process is an exponentiated Hurst-Kolmogorov process with Hurst parameter $H = 0.9$. **(left)** The sample moments are estimated from a single simulation of that process with length 64 000, where parts of this time series with sample size n from 10 to 64 000 are used for the estimation. Subsetting of the time series to sample size n was done either from the beginning to the end (thicker lines) or from the end to the beginning (finer lines). Continuous lines in the two cases represent the eighth moment estimates, $\sum_{i=1}^{n} x_i^8 /n$, and dashed lines represent maximum values, $\left(\max_{1\le i\le n}(x_i)\right)^8/n$. **(right)** Sampling distribution of the eighth moment estimator $\sum_{i=1}^{n} \underline{x}_i^8 /n$ estimated from 1000 simulated series of length 1000 each and visualized by the 99% prediction limits (percentiles), the median and the average, plotted as ratios to the true value. Theoretically, the ratio should be 1, but it is smaller by many orders of magnitude, and the convergence to 1 is very slow. The ratio to $\left(\max_{1\le i\le n}(x_i)\right)^8/n$, also plotted, is close to 1. (Source: Koutsoyiannis, 2019a.)

4.5 Sample mean estimator and effective sample size

According to (4.12), the estimator of the true mean μ is defined as:

$$\underline{\hat{\mu}} := \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i \tag{4.12}$$

Another common notation of the mean estimator is $\overline{\underline{x}}$. The estimator is unbiased ($\mathrm{E}\left[\underline{\hat{\mu}}\right] = \mathrm{E}\left[\underline{x}\right] = \mu$). Its numerical value $\hat{\mu} := (1/n)\sum_{i=1}^{n} x_i$, else denoted as $\overline{x}$, is called the *observed mean* or simply the *average*. If $\underline{x}_i$ is a (IID) sample of size n then the variance of the estimator is:

$$\mathrm{var}\left[\underline{\hat{\mu}}'_p\right] = \frac{\mathrm{var}\left[\underline{x}\right]}{n} = \frac{\gamma_1}{n} \tag{4.13}$$

regardless of the distribution function of $\underline{x}$. However, if $\underline{x}_i$ is a stochastic process (with dependence) then combining (3.13) and (4.12) we conclude that:

$$\underline{\hat{\mu}} = \underline{x}_1^{(n)} = \frac{\underline{X}(nD)}{nD} \tag{4.14}$$

where the superscript in parenthesis indicates that the discretization scale is nD. Consequently:

$$\mathrm{var}\left[\underline{\hat{\mu}}'_p\right] = \mathrm{var}\left[\frac{\underline{X}(nD)}{nD}\right] = \gamma(nD) = \gamma_n \tag{4.15}$$

Both equations (4.13) and (4.15) suggest that the estimator is consistent (assuming ergodicity). However, equations (4.13) and (4.15) may result in quite different values of the variance. By means of these two equations we can define the notion of the ''equivalent'' (or ''effective'') sample size n' in the classical statistics (IID) sense (Koutsoyiannis and Montanari, 2007). This is the sample size of a hypothetical IID sample of a variable $\underline{x}$ with variance γ_1 whose variance of the mean equals γ_n; symbolically:

$$\frac{\gamma_1}{n'} = \gamma_n \Leftrightarrow n' = \frac{\gamma_1}{\gamma_n} \tag{4.16}$$

As an example, in an HK process, in which $\gamma_n = \lambda(\alpha/nD)^{2-2H}$ (equation (3.81)), we will have:

$$n' = n^{2-2H} \tag{4.17}$$

In white noise ($H = 0.5$), clearly $n' = n$. However, if $H = 0.9$ and $n = 1000$ then $n' = 4$ (a big difference from 1000!). Thus, a time series of 1000 terms of that HK process is equivalent to a (classical, IID) sample of only 4 terms. This example shows the dramatic increase of uncertainty in case of dependence.

4.6 Climacogram estimator and its bias

The *typical variance estimator*:

$$\underline{\hat{\mu}}_2 \equiv \underline{\hat{\gamma}}_1 \coloneqq \frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^2 \tag{4.18}$$

is well known to be biased. It is also well known from elementary classical statistics books that the replacement of n with $n-1$ in the denominator of the right-hand side makes the estimator unbiased. Thus, the *classical variance estimator* is:

$$\underline{\hat{\gamma}}_1^* \coloneqq \frac{1}{n-1}\sum_{\tau=1}^{n}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)^2 = \frac{n}{n-1}\underline{\hat{\gamma}}_1 \tag{4.19}$$

This is also known as *sample variance* or *unbiased variance estimator*. However, the latter term is not correct: In stochastic processes describing natural phenomena, this slight change does not make the estimator unbiased. Here we use the term *typical* when we divide the sum by n (equation (4.18)) and *classical* when we divide by $n-1$ (equation (4.19)). We will use the same terminology for covariances below and we will explain the reasons that we prefer the typical over the classical.

In stochastic processes the bias can be determined analytically in terms the climacogram as follows (see also Koutsoyiannis 2003, 2011a, 2016):

$$\begin{aligned}\mathrm{E}\left[\underline{\hat{\gamma}}_1\right] &= \frac{1}{n}\,\mathrm{E}\left[\sum_{\tau=1}^{n}\left(\left(\underline{x}_\tau-\mu\right)-\left(\underline{x}_1^{(n)}-\mu\right)\right)^2\right] \\ &= \frac{1}{n}\mathrm{E}\left[\sum_{\tau=1}^{n}\left(\underline{x}_\tau-\mu\right)^2\right]-2\frac{1}{n}\,\mathrm{E}\left[\left(\underline{x}_1^{(n)}-\mu\right)\sum_{\tau=1}^{n}\left(\underline{x}_\tau-\mu\right)\right] \\ &\quad+\mathrm{E}\left[\left(\underline{x}_1^{(n)}-\mu\right)^2\right]\end{aligned} \tag{4.20}$$

Since $\sum_{\tau=1}^{n}\left(\underline{x}_\tau-\mu\right)=n\left(\underline{x}_1^{(n)}-\mu\right)$ we find after the algebraic manipulations:

$$\mathrm{E}\left[\underline{\hat{\gamma}}_1\right]=\gamma_1-\gamma_n=\left(1-\frac{\gamma_n}{\gamma_1}\right)\gamma_1=\left(1-\frac{1}{n'}\right)\gamma_1 \tag{4.21}$$

and

$$\mathrm{E}\left[\underline{\hat{\gamma}}_1^*\right]=\frac{n}{n-1}\left(\gamma_1-\gamma_n\right)=\frac{\left(1-\gamma_n/\gamma_1\right)}{1-1/n}\gamma_1=\frac{\left(1-1/n'\right)}{1-1/n}\gamma_1 \tag{4.22}$$

Likewise, for the climacogram at scale $k=\kappa D$, if the observation period is $L=nk$, the estimators become:

$$\underline{\hat{\gamma}}_\kappa\equiv\underline{\hat{\gamma}}(k)\coloneqq\frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau^{(\kappa)}-\underline{\hat{\mu}}\right)^2,\qquad\underline{\hat{\gamma}}_\kappa^*\equiv\underline{\hat{\gamma}}^*(k)\coloneqq\frac{n}{n-1}\underline{\hat{\gamma}}(k) \tag{4.23}$$

and their expectations are:

$$\mathrm{E}\left[\underline{\hat{\gamma}}(k)\right]=\gamma(k)-\gamma(L)=\left(1-\frac{\gamma(L)}{\gamma(k)}\right)\gamma(k),\qquad\mathrm{E}\left[\underline{\hat{\gamma}}^*(k)\right]=\frac{1-\gamma(L)/\gamma(k)}{1-k/L}\gamma(k) \tag{4.24}$$

The above equations show that there is no gain in using the *classical* estimator (dividing by $n-1$) of variance $\underline{\hat{\gamma}}_1^*$ (or $\underline{\hat{\gamma}}^*(k)$). The equations are simpler if we use the *typical* estimator $\underline{\hat{\gamma}}_1$ (or $\underline{\hat{\gamma}}(k)$) (dividing by n). As we will see below, the *typical* estimator is also preferable in fitting distributional parameters. Whatever estimator we use, there is estimation bias which should be taken into account in model fitting.

4.7 Covariance and autocovariance estimators

The *typical* and *classical* estimators of covariance:

$$\underline{\hat{c}}_{xy}\coloneqq\frac{1}{n}\sum_{\tau=1}^{n}\left(\underline{x}_\tau-\underline{\hat{\mu}}\right)\left(\underline{y}_\tau-\underline{\hat{\mu}}\right),\qquad\underline{\hat{c}}_{xy}^*\coloneqq\frac{1}{n-1}\sum_{\tau=1}^{n}\left(\underline{x}_\tau-\underline{\hat{\mu}}\right)\left(\underline{y}_\tau-\underline{\hat{\mu}}\right) \tag{4.25}$$

are both biased if $\underline{x}_\tau$ and $\underline{y}_\tau$ are stochastic processes non identical to white noise. For example, if they are HK processes with common Hurst parameter H, then the expectation of $\underline{\hat{c}}_{xy}$ is (Koutsoyiannis, 2003):

$$\mathrm{E}\left[\underline{\hat{c}}_{xy}\right]=\left(1-\frac{1}{n^{2-2H}}\right)c_{xy}=\left(1-\frac{1}{n'}\right)c_{xy} \tag{4.26}$$

In the case of autocovariance estimation, it well known that there is downward bias (Wallis and O'Connell, 1972; Salas, 1993, p. 19.10). The *typical* estimator of the lag η autocovariance is:

$$\underline{\hat{c}}_\eta := \frac{1}{n}\sum_{\tau=1}^{n-\eta}\left(\underline{x}_\tau - \underline{\hat{\mu}}\right)\left(\underline{x}_{\tau+\eta} - \underline{\hat{\mu}}\right) \tag{4.27}$$

and it has been a common practice to prefer it over the classical estimator (with division by $n-1$ or $n-\eta$), particularly when we use autocovariance to estimate the power spectrum. The expectation of $\underline{\hat{c}}_\eta$ is (see also Koutsoyiannis, 2003):

$$\begin{aligned}\mathrm{E}\left[\underline{\hat{c}}_\eta\right] &= \frac{1}{n}\,\mathrm{E}\left[\sum_{\tau=1}^{n-\eta}\left(\left(\underline{x}_\tau - \mu\right) - \left(\underline{x}_1^{(n)} - \mu\right)\right)\left(\left(\underline{x}_{\tau+\eta} - \mu\right) - \left(\underline{x}_1^{(n)} - \mu\right)\right)\right] \\ &= \frac{1}{n}\mathrm{E}\left[\sum_{\tau=1}^{n-\eta}\left(\underline{x}_\tau - \mu\right)\left(\underline{x}_{\tau+\eta} - \mu\right)\right] \\ &\quad - \frac{1}{n}\,\mathrm{E}\left[\left(\underline{x}_1^{(n)} - \mu\right)\sum_{\tau=1}^{n-\eta}\left(\left(\underline{x}_\tau - \mu\right) + \left(\underline{x}_{\tau+\eta} - \mu\right)\right)\right] + \mathrm{E}\left[\left(\underline{x}_1^{(n)} - \mu\right)^2\right]\end{aligned} \tag{4.28}$$

Since $\sum_{\tau=1}^{n-\eta}\left(\underline{x}_\tau - \mu\right) = (n-\eta)\left(\underline{x}_1^{(n-\eta)} - \mu\right)$, assuming that η is small in comparison with n so that we can interchange $n - \eta$ and n, and also extend the corresponding sums, we obtain after the algebraic manipulations:

$$\mathrm{E}\left[\underline{\hat{c}}_\eta\right] \approx c_\eta - \gamma_n = \left(1 - \frac{\gamma_n}{c_\eta}\right)c_\eta = \left(1 - \frac{1}{n'}\right)\gamma_1 \tag{4.29}$$

If we estimate the autocorrelation coefficient by:

$$\underline{\hat{r}}_\eta := \frac{\underline{\hat{c}}_\eta}{\underline{\hat{\gamma}}_1} \tag{4.30}$$

then this will be biased. An approximately unbiased estimator would be:

$$\underline{\tilde{r}}_\eta := \frac{\underline{\hat{c}}_\eta + \gamma_n}{\underline{\hat{\gamma}}_1 + \gamma_n} = \frac{\underline{\hat{r}}_\eta\underline{\hat{\gamma}}_1 + \gamma_n}{\underline{\hat{\gamma}}_1 + \gamma_n} = \left(1 - \frac{1}{n'}\right)\underline{\hat{r}}_\eta + \frac{1}{n'} \tag{4.31}$$

It is stressed that the use of autocovariance and (even more) the autocorrelation estimates should be avoided in the identification and fitting phases of a stochastic model. Identification and fitting are better served by the climacogram (see Digression 4.C).

Digression 4.C: The climacogram and the climacogram-based metrics compared to standard metrics

The most popular procedure in time series modelling, is to construct the empirical autocorrelogram of the time series using equation (4.27) and assess which stochastic process (e.g., of AR or ARMA type) is suitable and how many autocorrelation terms should be preserved. It is not too difficult to illustrate that this technique can completely distort the underlying process. Figure 4.2(a) depicts the autocorrelogram of a time series with length 100, which does not seem

to have any relationship with the theoretical autocorrelation function of the model from which it was constructed. The latter is the FHK one with parameters as in the caption of Figure 3.7. Clearly, the empirical autocorrelation does not give any hint that the time series stems from a process with persistence. With that autocorrelogram one would conclude that an AR(1) model with a lag-1 autocorrelation of about 0.4 would be appropriate.

The reasons for the failure of the autocorrelogram to capture the real behaviour of the process are two. First is the bias, as analysed in section 4.7. Second, from equation (3.29) it is seen that the autocorrelation is by nature the second derivative of the climacogram standardized by variance. Estimation of the second derivative from data is too uncertain and makes a very rough graph.

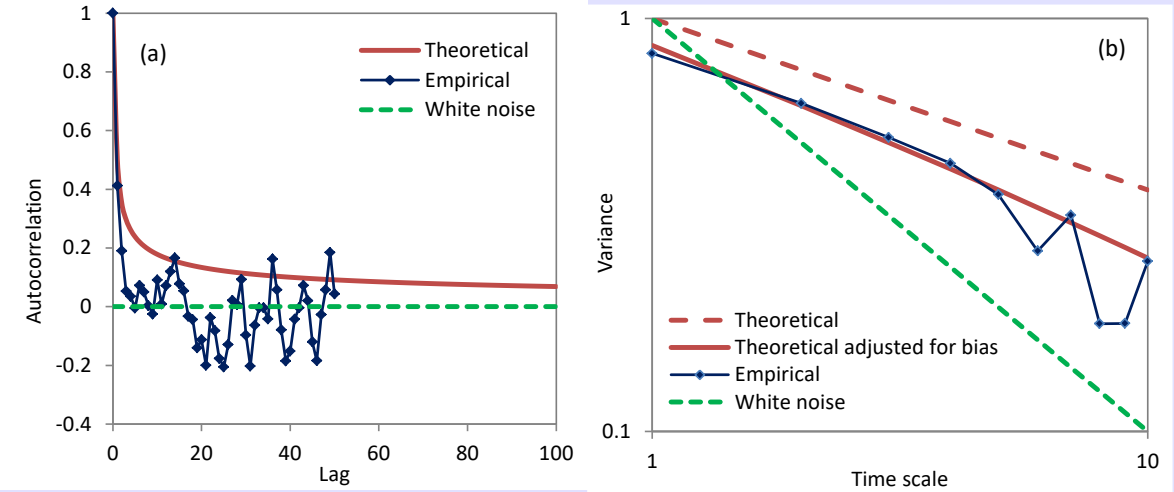

Figure 4.2 (left) Autocorrelogram and **(right)** climacogram of a time series of 100 terms generated from the FHK model with parameters as in the caption of Figure 3.7. (Source: Koutsoyiannis, 2016.)

The alternative of using the periodogram (the estimate of the power spectrum, which is the Fourier transform of the autocovariance; see section 4.10) is even worse as it entails an even rougher shape and more uncertain estimation than in the autocovariance (see also section 4.10 and Dimitriadis and Koutsoyiannis, 2015a).

It is, thus, much preferable to directly use the climacogram instead of the autocorrelogram for model identification. For our example time series, this is illustrated in Figure 4.2(b), which indicates that the long-term persistence is well captured by the empirical climacogram and the parameter H is correctly estimated (H = 0.79, based on the method presented in Koutsoyiannis, 2003, and Tyralis and Koutsoyiannis, 2011). Additional advantages of the climacogram are (a) its intactness on discretization, (b) its close relationship with entropy production and (c) its expandability to high-order moments.

4.8 Parameter estimation of distribution functions – The method of moments

Assuming a stochastic variable $\underline{x}$ with known distribution function but with unknown parameters $\boldsymbol{\theta} \coloneqq [\theta_1, \theta_2, \ldots, \theta_m]^{\mathrm{T}}$, we can denote the probability density function of $\underline{x}$ as a function $f(x, \boldsymbol{\theta})$. Here, we will examine the problem of the estimation of these parameters based on a sample vector $\underline{\boldsymbol{x}} \coloneqq [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]^{\mathrm{T}}$. Specifically, we will present the two most popular methods in statistics, namely the moments method and the maximum likelihood method. Several other general methods have been developed in statistics for parameter estimation, e.g. the maximum entropy method (e.g. Singh and Rajagopal, 1986) and the L-moments method (Hosking et al., 1985a,b; Hosking, 1990). Moreover, in practical

applications, other types of methods like graphical, tabulated, empirical and semi-empirical, have been devised. As will be seen in later chapters, here we prefer a different approach based on K-moments, over all above methods.

The method of moments is based on equating the theoretical moments of $\underline{x}$ with the corresponding sample estimates of noncentral moments. Thus, as m is the number of the unknown parameters of the distribution, we can write m equations of the form

$$\mu'_p = \hat{\mu}'_p, \qquad p = 1, \dots, m \tag{4.32}$$

where the theoretical moments μ'_p are functions of the unknown parameters given by:

$$\mu'_p = \int_{-\infty}^{\infty} x^p \, f(x, \boldsymbol{\theta}) \mathrm{d}x \tag{4.33}$$

Thus, the solution of the resulting system of the m equations gives the unknown parameters $(\theta_1, \theta_2, \dots, \theta_m)$. In general, the system of equations may not be linear and may not have an analytical solution. In the latter case the system of equations will be solved numerically.

This method is easy to apply. However, for distributions involving more than two parameters, the problem of knowability of moments intervenes and makes the method unreliable. Furthermore, when dealing with extremes we must have in mind that they are influenced by high-order moments and thus, relying on the lowest-order moment is not the best practice (see section 6.15).

Digression 4.D: Illustration of the method of moments

As an example of the implementation of the method of moments, we will determine the parameters of the normal distribution. The probability density function:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

has two parameter, μ and σ. Thus, we need two equations. Based on Table 2.3, these equations are:

$$\mu = \hat{\mu}, \qquad \sigma^2 + \mu^2 = \hat{\mu}_2 + \hat{\mu}^2 \Rightarrow \sigma^2 = \hat{\mu}_2$$

where we have used the identity $\mu'_2 = \mu_2 + \mu^2$. Consequently, the final estimates are

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad \sigma^2 \coloneqq \frac{1}{n}\sum_{\tau=1}^{n} (x_\tau - \hat{\mu})^2$$

This estimation μ is unbiased but that of σ^2 (and σ) is biased even in IID statistics (notice in the latter equation that the result contains the typical, rather the classical estimate).

As we have seen in this example, the application of the method of moments is very simple and this extends to other distribution functions.

4.9 Parameter estimation of distribution functions – The maximum likelihood method

While the method of moments is a method of approximation and has several weaknesses, as described in section 4.8, the method of maximum likelihood has a strong logical

background. We will initially present the method in a Bayesian framework and then we will see that it stands also out of that framework.

The problem that we have to resolve is to find the parameter vector $\boldsymbol{\theta}$ from the known observations $\underline{\boldsymbol{x}} = \boldsymbol{x}$. Since the observations $\boldsymbol{x}$ are known while the parameters $\boldsymbol{\theta}$ are unknown, we can regard the latter as stochastic variables $\underline{\boldsymbol{\theta}}$. This allows us to assign $\underline{\boldsymbol{\theta}}$ a probability density function $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and also express conditional densities by the Bayes theorem (equation (2.14)). This can be written in terms of densities as:

$$f_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})}{f_{\boldsymbol{x}}(\boldsymbol{x})}\, f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{4.34}$$

where we have replaced the events A and B with the vectors $\boldsymbol{x}$ and $\boldsymbol{\theta}$, respectively. The terminology used in the Bayesian framework is:

- Prior (before observation) probability density for $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$
- Posterior (after observation) probability density for $f_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{\theta}|\boldsymbol{x})$
- Likelihood for the conditional density $f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})$; this is the hypothesized model (i.e. distribution for $\underline{\boldsymbol{x}}$) given the parameters $\boldsymbol{\theta}$.

According to this terminology, we can write (4.34) in the following form:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior} \tag{4.35}$$

Since we have to assign $\underline{\boldsymbol{\theta}}$ a single value $\boldsymbol{\theta}$, the most rational choice for that value is the mode of the its distribution conditional on $\underline{\boldsymbol{x}} = \boldsymbol{x}$, i.e. the value that maximizes the posterior $f_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{\theta}|\boldsymbol{x})$. To find the mode we equate the derivative of the conditional density to 0, i.e.:

$$\frac{\mathrm{d}f_{\boldsymbol{\theta}|\boldsymbol{x}}(\boldsymbol{\theta}|\boldsymbol{x})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \Leftrightarrow \frac{1}{f_{\boldsymbol{x}}(\boldsymbol{x})}\left(\frac{\mathrm{d}f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\, f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})\frac{\mathrm{d}f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right) = \mathbf{0} \tag{4.36}$$

Since we know nothing about the prior $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, we can choose a so-called noninformative prior, which does not change with $\boldsymbol{\theta}$, i.e. $\mathrm{d}f_{\boldsymbol{\theta}}(\boldsymbol{\theta})/\mathrm{d}\boldsymbol{\theta} = 0$. In this case from (4.36) we obtain:

$$\frac{\mathrm{d}f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \tag{4.37}$$

which demands that also the likelihood be at maximum. In other words, we find $\boldsymbol{\theta}$, demanding that the density $f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})$ have a value as high as possible at the point $\underline{\boldsymbol{x}} = \boldsymbol{x}$.

If the vector $\boldsymbol{x}$ is part of a stochastic procees, determination of $f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})$ can be laborious. However, in IID statistics, $\boldsymbol{x}$ is a sample vector with independent items and thus the joint probability density function is:

$$f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{x|\boldsymbol{\theta}}(x_i|\boldsymbol{\theta}) \tag{4.38}$$

Thus, we seek a solution of:

$$\frac{\mathrm{d}\prod_{i=1}^{n} f_{x|\boldsymbol{\theta}}(x_i|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \tag{4.39}$$

We can also convert the product to a sum by taking the logarithm of $f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta})$:

$$L(\boldsymbol{x}|\boldsymbol{\theta}) \coloneqq \ln f_{\boldsymbol{x}|\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f_{x|\boldsymbol{\theta}}(x_i|\boldsymbol{\theta}) \tag{4.40}$$

The function $L(\ )$ is called the log-likelihood function. In this case, the condition of maximum is:

$$\frac{\mathrm{d}L(\boldsymbol{x}|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{1}{f_{x|\boldsymbol{\theta}}(x_i|\boldsymbol{\theta})} \frac{\mathrm{d}f_{x|\boldsymbol{\theta}}(x_i|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{0} \tag{4.41}$$

Both (4.39) and (4.41) are vector equations equivalent to m scalar equations. Solving either of them we obtain the values of the m unknown parameters.

Digression 4.E: Illustration of the maximum likelihood method

We will determine the parameters of the normal distribution from a sample using the maximum likelihood method. The probability density function of the normal distribution is:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The likelihood function is:

$$f(\boldsymbol{x}|\mu,\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

The log-likelihood function is:

$$L(\boldsymbol{x}|\mu,\sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Taking the derivatives with respect of the unknown parameters μ and σ and equating them to 0 we find

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0\,, \qquad \frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

and solving the system we obtain the final parameter estimates:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = \hat{\mu}_2$$

The results are precisely identical with those of Digression 4.D, despite the fact that the two methods are fundamentally different. The application of the maximum likelihood method is more complex than that of the method of moments. The coincidence of results found here is not the rule for all distribution functions. On the contrary, in most cases the two methods yield different results.

4.10 The estimation of power spectrum and the periodogram

We assume that a stochastic process $\underline{x}(t)$ is observed on a time-average basis at equidistant times $\tau D, \tau = 0, \dots, n-1$, where D is a time step (a total observation time $L = nD$). We have thus a time series with a finite number, n, of observations x_τ of the discrete-time process $\underline{x}_\tau$. If we study the process on the frequency domain, we have the following characteristic frequencies, dimensional (w) or dimensionless ($\omega = wD$):

Frequency resolution	$w_1 = 1/L = w_D/n$	$\omega_1 = w_1 D = D/L = 1/n$
Sampling frequency	$w_D = 1/D = n/L$	$\omega_D = w_D D = 1$
Nyquist frequency	$w_\mathrm{N} = 1/2D = n/2L$	$\omega_\mathrm{N} = w_\mathrm{N} D = 0.5$

As we will see, the Nyquist frequency ($\omega_\mathrm{N} = 0.5$) is the maximum frequency on which we can make estimates as beyond that the resulting spectrum estimates are repeated in a cyclic manner.

We are interested about estimators of the power spectrum of the discrete-time process $\underline{x}_\tau$. A first estimator can be established by utilizing the relationship between the power spectrum and the autocovariance function (equation (3.35)). From n observations we can estimate from equation (4.27) up to n autocovariance terms, $\underline{\hat{c}}_0, \underline{\hat{c}}_1, \dots, \underline{\hat{c}}_{n-1}$ (noting that most of them will not be reliably estimated). Then, by truncating equation (3.35) to a finite number of terms we can formulate an estimator of the spectrum in the form:

$$\underline{\hat{s}}_\mathrm{d}(\omega) = 2\underline{\hat{c}}_0 + 4\sum_{\eta=1}^{n-1} \underline{\hat{c}}_\eta \cos(2\pi\eta\omega) + 2\underline{\hat{c}}_n \cos(2\pi n\omega) \tag{4.42}$$

where we have put a last term for $\underline{\hat{c}}_n$ with a weight 2 (instead of 4), which, as we will see facilitates and accelerates calculations. If we have n data values x_τ, then $\underline{\hat{c}}_n \equiv 0$, but the calculation should stand in cases where use a fewer number of autocorrelations or in cases where we process true values rather than estimates (in the latter case, $c_n \neq 0$). While from first glance we can use this equation to estimate $\underline{\hat{s}}_\mathrm{d}(\omega)$ for any ω, the resulting values are not always consistent and therefore it is suggested to make estimates for a finite number of discrete frequencies $\omega_j = j\omega_1$ for some resolution ω_1 with j taking integer values as we will specify below.

The inversion of the formula to find the autocovariance estimates from power spectrum estimates is possible through the equation:

$$\underline{\hat{c}}_\eta = \omega_1 \left(\frac{\underline{\hat{s}}_\mathrm{d}(0) + (-1)^n \underline{\hat{s}}_\mathrm{d}(0.5)}{2} + \sum_{0<\omega_j<0.5} \underline{\hat{s}}_\mathrm{d}(\omega_j) \cos(2\pi\eta\omega_j) \right) \tag{4.43}$$

The estimation of $\underline{\hat{s}}_\mathrm{d}(\omega)$ is streamlined and accelerated if we use the *discrete Fourier transform* (DFT) and particularly its variant named *fast Fourier transform* (FFT), for which the required software exists on all computational environments. For a sequence of numbers $x_\tau, \tau = 0, \dots, N-1$, the DFT is defined as a sequence $u_j, j = 0, \dots, N-1$, where:

$$u_j = \frac{1}{N}\sum_{\tau=0}^{N-1} x_\tau \mathrm{e}^{-\mathrm{i}2\pi\tau j/N}, \qquad j = 0, \dots, N-1 \tag{4.44}$$

The sequence x_τ is recovered from the sequence u_j by the inverse DFT, which is:

$$x_\tau = \sum_{j=0}^{N-1} u_j \mathrm{e}^{\mathrm{i}2\pi j\tau/N}, \qquad \tau = 0, \dots, N-1 \tag{4.45}$$

The FFT is the DFT made by a fast computational algorithm; the fastest case is when n is a power of 2.

To utilize DFT and FFT in determining $\underline{\hat{s}}_\mathrm{d}(\omega)$ we write equation (4.42) as:

$$\underline{\hat{s}}_\mathrm{d}(\omega) = \sum_{\eta=0}^{n} 2\underline{\hat{c}}_\eta \cos(2\pi\eta\omega) + \sum_{\eta=1}^{n-1} 2\underline{\hat{c}}_\eta \cos(2\pi\eta\omega) \tag{4.46}$$

Setting $j = \eta$ for the first sum and $j = 2n - \eta$ for the second sum we have:

$$\underline{\hat{s}}_\mathrm{d}(\omega) = \sum_{j=0}^{n} 2\underline{\hat{c}}_j \cos(2\pi j\omega) + \sum_{j=n+1}^{2n-1} 2\underline{\hat{c}}_{2n-j} \cos(2\pi(2n-j)\omega) \tag{4.47}$$

The latter cosine, if ω is an integer multiple of $\omega_1 = 1/N$ where $N \coloneqq 2n$, equals $\cos(2\pi j\omega)$. By setting:

$$\underline{u}_j = \begin{cases} 2\underline{\hat{c}}_j, & 0 \le j \le n \\ 2\underline{\hat{c}}_{2n-j}, & n \le j \le N-1 \end{cases} \tag{4.48}$$

we can simplify (4.47) to:

$$x_\tau = \sum_{j=0}^{N-1} u_j \mathrm{e}^{\mathrm{i}2\pi j\tau/N}, \qquad \underline{\hat{s}}_\mathrm{d}(\omega) = \sum_{j=0}^{N-1} \underline{u}_j \cos(2\pi j\omega) \tag{4.49}$$

Considering that the imaginary part of $\underline{u}_j$ is zero, setting $\omega_\tau = \tau/N$, and comparing equations (4.45) and (4.49), we conclude that $\underline{\hat{s}}_\mathrm{d}(\omega_\tau)$ is the inverse DFT of $\underline{u}_j$. If we have taken care to choose n a power of 2, so will be N too and thus we can use the inverse FFT to calculate estimates $\hat{s}_\mathrm{d}(\omega_\tau)$ from estimates $\hat{c}_\eta$ for frequencies ω ranging from 0 to 0.5 with a resolution $1/N = 1/2n$. The inverse of (4.49) is:

$$\underline{u}_j = 2\underline{\hat{c}}_j = \frac{1}{N}\sum_{\tau=0}^{N-1} \underline{\hat{s}}_\mathrm{d}(\omega_\tau) \cos(2\pi j\omega_\tau), \qquad 0 \le j \le n \tag{4.50}$$

There is an alternative way to produce another estimator of the power spectrum using the DFT on the discrete-time process per se, rather than on its autocovariance. Specifically, the DFT of $\underline{x}_\tau$ is:

$$\underline{u}_j = \frac{1}{n}\sum_{\tau=0}^{n-1} \underline{x}_\tau \mathrm{e}^{-\mathrm{i}2\pi\tau j/n}, \qquad j = 0, \dots, n-1 \tag{4.51}$$

Assuming that $\underline{x}_\tau, \tau = 0, \dots, n-1$, are real-valued stochastic variables, their transformation $u_j, j = 0, \dots, n-1$, will be complex valued stochastic variables, i.e. $\underline{u}_j = \underline{u}_j^{\mathrm{R}} + \mathrm{i}\,\underline{u}_j^{\mathrm{I}}$, where $\underline{u}_j^{\mathrm{R}}$ and $\underline{u}_j^{\mathrm{I}}$ are real-valued. The inverse DFT of $\underline{u}_j$ recovers the real-valued $\underline{x}_\tau$. The sequence of the absolute values of $\underline{u}_j$ multiplied by $2n$:

$$\underline{S}_j := 2n\left|\underline{u}_j\right|^2 = 2n\left(\left(\underline{u}_j^{\mathrm{R}}\right)^2 + \left(\underline{u}_j^{\mathrm{I}}\right)^2\right) \tag{4.52}$$

is real valued and, as a function of $\omega_j = j/n$, is known as the *periodogram* of $\underline{x}_\tau$. It is another estimator of $s_{\mathrm{d}}(\omega)$ with a resolution $1/n$ (while in the estimator (4.49) this is $1/2n$). The two alternatives of estimating the power spectrum are schematically presented in Figure 4.3.

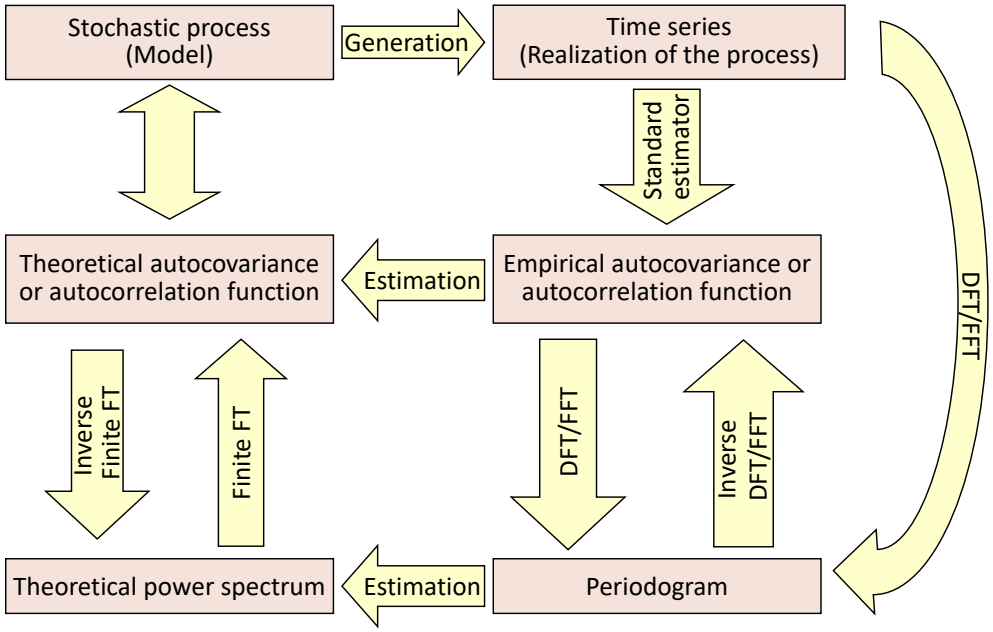

Figure 4.3 Schematic of the different paths to estimate the power spectrum.

For real-valued $\underline{x}_\tau$ the stochastic variables $\underline{u}_j$ and $\underline{S}_j$ have the following properties of symmetry:

$$\begin{gathered} \underline{u}_0 = \underline{u}_0^{\mathrm{R}} = \underline{\hat{\mu}}, \qquad \underline{u}_0^{\mathrm{I}} = 0 \\ \underline{u}_{n-j}^{\mathrm{R}} = \underline{u}_j^{\mathrm{R}}, \qquad \underline{u}_{n-j}^{\mathrm{I}} = \underline{u}_j^{\mathrm{I}}, \qquad \underline{S}_{n-j} = \underline{S}_j, \qquad 1 \le j \le n-1 \end{gathered} \tag{4.53}$$

In other words, the real component of $\underline{u}_j$ and $\underline{S}_j$ are symmetric with respect to $n/2$, while the imaginary component is antisymmetric. Consequently, if n is even, then $\underline{u}_{n/2}^{\mathrm{I}} = 0$. Because of the symmetries, starting with n real numbers x_τ we end up with n real numbers $\underline{u}_j^{\mathrm{R}}$ and $\underline{u}_j^{\mathrm{I}}$, and $n/2$ real numbers $\underline{S}_j$. The values of $\underline{S}_j$ for frequencies $\omega = j/n \le 0.5$ provide all extractable information while the larger frequencies do not add anything.

Other interesting properties of the periodogram and the related quantities are:

$$\frac{1}{n}\sum_{\tau=0}^{n-1}\underline{x}_\tau^2=\sum_{j=0}^{n-1}\left|\underline{u}_j\right|^2,\qquad \underline{\hat{\gamma}}_1=\frac{1}{n}\sum_{\tau=0}^{n-1}\left(\underline{x}_\tau-\underline{\hat{\mu}}\right)^2=\sum_{j=1}^{n-1}\left|\underline{u}_j\right|^2=\frac{1}{n}\sum_{1\le j\le n/2}\underline{S}_j-\frac{\underline{S}_{n/2}}{2n} \quad (4.54)$$

where if n is odd, the last term $\underline{S}_{n/2}$ is set to zero. The latter equation allows decomposing the variance estimate $\hat{\gamma}_1$ into partial components $\left|u_j\right|^2$, each corresponding to a particular frequency, which ranges from $\omega_1 = w_1 D = 1/n$ to $\omega_{\mathrm{N}} = w_{\mathrm{N}} D = 0.5$. The frequency 0 corresponds to the mean estimate and is not related to the variance. Any prominence (peak) in one or more $\left|\underline{u}_j\right|^2$ over the other is very often regarded as evidence of a periodic behaviour of the process with a frequency j/n (period n/j).

However, claims of periodicities without a deterministic explanation are usually meaningless. As evident from the notation in the entire section, all related concepts, including the periodogram, are estimators, i.e. stochastic variables, which produce estimates. Considered as a sequence of stochastic variables, the periodogram $\underline{S}_j$ is a nonstationary stochastic process indexed by $j = 1, \ldots, \lfloor n/2 \rfloor$. The same happens with the estimator $\underline{\hat{s}}_{\mathrm{d}}(\omega_j)$, which is a nonstationary stochastic process indexed by $j = 1, \ldots, n$, as well as with the covariance estimator $\underline{\hat{c}}_\eta$. The produced shapes in graphs of estimates indicate high variability and roughness, and thus possible peaks are most probably random effects. Note that by increasing the number of observations, the variability and roughness are not necessarily decreasing (cf. (4.52), where $\left|\underline{u}_j\right|^2$ is multiplied by $2n$).

An illustration is given in Figure 4.4 for a time series generated from the discrete-time HK process, where several peaks appear, all of which are random effects. A simple technique to see that these are random effects is to split the time series into two halves, three thirds, etc. and inspect whether the peaks appear systematically in all cases (Koutsoyiannis and Georgakakos, 2006). Splitting the time series and taking the average of the different parts for the same frequency is a method of smoothing the periodogram (for details and other smoothing methods see Papoulis, 1991). The least square trend (power law) of the spectrum estimates from autocovariance is also shown in the log-log spectrum plot of Figure 4.4 (bottom-right). The slope is -1.24, an inconsistent value as theoretically the slope cannot be steeper than -1 (the slope of the theoretical curve, also shown in the figure, is $1 - 2H = -0.6 > -1$). This inconsistency is not expected to be resolved by the aforementioned smoothing of the power spectrum. For these reasons, it could be suggested to use the climacospectrum, instead of the power spectrum for estimation of slopes (Koutsoyiannis, 2017).

4.11 Interval estimation and confidence intervals

An *interval estimate* of a parameter λ of a distribution function is an interval of the form (θ_1, θ_2), where θ_1 and θ_2 are functions of the observed sample vector $\boldsymbol{x}$, i.e., $\theta_1 = g_1(\boldsymbol{x})$ and $\theta_2 = g_2(\boldsymbol{x})$. If we replace the observed sample with the sample (or the part of a stochastic process), then the interval's limits become stochastic variables, $\underline{\theta}_1 = g_1(\underline{\boldsymbol{x}})$ and $\underline{\theta}_2 = g_2(\underline{\boldsymbol{x}})$. The interval $(\underline{\theta}_1, \underline{\theta}_2)$ is an interval estimator of the parameter λ.

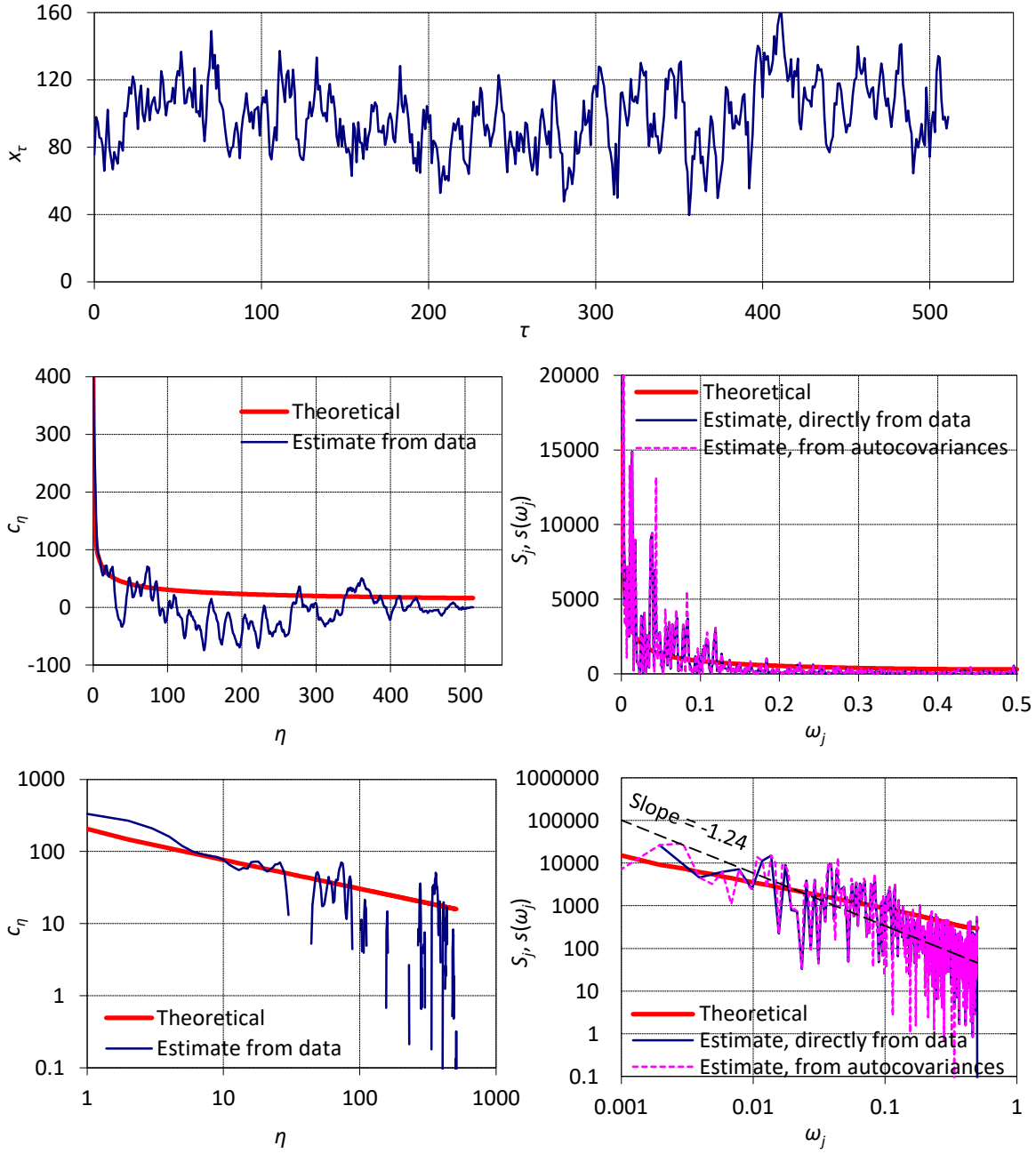

Figure 4.4 (upper) A plot of a time series with $n = 512$ terms generated from the Gaussian HK model with $H = 0.8, \mu = 100, \gamma_1 = 400$. **(middle)** The autocovariance and power spectrum of the generating stochastic process and their estimates. **(lower)** Same as middle but with logarithmic axes. The least square trend (power law) of the estimates from autocovariance, with slope $= -1.24$ is also plotted in the spectrum panel.

We say that the interval $(\underline{\theta}_1, \underline{\theta}_2)$ is a *C-confidence interval* of the parameter λ if:

$$P\{\underline{\theta}_1 < \lambda < \underline{\theta}_2\} = C \tag{4.55}$$

where C is a given constant $(0 < C < 1)$ called the *confidence coefficient*, and the limits $\underline{\theta}_1, \underline{\theta}_2$ are called *C-confidence limits*. Usually we choose values of C near 1 (e.g. 0.9, 0.95, 0.99, so that the inequality in (4.55) become near certainty). In practice the term

confidence limits is often (loosely) used to describe the numerical values of the statistics $\underline{\theta}_1, \underline{\theta}_2$, whereas the same happens for the term confidence interval.

In order to provide a method for the calculation of a confidence interval, we will assume that the statistic $\underline{\theta} = g(\underline{\boldsymbol{x}})$ is a point estimator of the parameter λ with distribution function is $F_\theta(\theta)$. Based on this distribution function it is possible to calculate two positive numbers ξ_1 and ξ_2, so that the estimation error $\underline{\theta} - \lambda$ lie in the interval $(-\xi_1, \xi_2)$ with probability C, i.e.:

$$P\{\lambda - \xi_1 < \underline{\theta} < \lambda + \xi_2\} = C \tag{4.56}$$

and at the same time the interval $(-\xi_1, \xi_2)$ be as small as possible. If the distribution of $\underline{\theta}$ is symmetric then the interval $(-\xi_1, \xi_2)$ has minimum length for $\xi_1 = \xi_2$. For asymmetric distributions, it is difficult to calculate the minimum interval, thus we simplify the problem by splitting (4.56) into two equations, namely, $P\{\underline{\theta} < \lambda - \xi_1\} = P\{\underline{\theta} > \lambda + \xi_2\} = (1 - C)/2$. Equation (4.56) can be written as:

$$P\{\underline{\theta} - \xi_2 < \lambda < \underline{\theta} + \xi_1\} = C \tag{4.57}$$

Consequently, the confidence limits we are seeking are $\underline{\theta}_1 = \underline{\theta} - \xi_2$ and $\underline{\theta}_2 = \underline{\theta} - \xi_1$.

Although equations (4.56) and (4.57) are equivalent, their statistical interpretation is different. The former is a *prediction*, i.e. it gives the *prediction interval** of the stochastic variable $\underline{\theta}$. The latter is an *interval parameter estimator*, i.e. it gives the confidence limits of the unknown parameter λ, which is not a stochastic variable.

Classical statistical texts provide expressions for interval estimators of some common parameters, such as the mean and variance of the normal distribution of IID samples. However, in most real-world cases we deal with problems much more demanding than such idealized cases. The distributions may be non-normal, the parameter of interest may not be the mean or the variance, and instead of a sample we may have a stochastic process. Then analytical calculation of confidence limits becomes impossible. Naturally, the method of choice for such (that is, most) cases is the Monte Carlo simulation. General methodologies for tackling the problem have been proposed by Tyralis et al. (2013) and Tyralis and Koutsoyiannis (2014).

4.12 Order statistics

Let $\underline{x}$ be a stochastic variable and $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ be IID copies of it, forming a sample. We can rearrange them in increasing order of magnitude such that $\underline{x}_{(i:n)}$ be the ith smallest of the n, i.e.:

$$\underline{x}_{(1:n)} \leq \underline{x}_{(2:n)} \leq \cdots \leq \underline{x}_{(n:n)} \tag{4.58}$$

The stochastic variable $\underline{x}_{(i:n)}$ is termed the ith *order statistic*. The minimum and maximum are, respectively,

* The terms confidence limits, confidence interval, confidence coefficient etc. are also used for this prediction form of the equation.

$$\underline{x}_{(1:n)} = \min\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n\right), \qquad \underline{x}_{(n)} \coloneqq \underline{x}_{(n:n)} = \max\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n\right) \tag{4.59}$$

and represent special cases of the order statistics, the lowest and the highest.

If $f(x)$ and $F(x)$ are respectively the density and distribution function of $\underline{x}$, then the density function of $\underline{y} \coloneqq \underline{x}_{(i:n)}$ is (Papoulis 1990):

$$f_y(y) \coloneqq f_{(i:n)}(y) = (n-i+1)\binom{n}{i-1}\left(F(y)\right)^{i-1}\left(1-F(y)\right)^{n-i} f(y) \tag{4.60}$$

Now if we define the stochastic variable $\underline{u} \coloneqq F\left(\underline{y}\right) = F(\underline{x}_{(i:n)})$, then according to (2.11):

$$f_u(u) \coloneqq \frac{f_y\left(F^{-1}(u)\right)}{f\left(F^{-1}(u)\right)} = (n-i+1)\binom{n}{i-1}u^{i-1}(1-u)^{n-i} = \frac{u^{i-1}(1-u)^{n-i}}{\mathrm{B}(i, n-i+1)} \tag{4.61}$$

This is the density of the *Beta distribution* function and hence:

$$F_{(i:n)}(x) = P\left\{\underline{x}_{(i:n)} \le x\right\} = P\left\{\underline{u} \le F(x)\right\} = \frac{\mathrm{B}_{F(x)}(i, n-i+1)}{\mathrm{B}(i, n-i+1)} \tag{4.62}$$

For the special cases of the minimum and maximum we have, respectively,

$$F_{(1:n)}(x) = \frac{\mathrm{B}_{F(x)}(1, n)}{\mathrm{B}(1, n)} = 1 - \left(1-F(x)\right)^n, F_{(n:n)}(x) = \frac{\mathrm{B}_{F(x)}(n, 1)}{\mathrm{B}(n, 1)} = \left(F(x)\right)^n \tag{4.63}$$

As we will see in Chapter 5 and Chapter 6, the order statistics are quite important for studying extremes.

4.13 Samples vs. time series and forecast-oriented estimation

As we have seen, in the classical statistics, samples are by definition sets of IID stochastic variables. Classical statistical estimations make use of the entire vector of available observations. But what if instead of a sample we have a stochastic process with time dependence and instead of an observed sample we have a time series? Apparently, things can be quite different and generally we should avoid uncritical use of classical statistics. To illustrate the difference, we consider the following problem: How many past terms will we use to estimate an average that is representative of the future mean for a period of length κ? This is not necessarily the "global" average estimated by the entire time series of observations.

The statistic sought is the "local" mean of the future period of length κ conditional on the present and past, i.e.:

$$\underline{\mu}_\kappa \coloneqq \mathrm{E}\left[\frac{1}{\kappa}\left(\underline{x}_1 + \dots + \underline{x}_\kappa\right) | \underline{x}_0, \underline{x}_{-1}, \dots\right] \tag{4.64}$$

Let us assume that we have a large number n of observations of the present and past but we choose to use $\nu \le n$ of them for the estimation:

$$\hat{\underline{\mu}}_\nu \coloneqq \frac{1}{\nu}\left(\underline{x}_0 + \underline{x}_{-1} + \dots + \underline{x}_{-\nu+1}\right) \tag{4.65}$$

To answer the question, it suffices to find ν which minimizes the mean square error $A(\kappa,\nu) \coloneqq \mathrm{E}\left[\left(\underline{\hat{\mu}}_\nu - \underline{\hat{\mu}}_\kappa\right)^2\right]$, which can be written as:

$$\begin{aligned} A(\kappa,\nu) &= \mathrm{E}\left[\left(\frac{1}{\kappa}\left(\underline{x}_1 + \cdots + \underline{x}_\kappa\right) - \frac{1}{\nu}\left(\underline{x}_0 + \underline{x}_{-1} + \cdots + \underline{x}_{-\nu+1}\right)\right)^2\right] \\ &= \mathrm{E}\left[\left(-\frac{\underline{x}_1}{\nu} - \cdots - \frac{\underline{x}_\nu}{\nu} + \frac{\underline{x}_{\nu+1}}{\kappa} + \cdots + \frac{\underline{x}_{\nu+\kappa}}{\kappa}\right)^2\right] \end{aligned} \tag{4.66}$$

As demonstrated in Appendix 4-I, this is expressed in terms of the climacogram as:

$$A(\kappa,\nu) = \left(\frac{1}{\kappa} + \frac{1}{\nu}\right)\left(\kappa\,\gamma(\kappa) + \nu\,\gamma(\nu) - (\nu+\kappa)\,\gamma(\nu+\kappa)\right) \tag{4.67}$$

We will discuss now a few examples. First is the Hurst-Kolmogorov process, for which $\gamma(\kappa) = \lambda(\kappa/\alpha)^{2H-2}$. As explained in Appendix 4-I, the value of ν that minimizes A is:

$$\nu = \frac{\kappa}{(\max(0, 2.5H - 1.5))^{2.5}} \tag{4.68}$$

If $H \leq 0.6$, this yields $\nu = \infty$, which means that the future mean estimate is the average of the entire set of n observations, the global mean. However, if $H > 0.6$, then it can be $\nu < n$ and, hence, we should use a local mean with fewer terms than n to estimate the future mean. As $H \to 1$, $\nu \to 1$, too. A graphical illustration of equation (4.68) is given in Figure 4.5 (left).

We recall, though, that the Hurst-Kolmogorov process entails infinite instantaneous variance and thus it is not an ideal model for real-world processes. The second example is a filtered Hurst-Kolmogorov process in its simplest Cauchy form (FHK-C) with $M = 1/2$, i.e., $\gamma(\kappa) = \lambda(1 + \kappa/\alpha)^{2H-2}$. This has finite instantaneous variance, equal to λ. Studying this process and in particular considering the specific values $A(\kappa, 1)$, $A(\kappa, 2)$, as given by (4.67), and $A(\kappa, \infty) = \gamma(\kappa)$, we will see that there are cases where:

$$A(\kappa, 1) \leq \min\big(A(\kappa, 2), A(\kappa, \infty)\big) \tag{4.69}$$

In such cases the resulting optimal ν equals one, which means that only the present value should count for the future mean. A systematic numerical investigation on equation (4.69) suggested that $\nu = 1$ is optimal when:

$$\kappa \leq \kappa_1 \approx 2.3(\alpha + 1)H^2 - 1 \tag{4.70}$$

Combining the above results, we find that an approximate general solution for the above FHK-C model is:

$$\nu \approx \max\left(1, \frac{\kappa - \kappa_1}{(\max(0, 2.5H - 1.5))^{2.5}}\right) = \max\left(1, \frac{\kappa - 2.3(\alpha+1)H^2 - 1}{(\max(0, 2.5H - 1.5))^{2.5}}\right) \tag{4.71}$$

Characteristic results are given graphically in Figure 4.5 (right). It can be seen that the result $\nu = 1$ is not uncommon as it appears for many parameter combinations. More generally, finite values of ν of the order of κ or somewhat larger are common for $\kappa \leq 10$ (for example for $H = 0.75$ and $\alpha = 10$, the optimal ν is 1 for $\kappa = 10$ and increases to $\nu = 20$

for $\kappa = 15$). The case $H = 0.5$ is virtually equivalent to a Markov process. As shown in Figure 4.5 (right) for this case (particularly for $\alpha = 10$), the plot is a vertical line at $\kappa = \kappa_1$ and this means the optimal value is either $\nu = 1$ (for $\kappa \le \kappa_1$) or $\nu = n$ (for $\kappa > \kappa_1$). In order for this to happen, κ_1 must be ≥ 1, which happens when $\alpha \ge 2.5$ (otherwise, $\nu = n$ for any κ).

Note that here we considered the question: Which of the local past averages is most representative as an estimate of the future average? We did not consider weighted averages of past values, even though this could reduce estimation variance. Therefore, the case where the resulting optimal value is $\nu = 1$ does not suggest that the process is a martingale*. This analysis aims to show the differences of global and local time averages and the fact that the latter may provide better prediction for the future in processes with dependence. A detailed study on the subject using real-world (rainfall) data has been made by Iliopoulou and Koutsoyiannis (2020). An illustration using weighted rather than standard averages is given in Digression 4.F.

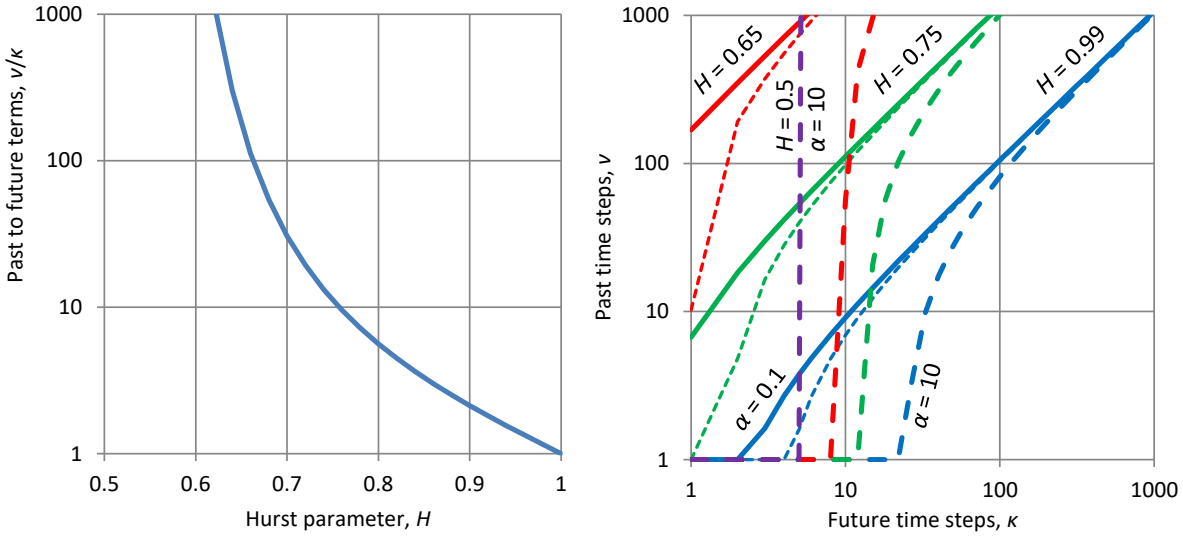

Figure 4.5 (left) Graphical illustration of equation (4.68) assuming a large length of time series n. **(right)** Characteristic curves of optimal ν vs κ for the indicated values of parameters H and α as found by numerical analysis (and approximated by equation (4.71)); continuous, dotted and dashed lines correspond to $\alpha = 0.1$, 1 and 10, respectively (for the case $H = 0.5$ the curves for $\alpha = 0.1$ and 1 fall out of the graph and therefore only that for $\alpha = 10$ is shown, which is a vertical straight line).

Digression 4.F: Forecast-oriented estimation using weights

If we use weights b_i to estimate a weighted average of past values which will be representative of the future average, then equation (4.66) is replaced by:

$$A(\boldsymbol{b}, \kappa, \nu) = \mathrm{E}\left[\left(\frac{1}{\kappa}(\underline{x}_1 + \cdots + \underline{x}_\kappa) - (b_1\underline{x}_0 + b_2\underline{x}_{-1} + \cdots + b_\nu\underline{x}_{-\nu+1})\right)^2\right] \tag{4.72}$$

Assuming that $b_1 + \cdots + b_\nu = 1$, and setting $\underline{y} \coloneqq (\underline{x}_1 + \cdots + \underline{x}_\kappa)/\kappa, \underline{w}_i \coloneqq \underline{y} - \underline{x}_i$, we can write:

* A martingale is a stochastic process in which $\mathrm{E}[\underline{x}_1 | \underline{x}_0, \underline{x}_{-1}, \ldots] = \underline{x}_0$.

$$A(\boldsymbol{b},\kappa,\nu)=\mathrm{E}\left[\left(b_1(\underline{y}-\underline{x}_1)+\cdots+b_\nu(\underline{y}-\underline{x}_\nu)\right)^2\right] \tag{4.73}$$

or in vector form:

$$A(\boldsymbol{b},\kappa,\nu)=\mathrm{E}\left[\left(\boldsymbol{b}^{\mathrm{T}}\underline{\boldsymbol{w}}\right)^2\right]=\mathrm{E}\left[\boldsymbol{b}^{\mathrm{T}}\underline{\boldsymbol{w}}\underline{\boldsymbol{w}}^{\mathrm{T}}\boldsymbol{b}\right]=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{b} \tag{4.74}$$

where $\boldsymbol{b}\coloneqq[b_1,\dots,b_\nu]^{\mathrm{T}}$, $\underline{\boldsymbol{w}}\coloneqq\left[\underline{w}_1,\dots,\underline{w}_\nu\right]^{\mathrm{T}}$ and $\boldsymbol{C}\coloneqq\mathrm{E}\left[\underline{\boldsymbol{w}}\boldsymbol{w}^{\mathrm{T}}\right]$. The elements of $\boldsymbol{C}$ are:

$$C_{ij}=\mathrm{E}\left[\underline{w}_i\underline{w}_j\right]=\mathrm{E}\left[\left(\underline{y}-\underline{x}_i\right)\left(\underline{y}-\underline{x}_j\right)\right]=\mathrm{E}\left[\underline{y}^2\right]-\mathrm{E}\left[\underline{x}_i\underline{y}\right]-\mathrm{E}\left[\underline{x}_j\underline{y}\right]+\mathrm{E}\left[\underline{x}_i\underline{x}_j\right] \tag{4.75}$$

The first term of the sum (assuming zero mean, without loss of generality) is $\mathrm{E}\left[\underline{y}^2\right]=\gamma(\kappa)=\Gamma(\kappa)/\kappa^2$. The last term is:

$$\mathrm{E}\left[\underline{x}_i\underline{x}_j\right]=c_{[i-j|}=\frac{1}{2}\left(\Gamma(|i-j+1|)+\Gamma(|i-j-1|)-2\Gamma(|i-j|)\right) \tag{4.76}$$

To find the middle terms of the sum we observe that $\underline{X}_{\kappa-1}=\underline{X}_\kappa-\underline{x}_\kappa$ and hence:

$$\mathrm{E}\left[\underline{X}_\kappa\underline{x}_\kappa\right]=\frac{\mathrm{E}\left[\underline{X}_\kappa^2\right]+\mathrm{E}\left[\underline{x}_\kappa^2\right]-\mathrm{E}\left[\underline{X}_{\kappa-1}^2\right]}{2}=\frac{\Gamma(\kappa)-\Gamma(\kappa-1)+\Gamma(1)}{2} \tag{4.77}$$

while because of symmetry $\mathrm{E}\left[\underline{X}_\kappa\underline{x}_\kappa\right]=\mathrm{E}\left[\underline{X}_\kappa\underline{x}_1\right]$. Thus,

$$\begin{aligned}\mathrm{E}\left[\underline{x}_i\underline{y}\right]&=\frac{1}{\kappa}\mathrm{E}\left[\underline{x}_i\left(\underline{X}_{\nu+\kappa}-\underline{X}_\nu\right)\right]=\frac{1}{\kappa}\mathrm{E}\left[\underline{x}_1\left(\underline{X}_{\nu+\kappa-i+1}-\underline{X}_{\nu-i+1}\right)\right]\\&=\frac{1}{2\kappa}\left(\Gamma(\nu+\kappa-i+1)-\Gamma(\nu+\kappa-i)-\Gamma(\nu-i+1)+\Gamma(\nu-i)\right)\end{aligned} \tag{4.78}$$

Once we have calculated the elements of the symmetric matrix $\boldsymbol{C}$ as above, what it remains is to:

$$\text{minimize } A'(\boldsymbol{b},\kappa,\nu)\coloneqq A(\boldsymbol{b},\kappa,\nu)+\lambda\left(\mathbf{1}^{\mathrm{T}}\boldsymbol{b}-1\right)=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{b}+\lambda\left(\mathbf{1}^{\mathrm{T}}\boldsymbol{b}-1\right) \tag{4.79}$$

The solution is given by:

$$\frac{\partial A'(\boldsymbol{b},\kappa,\nu)}{\partial\boldsymbol{b}}=2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C}+\lambda\mathbf{1}^{\mathrm{T}}=\mathbf{0},\qquad\frac{\partial A'(\boldsymbol{b},\kappa,\nu)}{\partial\lambda}=\mathbf{1}^{\mathrm{T}}\boldsymbol{b}-1=0 \tag{4.80}$$

where $\mathbf{0}$ and $\mathbf{1}$ are vectors with all elements equal to zero and one, respectively. This solution can be written in a concise form as:

$$\boldsymbol{b}'=\boldsymbol{C}'^{-1}\boldsymbol{d} \tag{4.81}$$

where:

$$\boldsymbol{b}'\coloneqq\begin{bmatrix}\boldsymbol{b}\\\lambda\end{bmatrix},\qquad\boldsymbol{d}\coloneqq\begin{bmatrix}\mathbf{0}\\1\end{bmatrix},\qquad\boldsymbol{C}'\coloneqq\begin{bmatrix}\boldsymbol{C}&\mathbf{1}\\\mathbf{1}^{\mathrm{T}}&0\end{bmatrix} \tag{4.82}$$

Apparently, this case is more complicated than the one studied in section 5.6 and no analytical, exact or approximate, relationship can be formulated. However, it is interesting to see a numerical illustration, such as that depicted in Figure 4.6.

This figure allows to make the following observations:

- When there is no persistence (H = 0.5) and almost no dependence (α = 0.1), the weights are almost equal, $b_i\approx 1/100=0.01$. This is equivalent to choosing the global average of the past for inferring the future average.
- When α is small (α = 0.1), i.e. the behaviour is close to the standard HK process, then the weights form a curve with almost constant slope, with higher H corresponding to steeper negative slope.
- The weight of the first term (the present) is highest.

- The weight of the last term (x_{-99}) is higher than the adjacent ones. The explanation is that it represents the unknown (and thus neglected) terms beyond x_{-99}.
- The sequences with a large time scale parameter α have a negative weight for $\tau = 1$. This means that that the model takes into account the latest "trend", in addition to the latest value.

To explain the last point, let us examine the case of the prediction based on the present value x_0 and a single past value x_{-1}. If we used the "global" mean for the prediction, we would have $x_1 = (x_0 + x_{-1})/2$. If we used just the "trend", then we would have $x_1 = 2x_0 - x_{-1}$. If we took the mean of the two, we would have $x_1 = 1.25x_0 - 0.25x_{-1}$. In the last two cases the weight of the term x_{-1} is negative.

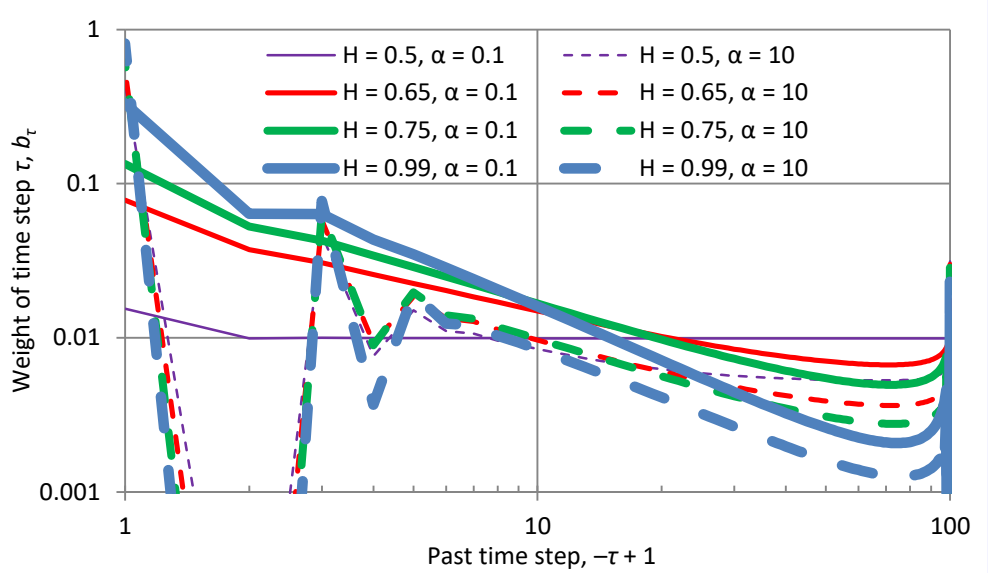

Figure 4.6 Illustration of the weights of the present and past values of time series of length $\nu = 100$ for the prediction of future mean for length $\kappa = 10$ assuming an FHK-C process with $M = 0.5$ and the indicated values of parameters H and α. Note that for $\alpha = 10$ (indicating high correlation at small lags) and $\tau = -1$, the weights are negative and cannot be shown in the logarithmic plot.

Appendix 4-I: Proof of equations (4.67)-(4.68)

From equation (4.66) we find, in terms of the cumulative process $\underline{X}_i$:

$$\begin{aligned} A(\kappa,\nu) &= \mathrm{E}\left[\left(\frac{1}{\kappa}\left(\underline{X}_{\nu+\kappa}-\underline{X}_{\nu}\right)-\frac{1}{\nu}\underline{X}_{\nu}\right)^2\right] = \mathrm{E}\left[\left(\frac{1}{\kappa}\underline{X}_{\nu+\kappa}-\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)\underline{X}_{\nu}\right)^2\right] \\ &= \frac{1}{\kappa^2}\mathrm{E}\left[\underline{X}_{\nu+\kappa}^2\right]+\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)^2\mathrm{E}\left[\underline{X}_{\nu}^2\right]-\frac{2}{\kappa}\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)\mathrm{E}\left[\underline{X}_{\nu+\kappa}\underline{X}_{\nu}\right] \end{aligned} \tag{4.83}$$

On the other hand, we have:

$$\mathrm{E}\left[\underline{X}_{\kappa}^2\right] = \mathrm{E}\left[\left(\underline{X}_{\nu+\kappa}-\underline{X}_{\nu}\right)^2\right] = \mathrm{E}\left[\left(\frac{1}{\kappa}\underline{X}_{\nu+\kappa}-\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)\underline{X}_{\nu}\right)^2\right] = \mathrm{E}\left[\underline{X}_{\nu+\kappa}^2\right]+\mathrm{E}\left[\underline{X}_{\nu}^2\right]-2\mathrm{E}\left[\underline{X}_{\nu+\kappa}\underline{X}_{\nu}\right] \tag{4.84}$$

Thus:

$$\begin{aligned} A(\kappa,\nu) &= \frac{1}{\kappa^2}\mathrm{E}\left[\underline{X}_{\nu+\kappa}^2\right]+\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)^2\mathrm{E}\left[\underline{X}_{\nu}^2\right]+\frac{1}{\kappa}\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)\left(\mathrm{E}\left[\underline{X}_{\kappa}^2\right]-\mathrm{E}\left[\underline{X}_{\nu+\kappa}^2\right]-\mathrm{E}\left[\underline{X}_{\nu}^2\right]\right) \\ &= \frac{1}{\kappa}\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)\mathrm{E}\left[\underline{X}_{\kappa}^2\right]+\frac{1}{\nu}\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)\mathrm{E}\left[\underline{X}_{\nu}^2\right]-\frac{1}{\kappa\nu}\mathrm{E}\left[\underline{X}_{\nu+\kappa}^2\right] \end{aligned} \tag{4.85}$$

Hence:

$$A(\kappa,\nu)=\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)\left(\frac{\Gamma(\kappa)}{\kappa}+\frac{\Gamma(\nu)}{\nu}-\frac{\Gamma(\nu+\kappa)}{\nu+\kappa}\right) \tag{4.86}$$

which can be written in the form of (4.67).

For an HK process:

$$\frac{A(\kappa,\nu)}{\lambda\, a^{2-2H}}=\left(\frac{1}{\kappa}+\frac{1}{\nu}\right)(\kappa^{2H-1}+\nu^{2H-1}-(\nu+\kappa)^{2H-1}) \tag{4.87}$$

Assuming that κ and H are specified, we minimize the quantity:

$$\frac{A(\kappa,\nu)\kappa^{1-2H}}{\lambda\, a^{2-2H}}=\left(\frac{\nu}{\kappa}\right)^{-1}\left(\frac{\nu}{\kappa}+1\right)\left(1+\left(\frac{\nu}{\kappa}\right)^{2H-1}-\left(\frac{\nu}{\kappa}+1\right)^{2H-1}\right) \tag{4.88}$$

which is a function if ν/κ. For approximation we assume that ν/κ is a real number and we take the derivative with respect to it, which we equate to zero to find the value that minimizes A. This can be solved only numerically. By performing a systematic numerical investigation (finding optimal values of ν/κ for different H) we are able to fit equation (4.68) on the results with a small error.

Chapter 5. Return period

5.1 Definitions and insights on return period

We have already introduced the concept of the return period T in section 1.5, where we have seen that it is inversely proportional to the probability P_1 that a dangerous event A would occur in a time unit D. Although this relationship (equation (1.7)) is almost obvious, here we will approach it again in a rigorous manner and examine several variants of it.

First, we define the concept. For a specific event A, which is a subset of a basic set Ω, we define the *return period*, T, as the mean time between consecutive occurrences of the event A. This is a standard term in engineering applications (in engineering hydrology in particular) but needs some clarification to avoid common misuses and frequent confusion.

We will initially consider the discrete time version and we will later see how we can reformulate it in continuous time. Let B be the complementary event of A ($B \coloneqq \Omega - A$). We denote:

$$P_1 \coloneqq P(A), \qquad P_{11} \coloneqq P(A_0, A_1) \tag{5.1}$$

We examine sequences of events B, possibly in contact with events A, and in particular the following sequences and their corresponding probabilities:

$$\begin{aligned} P_B(n) &\coloneqq P(B_1, B_2, \dots\ B_n) \\ P_{B|}(n) &\coloneqq P(B_1, B_2, \dots\ B_n, A_{n+1}) \\ P_{|B}(n) &\coloneqq P(A_1, B_2, \dots\ B_n, B_{n+1}) \\ P_{|B|}(n) &\coloneqq P(A_0, B_1, B_2, \dots\ B_n, A_{n+1}) \end{aligned} \tag{5.1}$$

Since:

$$\begin{aligned} P_{|B}(n) &= P(A_1, B_2, \dots\ B_n, B_{n+1}) = P(B_2, B_3, \dots, B_{n+1}) - P(B_1, B_2, \dots, B_{n+1}) \\ P_{B|}(n) &= P(B_1, B_2, \dots\ B_n, A_{n+1}) = P(B_1, B_2, \dots, B_n) - P(B_1, B_2, \dots, B_{n+1}) \end{aligned} \tag{5.2}$$

while, by virtue of stationarity, $P(B_2, B_3, \dots\ B_{n+1}) = P(B_1, B_2, \dots, B_n)$, we conclude that:

$$P_{|B}(n) = P_{B|}(n) \tag{5.3}$$

Special cases of the probabilities of the above event sequences are:

$$\begin{aligned} P_B(0) = 1, \qquad P_B(1) = 1 - P_1, \qquad P_B(2) = 1 - 2P_1 + P_{11} \\ P_{B|}(0) = P_1, \qquad p_{B|}(1) = P_1 - P_{11} \\ p_{|B|}(0) = P_{11} \end{aligned} \tag{5.4}$$

It is easy to see that:

$$\begin{aligned} P_{B|}(n) &= P_B(n) - P_B(n+1) \\ P_{|B|}(n) &= P_{B|}(n) - P_{B|}(n+1) = P_B(n) - 2P_B(n+1) + P_B(n+2) \end{aligned} \tag{5.5}$$

Assuming that the event A has happened at time 0, if its next occurrence is at time n, we can easily derive that:

$$P\{\underline{n} = n\} = P(B_1, B_2, \dots, B_{n-1}, A_n | A_0) = \frac{P(A_0, B_1, B_2, \dots, B_{n-1}, A_n)}{P(A_0)} = \frac{P_{|B|}(n-1)}{P_1} \quad (5.6)$$

The expected value of $\underline{n}$ is:

$$\mathrm{E}[\underline{n}] = \sum_{n=0}^{\infty} n\, P\{\underline{n} = n\} = \frac{1}{P_1} \sum_{n=0}^{\infty} n\, P_{|B|}(n-1) \quad (5.7)$$

where using (5.5) we easily see that the sum is evaluated to 1, and thus:

$$\mathrm{E}[\underline{n}] = \frac{T}{D} = \frac{1}{P_1} \quad (5.8)$$

This is the standard relationship between return period and probability. The proof we have given has not assumed anything but stationarity, so it is quite generic. In particular, there is no assumption of independence or any particular type of dependence (see also Koutsoyiannis, 2008; Volpi et al. 2015). We stress that, according to the definition, the return period is the mean time between consecutive occurrences of the event A. We have assumed for the above proof that the event A has happened at time 0 (present time).

If we do not have any information about the present time, we can derive the mean time to the first occurrence of the event A unconditionally, denoted as $T_\mathrm{u} := D\mathrm{E}[\underline{n}_\mathrm{u}]$, where $\underline{n}_\mathrm{u}$ is the number of time intervals of length D until the next occurrence of A. Its probability mass is:

$$P\{\underline{n}_\mathrm{u} = n\} = P(B_1, B_2, \dots, B_{n-1}, A_n) = P_{B|}(n-1) = P_B(n-1) - P_B(n) \quad (5.9)$$

Thus, its expected value is:

$$\mathrm{E}[\underline{n}_\mathrm{u}] = \sum_{n=0}^{\infty} n \left(P_B(n-1) - P_B(n)\right) \quad (5.10)$$

and simplifies to:

$$\mathrm{E}[\underline{n}_\mathrm{u}] = \frac{T_\mathrm{u}}{D} = \sum_{n=0}^{\infty} P_B(n) \quad (5.11)$$

In case of independence, $P_B(n) = P_B(1)^n = (1 - P_1)^n$ and thus:

$$\mathrm{E}[\underline{n}_\mathrm{u}] = \frac{T_\mathrm{u}}{D} = \frac{1}{P_1} \quad (5.12)$$

or $T_\mathrm{u} \equiv T$. If the succession of events A and B is modelled as a Markov chain, from Koutsoyiannis (2006a, equation (2)) we find:

$$P_B(n) = P_B(1) \left(\frac{P_B(2)}{P_B(1)}\right)^{n-1} \quad (5.13)$$

and hence:

$$\mathrm{E}[\underline{n}_{\mathrm{u}}] = \frac{T_{\mathrm{u}}}{D} = \frac{\big(P_B(1)\big)^2}{P_B(2)\big(P_B(1) - P_B(2)\big)} = \frac{(1-P_1)^3}{(P_1 - P_{11})(1 - 2P_1 + P_{11})} \tag{5.14}$$

The difference from T/D, after algebraic manipulations, is:

$$\mathrm{E}[\underline{n}_{\mathrm{u}}] - \mathrm{E}[\underline{n}] = \frac{(P_{11} - P_1^2)(1 - 3P_1 + P_1^2 + P_{11})}{P_1(P_1 - P_{11})(1 - 2P_1 + P_{11})} \tag{5.15}$$

For $P_1 \leq 0.5$ (an obvious condition to characterize the event A as dangerous) and $P_1^2 \leq P_{11} \leq P_1$ (meaning positive dependence), it can be verified that $1 - 3P_1 + P_1^2 + P_{11} > 0$. Thus $\mathrm{E}[\underline{n}_{\mathrm{u}}] \geq \mathrm{E}[\underline{n}]$ or $T_{\mathrm{u}} \geq T$. For other types of dependence, assuming $B \coloneqq \{\underline{x} \leq x\}$, so that $P(B) = F(x)$ and $P_B(n) = P\{\max(\underline{x}_1, \dots, \underline{x}_n) \leq x\}$, we can also evaluate $P_B(n)$ from the properties of the stochastic process $\underline{x}_i$ and then $\mathrm{E}[\underline{n}_{\mathrm{u}}]$ from equation (5.11). But this can be laborious.

Now let us make the calculations again on the condition that the latest occurrence of A has been observed at time $-m$ in the past. In this case we have the conditional return period $T_m \coloneqq D\mathrm{E}[\underline{n}|m]$, which can be determined from the conditional probability:

$$\begin{aligned} P\{\underline{n} = n|m\} &= P(B_1, B_2, \dots, B_{n-1}, A_n | B_0, B_{-1} \dots, B_{-m+1}, A_{-m}) \\ &= \frac{P(A_{-m}, B_{-m+1}, \dots B_0, B_1, \dots, B_{n-1}, A_n)}{P\big((A_{-m}, B_{-m+1}, \dots B_0)\big)} = \frac{P_{|B|}(m+n-1)}{P_{B|}(m)} \end{aligned} \tag{5.16}$$

The required expected value is:

$$\mathrm{E}[\underline{n}|m] = \sum_{n=0}^{\infty} n\, P\{\underline{n} = n|m\} = \frac{1}{P_{B|}(m)} \sum_{n=0}^{\infty} n\, P_{|B|}(m+n-1) \tag{5.17}$$

which by virtue of equation (5.5) is written as:

$$\mathrm{E}[\underline{n}|m] = \frac{1}{P_B(m) - P_B(m+1)} \sum_{n=0}^{\infty} n\,(P_B(m+n-1) - 2P_B(m+n) + P_B(m+n+1))$$

and reduces to:

$$\mathrm{E}[\underline{n}|m] = \frac{P_B(m)}{P_B(m) - P_B(m+1)} \tag{5.18}$$

In case of independence, $P_B(m) = P_B(1)^m = (1-P_1)^m$ and hence:

$$\mathrm{E}[\underline{n}|m] = \frac{T_m}{D} = \frac{(1-P_1)^m}{(1-P_1)^m - (1-P_1)^{m+1}} = \frac{1}{P_1} \tag{5.19}$$

or $T_m \equiv T$, which is expected because in an independent sequence of events conditioning on the past does not change anything. In the case of a Markov chain, using (5.13) we obtain:

$$\mathrm{E}[\underline{n}|m] = \frac{T_m}{D} = \frac{P_B(1)}{P_B(1) - P_B(2)} = \frac{1 - P_1}{P_1 - P_{11}} \tag{5.20}$$

This is independent of m because, due to the Markov property, the past is irrelevant once the present is known. The difference from $\mathrm{E}[\underline{n}_\mathrm{u}]$, after algebraic manipulations, is:

$$\mathrm{E}[\underline{n}|m] - \mathrm{E}[\underline{n}_\mathrm{u}] = \frac{(1-P_1)(P_{11}-P_1^2)}{(P_1-P_{11})(1-2P_1+P_{11})} \tag{5.21}$$

It can be verified that when $P_1 \le 0.5, P_1^2 \le P_{11} \le P_1$ (as above), the difference is positive. Thus $\mathrm{E}[\underline{n}|m] \ge \mathrm{E}[\underline{n}_\mathrm{u}] \ge \mathrm{E}[\underline{n}]$ or $T_m \ge T_\mathrm{u} \ge T$. For $m = 0$, $\mathrm{E}[\underline{n}|0] = \mathrm{E}[\underline{n}]$, as expected. However, for $m \ge 1$ the difference can be substantial, as seen in an illustration in Figure 5.1, and tends to infinity as $P_{11} \to P_1$.

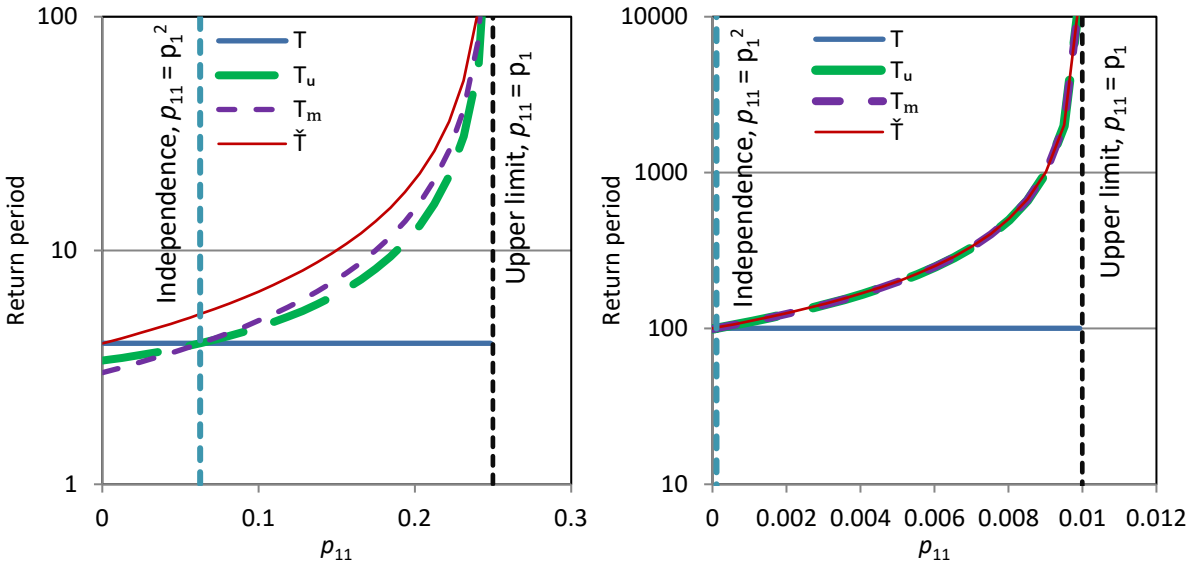

Figure 5.1 Illustration of the different variants of return period for probability of a dangerous event A **(left)** $P_1 = 0.25$ and **(right)** $P_1 = 0.01$ as a function of the probability if two consecutive dangerous events P_{11}. For T_u and T_m a Markov chain model is assumed.

Let us now take a further step, which is quite important. We first assume that the basic time step D, which we used to define the dangerous event A, for example a mean river discharge of 1000 m^3/s, is a day. Up to now, we have tacitly assumed that if the event A, occurs on two consecutive days, this constitutes two occurrences of A. But let us think that we radically reduce the time unit to, say, D = 1 min. The theory should apply also in this case. Should the occurrence of mean discharge of 1000 m^3/s for two consecutive minutes be regarded as two occurrences of a dangerous event? And what about if our time step is D = 1 s or 1 μs? A reasonable answer would be that the continuation of a dangerous event for a number of consecutive steps should not be regarded an occurrence of a new event.

This requires some modified analysis. According to this consideration, in the sequence $B_1, B_2, B_3, A_4, B_5, B_6, A_7, A_8, A_9, B_{10}$ we have two occurrences of A (A_4 and A_7-A_8-A_9) rather than four. Assuming a long period of length $L = lD$, in which we have n occurrences of A, we expect that $n = P_1 l$. A number of them will be continuations of the previous occurrences, namely, $n_1 = P_{11} l$. Thus, we may write:

$$\check{T} = \frac{L}{n - n_1} = \frac{D}{P_1 - P_{11}} \tag{5.22}$$

We can prove that this result is accurate using formal probability theory. To this aim in a sequence of events A and B we replace all but the first A with B, thus forming a sequence of modified events $\check{A}$ and $\check{B}$.For example the sequence given in the previous paragraph becomes $\check{B}_1, \check{B}_2, \check{B}_3, \check{A}_4, \check{B}_5, \check{B}_6, \check{A}_7, \check{B}_8, \check{B}_9, \check{B}_{10}$. Now if we define $\check{P}_1 \coloneqq P(\check{A})$, $\check{P}_{11} \coloneqq P(\check{A}_0, \check{A}_1)$, we readily find that their values are:

$$\check{P}_1 \coloneqq P(\check{A}) = P_1 - P_{11}, \qquad \check{P}_{11} \coloneqq P(\check{A}_0, \check{A}_1) = 0 \tag{5.23}$$

We can apply all previous results replacing P_1 and P_{11} with $\check{P}_1$ and $\check{P}_{11}$ and find the respective quantities for the sequence of modified events $\check{A}$ and $\check{B}$. In particular we find that the return period is:

$$\mathrm{E}[\underline{\check{n}}] = \frac{\check{T}}{D} = \frac{1}{\check{P}_1} = \frac{1}{P_1 - P_{11}} \tag{5.24}$$

Comparing with the previous results we infer that for independent processes and Markov chains, $\check{T} \geq T_m \geq T_{\mathrm{u}} \geq T$. In particular, the inequality $\check{T} \geq T$ is obviously valid for any process.

The sequence of modified events $\check{A}$ and $\check{B}$ cannot be independent, because the occurrence of $\check{A}_i$ excludes the occurrence of $\check{A}_{i+1}$; this signifies negative dependence. However, this sequence can be a Markov chain but now, because of the negative dependence, the conditional and unconditional return periods will be smaller ($\check{T}_m \leq \check{T}$). For example, it can be shown (homework) that for $m > 0$, $\check{T}_m = \check{T} - D$. This allows us to conjecture that $\check{T}$ represents an upper bound of all variants or return period, conditional and unconditional. If the sequence of events A and B is a Markov chain, then that of $\check{A}$ and $\check{B}$ is not and therefore the conditional $\check{T}_m$ should depend on m. Its analytical determination can be laborious. Nonetheless, stochastic simulation can readily provide its behaviour. A couple of examples is depicted in Figure 5.2, where it can be seen that $\check{T}_0 = \check{T}$, while $\check{T}_m$ tends to T_m as m increases.

Extreme events that are of interest in geophysics (and, in particular, hydroclimatic processes) are usually of two types, highs (e.g. storms and floods) or lows (e.g. droughts). In the former case the dangerous event is the exceedance of a certain threshold value x, usually related to a failure of a system, operational or structural. In this case the dangerous and non-dangerous events are defined as $A_i \coloneqq \{\underline{x}_i > x\}, B_i \coloneqq \{\underline{x}_i \leq x\}$, respectively, where $\underline{x}_i$ is a stochastic process quantifying a natural process (e.g. river discharge).

It is interesting to observe that, by virtue of (5.2) (or (5.3)):

$$P(A_1, B_2) = P(B_1, A_2) \tag{5.25}$$

and since

$$P(A_1, B_2) = P(A_1) - P(A_1, A_2), \qquad P(B_1, A_2) = P(B_1) - P(B_1, B_2) \tag{5.26}$$

we obtain

$$P(A_1) - P(A_1, A_2) = P(B_1) - P(B_1, B_2) \tag{5.27}$$

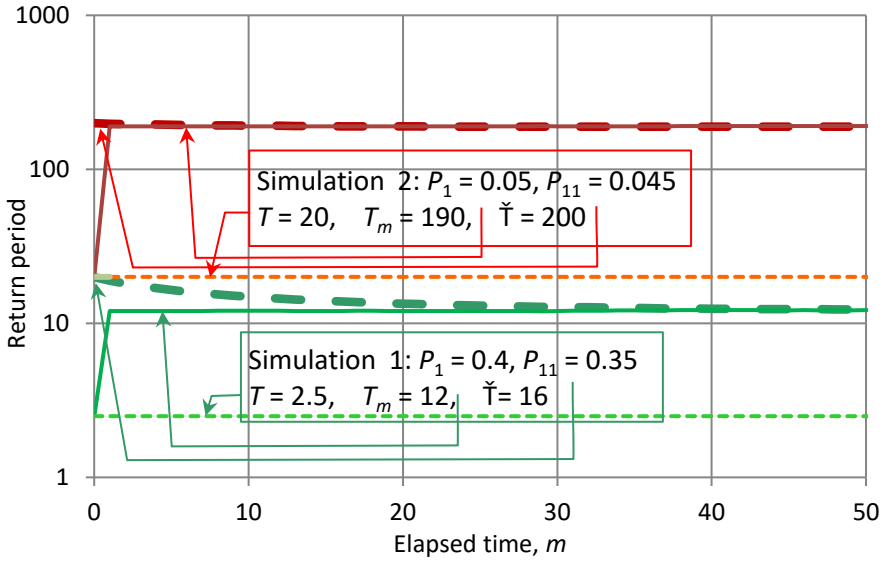

Figure 5.2 Simulation results for conditional return periods T_m and $\check{T}_m$ for two Markov chains with the indicated characteristics.

We can use the last property to express the return periods in terms of the distribution function of the process $\underline{x}_i$. In the case that extremes are maxima, we have $P_1 = P(A_i) = P\{\underline{x}_i > x\} = \bar{F}(x)$ and $P(B_i) = P\{\underline{x}_i \geq x\} = F(x)$. Likewise, $P_{11} = P(A_1, A_2) = P\{\underline{x}_1 > x, \underline{x}_2 > x\} = \bar{F}_z(x)$, where $\underline{z} := \min(\underline{x}_1, \underline{x}_2)$, and $P(B_1, B_2) = P\{\underline{x}_1 \leq x, \underline{x}_2 \leq x\} = F_y(x)$, where $\underline{y} := \max(\underline{x}_1, \underline{x}_2)$. Hence:

$$\frac{T}{D} = \frac{1}{P_1} = \frac{1}{\bar{F}(x)} = \frac{1}{1 - F(x)}, \qquad \frac{\check{T}}{D} = \frac{1}{P_1 - P_{11}} = \frac{1}{\bar{F}(x) - \bar{F}_z(x)} = \frac{1}{F(x) - F_y(x)} \tag{5.28}$$

where we have taken advantage of the symmetry relationship (5.27).

In the case that extremes are minima, working as above but interchanging the definitions of A_i and B_i and denoting the return period as $\overline{T}$ to distinguish it from T, we find:

$$\frac{\overline{T}}{D} = \frac{1}{P_1} = \frac{1}{F(x)}, \qquad \frac{\check{T}}{D} = \frac{1}{P_1 - P_{11}} = \frac{1}{F(x) - F_y(x)} \tag{5.29}$$

The identical expression of $\check{T}$ for maxima and minima is remarkable.

In the case of an independent process, $F_y(x) = \big(F(x)\big)^2$ and thus:

$$\frac{\check{T}}{D} = \frac{1}{F(x)\big(1 - F(x)\big)} = \frac{1}{F(x)\bar{F}(x)} \tag{5.30}$$

both for maxima and minima. Furthermore, it can be readily seen that

$$\frac{\check{T}}{D} = \frac{T}{D} + \frac{\overline{T}}{D} = \frac{T}{D}\frac{\overline{T}}{D} \tag{5.31}$$

Interestingly, the first term equals both the sum and the product of the two other terms.

5.2 Most useful variants of return period

Recapitulating the results of section 5.1, we distinguish the two most important concepts that will be used further: the *return period* T (for maxima) or $\overline{T}$ (for minima), and the *distinct return period* $\breve{T}$. Their properties, advantages and disadvantages are summarized in Table 5.1.

Table 5.1 Properties of the two most important variants of return period.

Property	Return period T or $\overline{T}$	Distinct return period $\breve{T}$
Definition	Mean time between consecutive exceedances (for maxima) or non-exceedances (for minima) of a threshold value x.	Mean time between consecutive distinct exceedances or non-exceedances of a threshold value x, (after an interruption with an opposite event).
Equation to derive	For maxima: $\frac{T}{D} = \frac{1}{\bar{F}(x)} = \frac{1}{1-F(x)}$ For minima: $\frac{\overline{T}}{D} = \frac{1}{F(x)}$	$\frac{\breve{T}}{D} = \frac{1}{\bar{F}(x) - \bar{F}_z(x)} = \frac{1}{F(x) - F_y(x)}$
Requirements to derive	Marginal distribution, $F(x)$	Order-two distribution $F(x_1, x_2)$ or marginal distributions of x and the maximum of two consecutive x.
Relationship with $F(x)$	One to one correspondence with $F(x)$ but different for maxima and minima	Symmetry with respect to maxima and minima but one-to-two correspondence with $F(x)$.
Discrete vs. continuous time	Can only work in discrete time.	Better behaviour for multiscale description, offering extensibility to continuous time.
Estimation from time series	Easy (and potentially unbiased) estimation from time series.	(Not explored yet)

In addition, we introduce the concept of the excess return period defined as the differences of return period from the time unit, i.e., the quantities:

$$T - D, \qquad \overline{T} - D \tag{5.32}$$

These are particularly useful in probability plots (see Digression 5.A) as both range in $(0, \infty)$ (while both T and $\overline{T}$ take values >D). These quantities have the following properties of symmetry:

$$\frac{T-D}{D} = \frac{F(x)}{\bar{F}(x)}, \qquad \frac{\overline{T}-D}{D} = \frac{\bar{F}(x)}{F(x)}, \qquad (T-D)\left(\overline{T}-D\right) = D^2 \tag{5.33}$$

Digression 5.A: Visualizing probability through return periods

It has been a common practice in geophysics and even more in engineering practice to use the return period T, instead of the distribution function F, for probability plots. Representative examples are given in the left column of Figure 5.3 for several distribution functions. In both panels the return period is on the horizontal axis, which is logarithmic. The distribution quantile is shown on the vertical axis on a linear axis (upper panel) or on a logarithmic axis (lower panel).

The former option is good for light-tailed distributions such as normal and exponential. Heavy-tailed ones are better depicted on the double logarithmic plot, on which the slope on the right equals the (upper-)tail index. If we also wish to visualize the lower-tail index, we should replace the return period, T, with the excess return period, $T - D$. In this case, as shown in the right column of Figure 5.3, the right part of the distribution is not affected, but the left is substantially changed, so that the slope on the left on the log-log plot (lower right panel) equals $1/\zeta$, where ζ is the lower-tail index. The values of the asymptotic slopes, left and right, for the excess return period plots and for several common distributions are shown in Table 5.2.

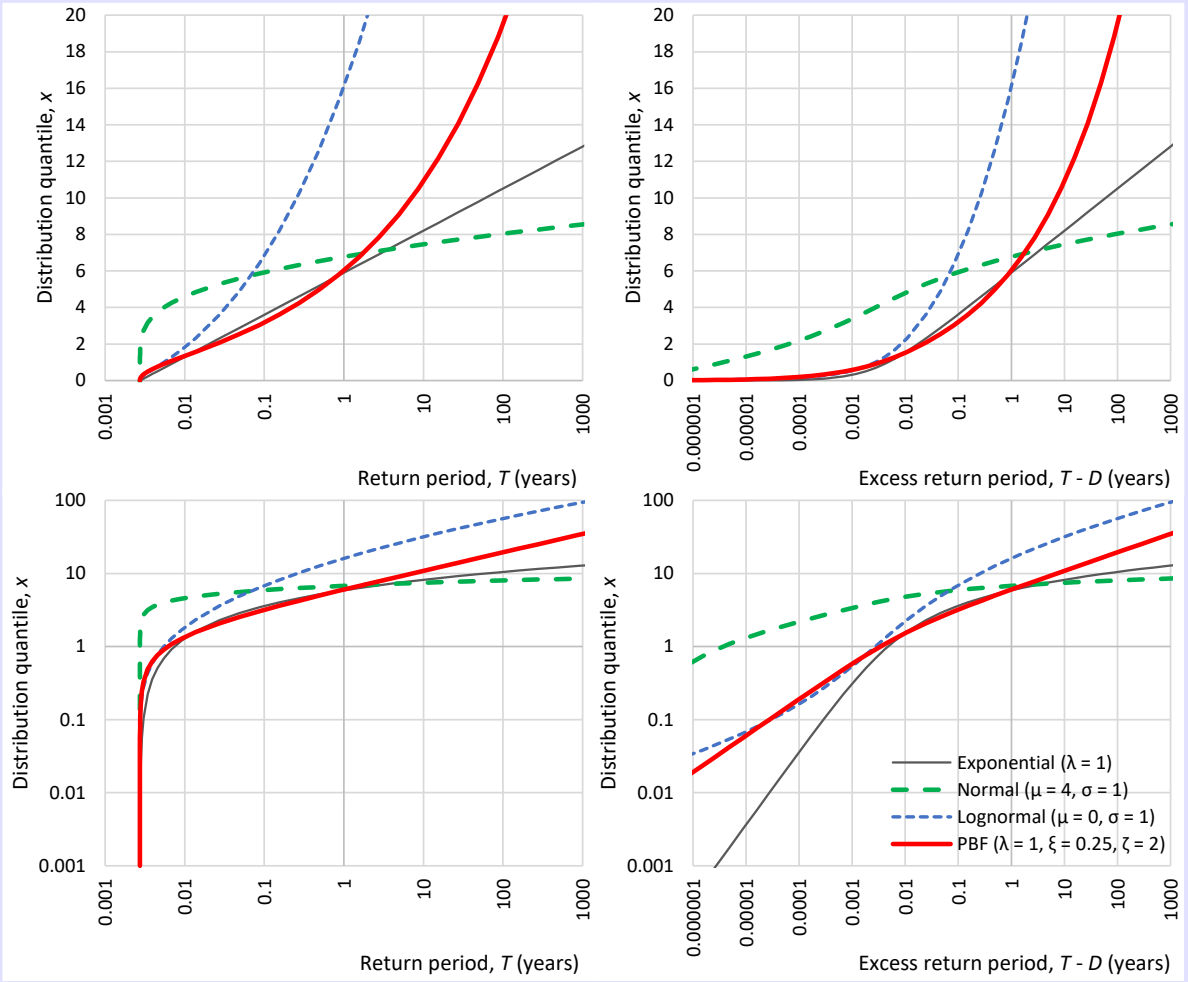

Figure 5.3 Variants of probability plots using the return period **(left column)** or the excess return period **(right column)** instead of the distribution function.

Table 5.2 Asymptotic slopes of the plots of distribution quantiles vs. excess return periods for several common distributions (for the definition of the distributions and their parameters see Table 2.4).

Distribution	Left slope $\lim_{T\to D} \frac{\mathrm{d}x}{\mathrm{d}\ln(T-D)}$	Right slope $\lim_{T\to\infty} \frac{\mathrm{d}x}{\mathrm{d}\ln(T-D)}$	Left slope, log-log $\lim_{T\to D} \frac{\mathrm{d}\ln x}{\mathrm{d}\ln(T-D)}$	Right slope, log-log $\lim_{T\to\infty} \frac{\mathrm{d}\ln x}{\mathrm{d}\ln(T-D)}$
Exponential	0	λ	1	0
Normal	0	0	0	0
Lognormal	0	∞	0	0
Gamma	0	λ	$1/\zeta$	0
Weibull	0	∞	$1/\zeta$	0
Pareto	0	∞	1	ξ
PBF	0	∞	$1/\zeta$	ξ

5.3 Reliability and probability of failure

Among the probabilities of sequences of events that have been introduced and discussed, most important is the probability of zero dangerous events in a period of n time steps, termed *reliability*:

$$P_B(n) \coloneqq P(B_1, B_2, \dots B_n) = P\{\underline{x}_1 \le x, \dots, \underline{x}_n \le x\} = F_{y_n}(x) \tag{5.34}$$

where $\underline{y}_n \coloneqq \max(\underline{x}_1, \dots, \underline{x}_n)$. Its complement from one, i.e.

$$P_{\mathrm{F}}(n) \coloneqq 1 - P_B(n) = \bar{F}_{y_n}(x) \tag{5.35}$$

is called *probability of failure* and is equally important. This is also known in literature as *risk* (e.g. Chow et al., 1988) or *risk of failure* (Serinaldi, 2015). However, here we preferred the more accurate term *probability of failure* instead of risk, because the latter term has acquired a broader meaning, incorporating, in addition to probability, exposure and vulnerability (Kron et al. 2019; see also Chapter 11).

The probability of failure is a concept best suited for design studies and risk assessments. However, it is not easy to handle as it needs the derivation of the any-order distribution of a stochastic process or at least the marginal distributions of maxima of any order—and, as we have seen in Digression 2.J, the convergence of this distribution to the related asymptote is too slow. For that reason, the design and risk assessment studies, as well as the legislation related to management of extreme events (e.g. the European Flood Directive; European Commission, 2007) are more commonly based on the concept of return period whose evaluation only needs the marginal distribution (in the standard variant) or the second-order distribution at most (in the variant of the distinct return period). Common values adopted in engineering design (depending on the importance of the structure and the consequences in case of failure) are shown in Table 5.3.

Table 5.3 Return periods (T) most commonly used in engineering design for high flows and corresponding exceedance probability ($\bar{F}$, equal to the probability of occurrence of a dangerous event P_1), and non-exceedance probability (F). The time unit is assumed D = 1 year.

T (years)	$\bar{F} = P_1$	F	T (years)	$\bar{F} = P_1$	F
2	0.50	0.50	500	0.002	0.998
5	0.20	0.80	1000	0.001	0.999
10	0.10	0.90	5 000	0.0002	0.9998
20	0.05	0.95	10 000	0.00001	0.9999
50	0.02	0.98	50 000	0.000002	0.99998
100	0.01	0.99	100 000	0.000001	0.99999

Note: If the dangerous event is a low one (e.g., low flow, low temperature), we must interchange the columns F and $\bar{F}$ of the table.

5.4 Relationship of probability of failure and return period

It is easy to see that in case that the process of interest is independent in time, the following relationship holds true:

$$1 - P_{\mathrm{F}}(n) = P_B(n) = (1 - P_1)^n \tag{5.36}$$

Thus, if we specify a length n (e.g. nD is the design life span of a project) and the probability of failure $P_{\mathrm{F}}(n)$, then the design return period is

$$\frac{T}{D} = \frac{1}{P_1} = \frac{1}{1 - \left(1 - P_{\mathrm{F}}(n)\right)^{1/n}} \tag{5.37}$$

This is a standard relationship used in hydrological design.

In a process with dependence in time we can modify this relationship by introducing the notion of equivalent length n':

$$\frac{T}{D} = \frac{1}{P_1} = \frac{1}{1 - \left(1 - P_{\mathrm{F}}(n)\right)^{1/n'}} \tag{5.38}$$

where obviously $n' = n$ for the case of independence. Solved for the probability of failure, this yields:

$$P_{\mathrm{F}}(n) = 1 - (1 - P_1)^{n'} = 1 - \left(1 - \frac{D}{T}\right)^{n'} \tag{5.39}$$

For a Markov chain we have

$$P_B(n) = P_B(1)\left(\frac{P_B(2)}{P_B(1)}\right)^{n-1} = \left(P_B(1)\right)^{n'} \tag{5.40}$$

so that

$$n' = 1 + \left(\frac{1}{\zeta} - 1\right)(n - 1), \qquad \zeta := \frac{\ln\left(P_B(1)\right)}{\ln\left(P_B(2)\right)} \tag{5.41}$$

If $\zeta = 1/2$, then we recover $n' = n$, the case of independence.

Koutsoyiannis (2006a), based on maximum entropy considerations, introduced the quasi-Markov (but in essence non-Markov) structure in which:

$$1 - P_{\mathrm{F}}(n) = P_B(n) = (1 - P_1)^{\left(1 + (\zeta^{-1/\theta} - 1)(n-1)\right)^{\theta}} \tag{5.42}$$

In this case:

$$n' = \left(1 + \left(\zeta^{-1/\theta} - 1\right)(n - 1)\right)^{\theta} \tag{5.43}$$

This is a two-parameter relationship with each of the parameters ζ, θ ranging in (0,1). It can readily be seen that the Markov chain is a special case attained when $\theta = 1$. Iliopoulou and Koutsoyiannis (2019), based on a systematic simulation study, have shown that it can

effectively describe an HK process $\underline{x}_i$ for several marginal distributions $F(x)$ and thresholds x.

For processes with positive dependence we will have $n' < n$ and the resulting return period will be smaller than in the case of independence (see Figure 5.4). Thus, if we design by specifying the probability of failure and we neglect the existing dependence of the process, our design will be on the safe side in terms of the resulting return period.

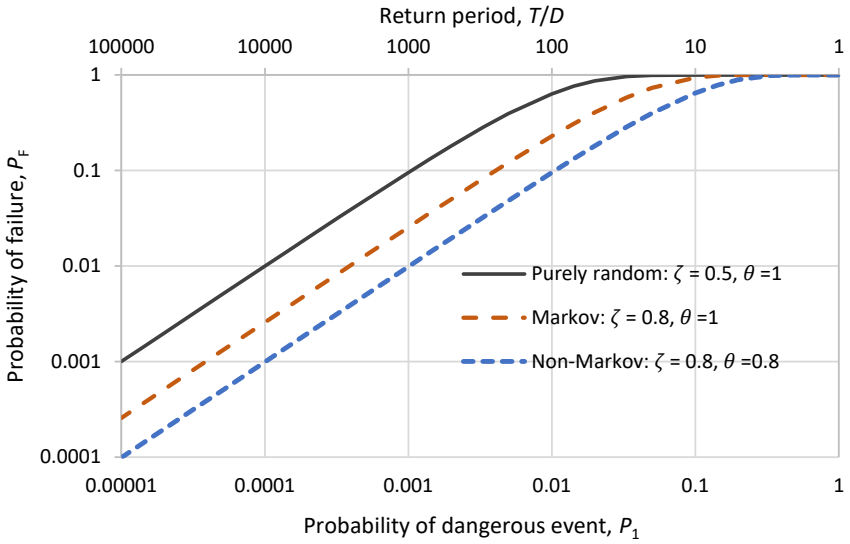

Figure 5.4 Probability of failure as a function of the probability of occurrence of a dangerous event P_1 or the return period T for a period nD where n = 100, and for different stochastic structures of the underlying process.

Digression 5.B: Avoiding misuses of return period

We have insisted in our discussion of the return period on the fact that it is a dimensional quantity with units of time. The notation should be consistent with this fact. It is very common to express the return period in years; it is also common to analyse annual time series, i.e. with D = 1 year. Sometimes dimensionality is forgotten, essentially identifying T with E[$\underline{n}$], or treating the return period as the reciprocal of the exceedance probability. However, this is not dimensionally consistent and not general enough (it does not cover the case of low flows).

Perhaps the word *period* in the term *return period* is not quite proper as it may mislead people to imply that there is some periodic behaviour in consecutive occurrences of events such as in exceedance or non-exceedances of threshold values in nature. In a stochastic process the time between consecutive occurrences of the event is a stochastic variable whose mean is the return period, T. For example, if the value 500 m^3/s of the annual maximum discharge in a river has a return period of 50 years, this does not mean that this value would be exceeded periodically once every 50 years. Rather it means that the average time between consecutive exceedances will be 50 years. An alternative term that has been used to avoid "period" is recurrence interval. However, sometimes (e.g. in Chow et al., 1988) this term has been given the meaning of the stochastic variable $\underline{n}D$ and not its mean T. Also, the notion of the return period should not be thought as a time period of the real world. For example, as seen in Table 5.3, engineers use return periods that in major constructions can be > 10 000 years. One should not compare such return periods with real durations, e.g. with the duration of the Holocene.

Nonetheless, the term *return period* by now has more than 100 years of history (Volpi et al., 2015) and we have kept it also in this text, despite the above caveats.

Digression 5.C: Approximation for the extremes of the normal distribution

The normal distribution is a key model with a great spectrum of applications. Its usefulness stems on the one hand from theoretical reasons (Central Limit Theory, Principle of Maximum Entropy) and on the other hand for the simplicity of its handling, particularly when dealing with multiple dependent variables. Also, the fact that time averaging of a stochastic process with normal distribution preserves the normal distribution, makes the process easiest to handle in discrete time. However, its behaviour that is related to extremes is not easy to handle in terms of precise analytical equations. Therefore, here we derive several approximations, useful for application.

First, in Appendix 5-I we derive the following approximation for $F_{\mathrm{N}}(x)$, the standard normal distribution function, N(0,1):

$$F_{\mathrm{N}}(x) \approx \begin{cases} 1-\frac{1}{2}\exp\left(-\frac{2}{3}x\left(1+\frac{2}{3}x\right)\right), & x \geq 0 \\ \frac{1}{2}\exp\left(\frac{2}{3}x\left(1-\frac{2}{3}x\right)\right), & x \leq 0 \end{cases} \tag{5.44}$$

By inverting it, the quantile function is:

$$x \approx \begin{cases} \frac{3}{4}\left(\sqrt{1-4\ln\left(2(1-F)\right)}-1\right), & F \geq 1/2 \\ -\frac{3}{4}\left(\sqrt{1-4\ln(2F)}-1\right), & x \leq 1/2 \end{cases} \tag{5.45}$$

As seen in Figure 5.5, the approximation is close to accurate. However, it is noted that the approximation is not consistent with the limiting behaviour on the distribution tail. A better approximation for the tail, in particular for $|x| \geq 6$, is given by a well-known approximation:

$$F_{\mathrm{N}}(x) \approx \begin{cases} 1-\frac{1}{\sqrt{2\pi}\,x}\exp\left(-\frac{x^2}{2}\right), & x \geq 6 \\ -\frac{1}{\sqrt{2\pi}\,x}\exp\left(-\frac{x^2}{2}\right), & x \leq -6 \end{cases} \tag{5.46}$$

This is consistent with the distribution tail, which means that:

$$\lim_{x\to\infty}\left(\frac{\bar{F}_{\mathrm{N}}(x)}{\frac{1}{\sqrt{2\pi}\,x}\exp\left(-\frac{x^2}{2}\right)}\right)=1 \tag{5.47}$$

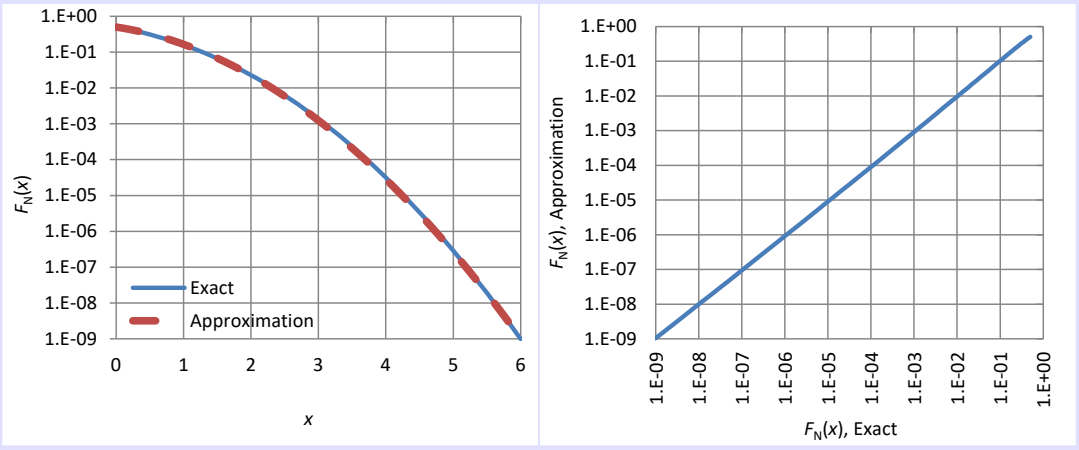

Figure 5.5 Comparison of the approximation of the normal distribution by equation (5.44) with the exact normal distribution.

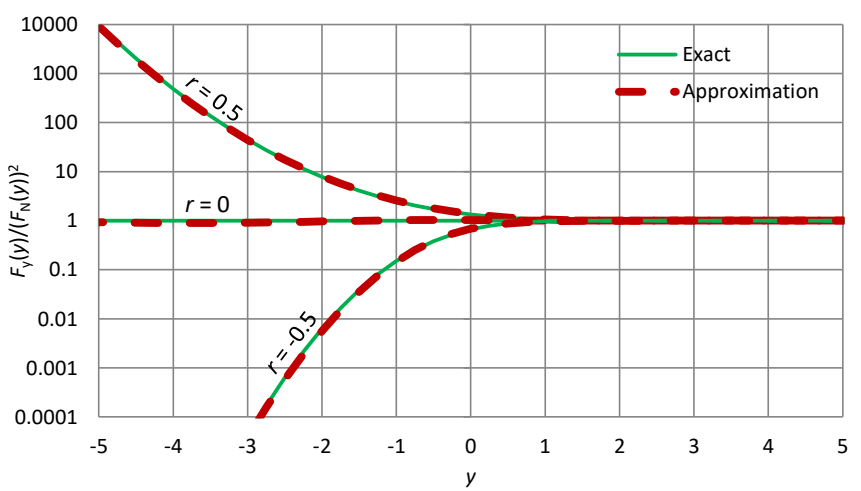

Figure 5.6 Comparison of the approximate and the exact distribution function of the maximum of two correlated variables with standard normal distribution and correlation coefficient r. The approximate distribution function is given by equation (5.48) and the exact is calculated by numerical integration of its density given in Appendix 5-II. The plots are of the ratios $F_y(y)/F_N(y)^2$, where $F_N(y)^2$ is the distribution function of the maximum of two uncorrelated variables.

Furthermore, in Appendix 5-II we find an approximation of the distribution function of the maximum of two correlated variables with standard normal distribution and correlation coefficient r. Denoting the maximum as $\underline{y} := \max(\underline{x}_1, \underline{x}_2)$, its distribution function is:

$$F_y(y) \approx F_N(y) - F_N(-|y|) + \exp\left(\frac{m^2}{2}\right)\frac{F_N(-s|y| - m)}{s} \tag{5.48}$$

where

$$m := 2\sqrt{\frac{1-r}{17+r}}, \qquad s := \frac{1}{3}\sqrt{\frac{17+r}{1+r}} \tag{5.49}$$

5.5 Return period and time scale

The analyses in section 5.1 have been based on a specified discrete time step, D, while reliability and probability of failure are defined in terms of a second characteristic time, the project life span $L = nD$. The study of extremes, thus, involves two operations on an instantaneous stochastic process $\underline{x}(t)$: taking the time average (for discretization on a time step D) and taking the maximum (over the period L). If we change the time step D the results may change and this signifies a problem as the choice of the time step is subjective, determined from the available time resolution of measurements or from modelling conventions. In order to make our analyses more objective we need proper transformations in descriptions among different time scales and also an analysis of the behaviour as time scale tends to zero. We note that the study of extremes in continuous time is a separate scientific field, the *level crossing theory* (e.g. Brill, 2017) but this may not be well suited for hydroclimatic extremes that are typically studied on time-averaged processes.

Replacing the fixed time constant D in equation (5.28) with a varying time scale k, we obtain the following expression for the return period $T^{(k)}$ and the distinct return period $\breve{T}^{(k)}$ at time scale k:

$$T^{(k)} = \frac{k}{1 - F^{(k)}(x)}, \qquad \breve{T}^{(k)} = \frac{k}{F^{(k)}(x) - F_y^{(k)}(x)} \tag{5.50}$$

where $F^{(k)}(x) \coloneqq P\left\{\underline{x}_1^{(k)} \le x\right\}$, $F_y^{(k)}(x) \coloneqq P\left\{\underline{y}^{(k)} \le x\right\}$, $\underline{y}^{(k)} \coloneqq \max\left(\underline{x}_1^{(k)}, \underline{x}_2^{(k)}\right)$, and $\underline{x}_\tau^{(k)}$ is the time-averaged process at time scale k. Now if we assume that the instantaneous process $\underline{x}_t$ has finite variance and is fully smooth (differentiable), it is reasonable to assume that a certain threshold x that is representative of a dangerous event at a certain small time scale k_0 will also be representative for any smaller time scale $k \le k_0$. As $k \to 0$, the distribution function $F^{(k)}(x)$ will tend to the one of the instantaneous process, $F(x)$. Thus, according to equation (5.50), the return period of the fixed threshold value x will vary in proportion to the time scale k and will be precisely zero for the instantaneous process ($k \to 0$), irrespective of the threshold value x. This looks paradoxical but it is a precise result, which suggests that the return period T is not a proper index to move across time scales and to study the instantaneous process. On the other hand, if it happens that $F_y^{(k)}(x) = F(x) - kC$, for some constant C, then it is possible to have a constant distinct return period $\breve{T}^{(k)}$ for a range of time scales, including the instantaneous one.

Before we continue on the latter issue, it is useful to discuss more about the smoothness of a process. In section 3.8 we have introduced the smoothness (or roughness or fractality) parameter M, $0 \le M \le 1$ and we noted that the value $M = 1/2$ signifies neutrality while lower values denote a rough process and higher values denote a smooth process with the value $M = 1$ corresponding to full smoothness. In the latter case the process is (mean-square) differentiable, a property meaning that the first and second derivatives of the autocovariance function exist (Papoulis, 1991, p. 337, 606); in particular the first derivative at zero should be $c'^{(0)} = 0$ (because $c(h)$ is even).

We will provide some illustrations using the FHK-C process whose climacogram is given in (3.87). It can be verified that for M = 1 the climacogram and the autocovariance function are:

$$\gamma(k) = \frac{\lambda}{(1 + (k/\alpha)^2)^{1-H}}, \qquad c(h) = \lambda \frac{1 + (5H-3)(h/\alpha)^2 + H(2H-1)(h/\alpha)^4}{(1 + (h/\alpha)^2)^{3-H}} \tag{5.51}$$

Further, it can be verified that the first two derivatives exist and $c'(0) = 0$ and $c''(0) = -12\lambda(1-H)$. We note that for $M < 1$ these derivatives do not exist at $h = 0$. Figure 5.7 depicts traces of the FHK-C process for three cases, the fully smooth, the neutral and a rough one. The meaning of the smoothness can better be realized by comparing the three traces.

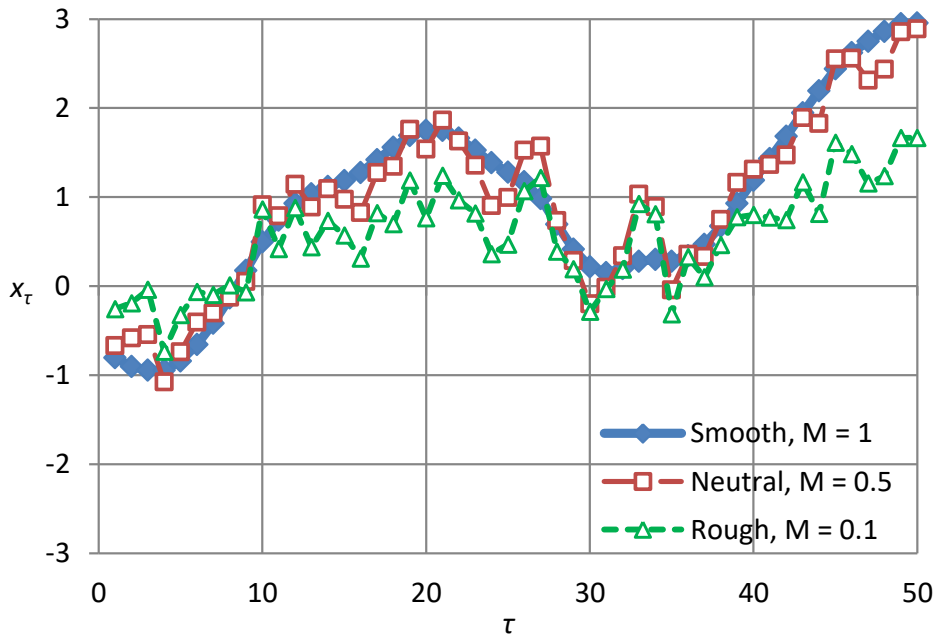

Figure 5.7 Illustration of traces from the FHK-C process for the indicated values of M and for H = 0.8, α = 10 and λ = 1; the same white noise sequence with normal distribution was used for all three cases and were transformed to FHK-C series by SMA filtering (see Chapter 7 for the details of the latter).

Having clarified the meaning of smoothness, we proceed to illustrate the relationships of the threshold x, the return period $T^{(k)}$ and the distinct return period $\breve{T}^{(k)}$ as time scale k varies. For this illustration we use the normal distribution with zero mean, as it is the easiest to handle theoretically, given that it is preserved at any time scale of averaging. We approach the distribution of the maximum of two consecutive variables, $\underline{y}^{(k)} := \max\left(\underline{x}_1^{(k)}, \underline{x}_2^{(k)}\right)$, by the approximation given in Digression 5.C and, in particular, in equation (5.48). We note that $T^{(k)}$ has lower mathematical limit corresponding to $\overline{F}^{(k)}(x) = 1$, which is

$$T_{\min}^{(k)} = k \tag{5.52}$$

However, under the condition $\overline{F}^{(k)}(x) \leq 1/2$ to characterize the event A as dangerous, the lower limits of both $T^{(k)}$ and $\breve{T}^{(k)}$ correspond to $\overline{F}^{(k)}(x) = 1/2$ and are:

$$T_{\min}^{(k)} = 2k, \qquad \breve{T}_{\min}^{(k)} = \frac{k}{1/2 - F_y^{(k)}\left(x_{1/2}\right)} \tag{5.53}$$

where $x_{1/2}$ is the median.

We set forward the hypothesis that in the transformations among scales, what should remain constant is the distinct return period $\breve{T}^{(k)}$. Starting with a fully smooth process, for which, as explained, it is reasonable to fix the threshold value for small scales, we observe in the panels of the upper row in Figure 5.8 that our hypothesis works: a constant $\breve{T}^{(k)}$ indeed yields a constant threshold x. However, if the process is not fully smooth, a

constant x for a changing scale is not a desideratum. Indeed, in a rough process, if we increase the time scale, the averaged process becomes smoother and thus the threshold value should be chosen smaller. Actually, this is the behaviour shown in the middle and lower row in Figure 5.8, and this confirms the reasonability of the hypothesis. Figure 5.9 shows the effect of the time scale parameter α in the variation with time scale k of the threshold x, the return period $T^{(k)}$ and the distinct return period $\breve{T}^{(k)}$. Here the model parameters except α were kept constant and the process studied is neutral in terms of smoothness (M = 0.5). The behaviour looks similar as in Figure 5.8.

Both figures suggest that there is an "optimal" time scale for definition of return period, in which the difference of $T^{(k)}$and $\breve{T}^{(k)}$ is minimum. This optimal time scale is of the order of magnitude of (but not exactly equal to) the model parameter α. If we specify the design return period at this scale, then the choice between T or $\breve{T}$ makes no big difference. But far from this scale, misspecification of the variant of return period that is used may have a dramatic effect. In particular, it is totally inappropriate to specify a design return period for a time scale much larger (say by an order of magnitude) than the optimal as in this case it will not represent anything related to the project safety but just the lower limit $T = 2k$.

5.6 Sample estimation of return period

In an observed sample of size n, the ith smallest value is an estimate of the nth order statistic, $\underline{x}_{(i:n)}$. What is the estimate of the return period of this value?

This question is very important in studying extremes, and its reply is particularly useful for the highest values in the sample. From the analysis on order statistics (section 4.12) we recall that the stochastic variable $\underline{u} \coloneqq F\left(\underline{x}_{(i:n)}\right)$ has beta distribution with parameters i and $n - i + 1$. From well-known results for the Beta distribution, the expected value is:

$$\mathrm{E}\left[\underline{u}\right] = \mathrm{E}\left[F\left(\underline{x}_{(i:n)}\right)\right] = \frac{i}{n+1} \tag{5.54}$$

It is remarkable that this does not depend of the distribution function $F(\,)$. Then an estimate of the return period $\underline{T}_{(i:n)} \coloneqq T\left(\underline{x}_{(i:n)}\right)$ could be:

$$\frac{T_{(i:n)}}{D} = \frac{1}{1 - \mathrm{E}\left[F\left(\underline{x}_{(i:n)}\right)\right]} = \frac{n+1}{n+1-i} \tag{5.55}$$

Hence, the return period estimate for the highest value, the maximum $\underline{T}_{(n)} \coloneqq T\left(\underline{x}_{(n:n)}\right)$, is given as:

$$\frac{T_{(n)}}{D} = n + 1 \tag{5.56}$$

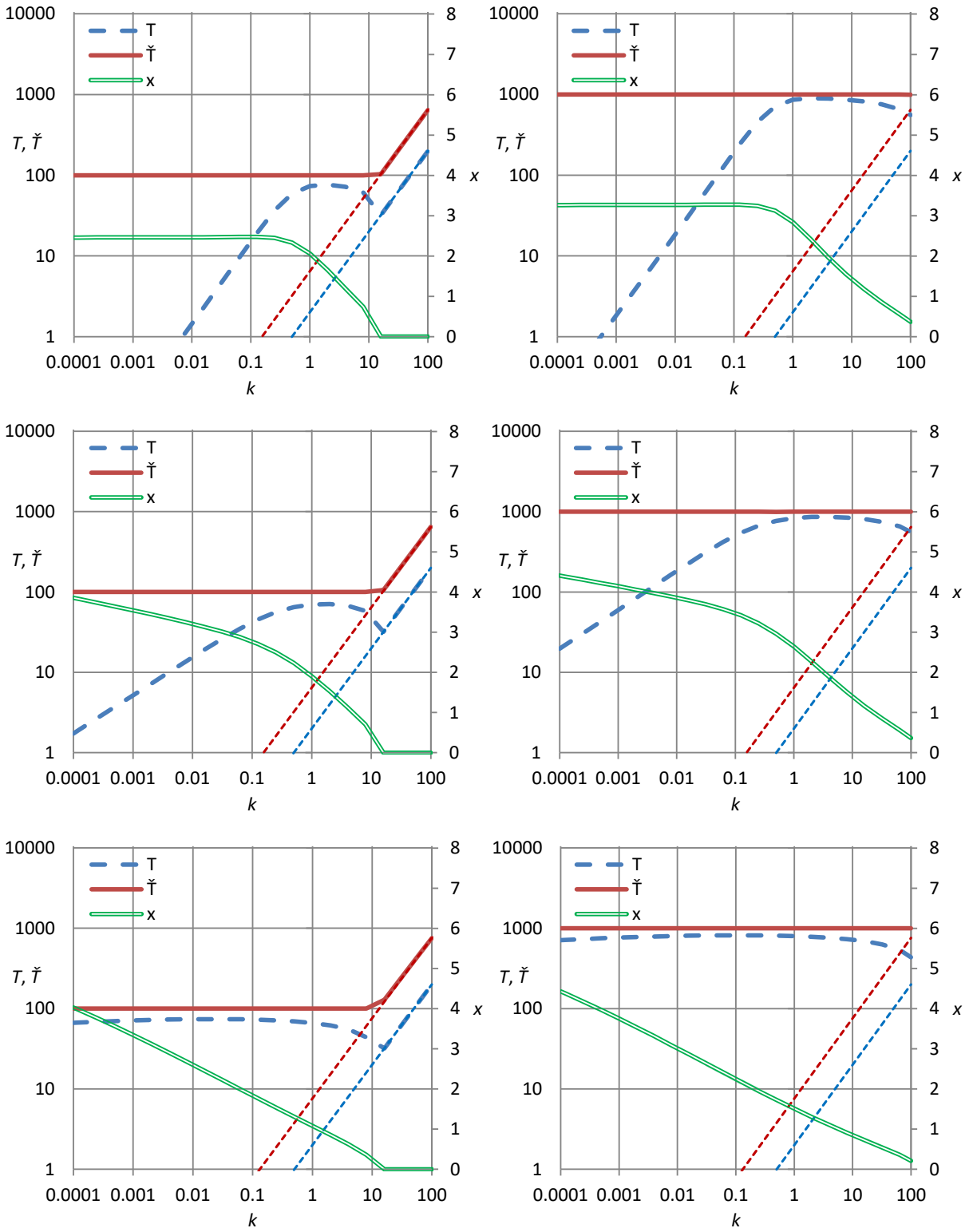

Figure 5.8 Illustration the variation of the threshold value $x^{(k)}$ and the return period $T^{(k)}$ vs. the time k for constant $\check{T}^{(k)} = 100$ **(left column)** and $\check{T}^{(k)} = 1000$ **(right column)** and for a fully smooth process (M = 1, **upper row**), a neutral process (M = 0.5, **middle row**) and a rough process (M = 0.1, **lower row**). In all cases the model is FHK-C with H = 0.8 and α = 1, and with standard normal distribution for the instantaneous process (λ = 1). Dotted lines represent the lowest feasible values of $T^{(k)}$ and $\check{T}^{(k)}$, when P_1= 0 5 and x = 0.

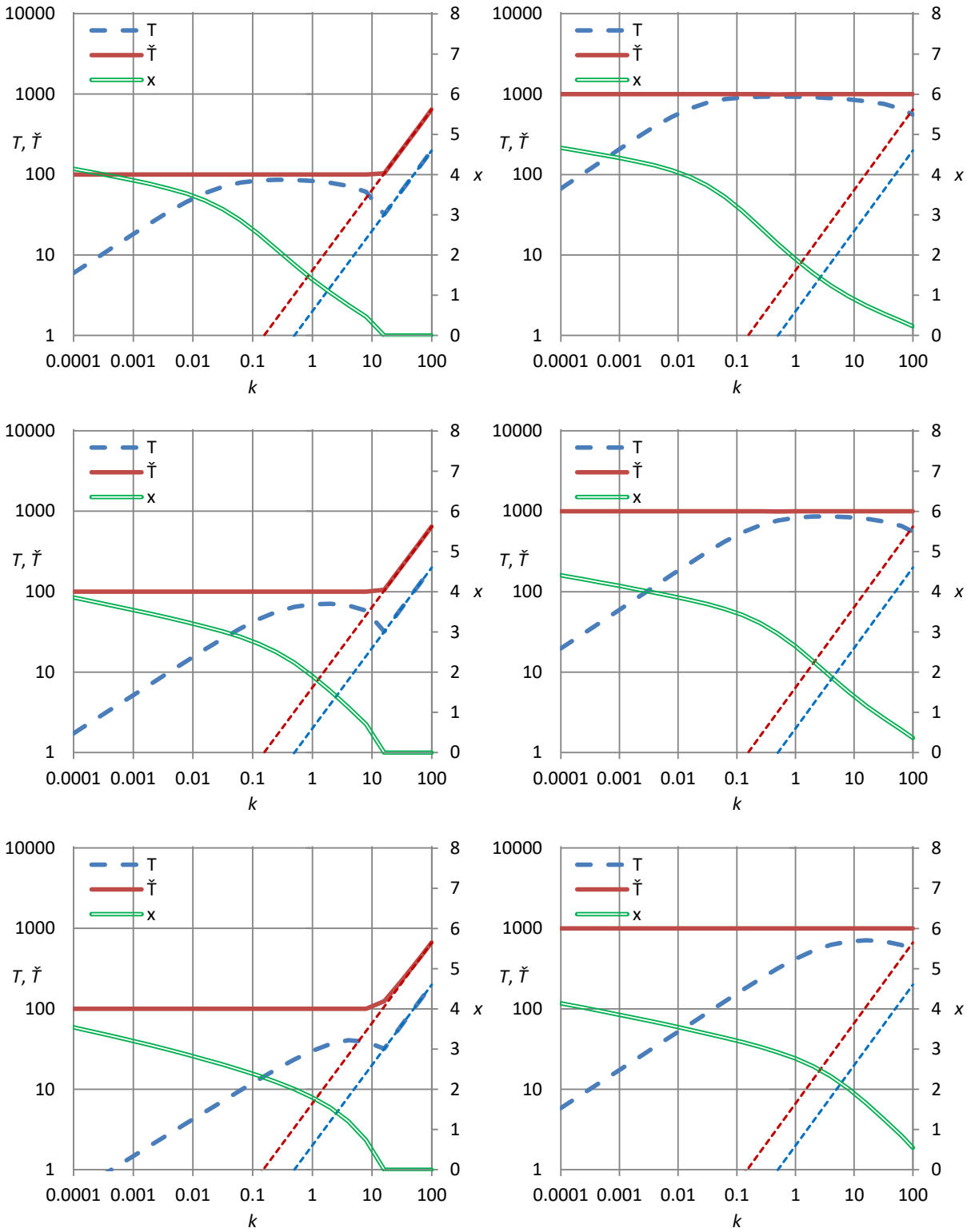

Figure 5.9 Illustration the variation of the threshold value $x^{(k)}$ and the return period $T^{(k)}$ vs. the time k for constant $\check{T}^{(k)} = 100$ **(left column)** and $\check{T}^{(k)} = 1000$ **(right column)** and for a neutral FHK-C process with standard normal distribution for the instantaneous process (λ = 1), M = 0.5, H = 0.8 and α = 0.1 (close to pure randomness; **upper row**), α = 1 **(middle row)** and α = 10 (high short-range dependence; **lower row**). Dotted lines represent the lowest feasible values of $T^{(k)}$ and $\check{T}^{(k)}$, when P_1= 0 5 and x = 0.

Equation (5.55) constitutes the most well-known and the most popular way of assigning return periods to sample values. It is known as the *Weibull plotting position* (Weibull, 1939). However, it is not the earliest, as Hazen (1914) had proposed a different formula, which looks similar but as far as the maximum observation is concerned, the difference in the assigned return period is dramatic, at a ratio of 2:1. Several other formulae have been proposed in the 20th century, which are listed in Table 5.4, along with their basic characteristics.

Table 5.4 Alternative formulae of plotting positions (in chronological order).

Name	$T_{(i:n)}/D$	$T_{(n)}/D$	Comments
Hazen (1914)	$\frac{n}{n+0.5-i}$	$2n$	Empirical
Weibull (1939)	$\frac{n+1}{n+1-i}$	$n+1$	Distribution free, unbiased for $F\left(\underline{x}_{(i:n)}\right)$
Blom (1958)	$\frac{n+1/4}{n+5/8-i}$	$(8/5)n+2/5$	Approximately unbiased quantiles for normal distribution
Tukey (1962)	$\frac{n+1/3}{n+2/3-i}$	$(3/2)n+1/2$	Distribution free, preserving median of $F\left(\underline{x}_{(i:n)}\right)$ (see text)
Gringorten (1963)	$\frac{n+0.12}{n+0.56-i}$	$1.786n+0.21$	Approximately unbiased quantiles for EV1 distribution
Cunnane (1978)	$\frac{n+1/5}{n+3/5-i}$	$(5/3)n+1/3$	Compromise for approximately unbiased quantiles for many distributions

All these relationships are of the form:

$$\frac{T_{(i:n)}}{D}=\frac{n+B}{n-i+A} \tag{5.57}$$

Also, all have a symmetry property. Namely, for the central element $i=m+1$ of a sample with size $n=2m+1$ they yield a return period T/D = 2. In that case, (5.57) yields $2=(2m+1+B)/(2m+1+A-m-1)$, from which we find that the symmetry property requires:

$$B=2A-1 \tag{5.58}$$

All formulae of Table 5.4 satisfy it.

Which of these different formulae should we follow? If we are interested in the small and intermediate items of the observed sample, all formulae give about the same results. But if we are interested in the largest values and particularly the maximum, then the differences are dramatic and the question is crucial. A first element of the answer is that we should avoid the Weibull formula. Certainly, it has some advantages, such as its extreme simplicity, its theoretical justification (which was provided after it was introduced) and the fact that it is distribution free. On the other hand, considering its theoretical justification, it does not make much sense to seek an unbiased estimate of $F\left(\underline{x}_{(i:n)}\right)$ while we aim to assign a return period to a certain value. Indeed, the Weibull formula provides an unbiased estimate of $F\left(\underline{x}_{(i:n)}\right)$, as well as of the exceedance

probability $1 - F\left(\underline{x}_{(i:n)}\right)$, but the estimate of the return period $\underline{T}_{(i:n)}$ given by equation (5.55) if far from unbiased, because of the nonlinearity of the latter equation.

The Tukey formula is also distribution free, and has the advantage that preserves the median of $F\left(\underline{x}_{(i:n)}\right)$ and any of its transformations, including $\underline{T}_{(i:n)}$. The median of the beta distribution of the variable $\underline{u} \coloneqq F\left(\underline{x}_{(i:n)}\right)$ with parameters i and $n - i + 1$ is $\mathrm{I}_{1/2}^{-1}(i, n - i + 1)$, where I^{-1} is the inverse Beta regularized function. A common approximation of that median is $(i - 1/3)/(n - 1/3)$, while for $i = n$ the median is 2^{-n}. Based on these, we can write:

$$\frac{T_{(i:n)}}{D} = \frac{1}{1 - \mathrm{median}\left[F\left(\underline{x}_{(i:n)}\right)\right]} = \begin{cases} \approx \dfrac{n + 1/3}{n - i + 2/3}, & i < n \\ \dfrac{2^{1/n}}{2^{1/n} - 1}, & i = n \end{cases} \tag{5.59}$$

Even using the upper of the two equations also for $i = n$, the error is small, $(3/2)\ln 2 - 1 \approx 0.04$ at most. We can adapt the constants in (5.59) according to (5.57)-(5.58) to make it simpler with a negligible additional error. Namely, if we replace the constant $2/3$ in the denominator with $\ln 2 = 0.693$ and calculate the constant in the numerator by (5.58), we conclude with the following equation which is asymptotically exact for $i = n$:

$$\frac{T_{(i:n)}}{D} = \frac{1}{1 - \mathrm{median}\left[F\left(\underline{x}_{(i:n)}\right)\right]} = \frac{n + 2\ln 2 - 1}{n - i + \ln 2} = \frac{n + 0.386}{n - i + 0.693} \tag{5.60}$$

Next, we seek unbiased estimates for the return period raised to a power ξ, namely of the expectation of the stochastic variable:

$$\underline{v} = \frac{\left(\underline{T}_{(i:n)}/D\right)^{\xi} - 1}{\xi} = \frac{\left(1 - F\left(\underline{x}_{(i:n)}\right)\right)^{-\xi} - 1}{\xi} \tag{5.61}$$

The case $\xi = -1$ corresponds to the unbiased estimation of $F\left(\underline{x}_{(i:n)}\right)$, which is already discussed. The case $\xi = 1$ corresponds to the unbiased estimation of the return period. The case $\xi = 0$ corresponds to the unbiased estimation of the logarithm of return period Thus, this general setting allows studying at once several cases for ξ varying in $[-1,1]$.

As $\underline{u} \coloneqq F\left(\underline{x}_{(i:n)}\right)$ has Beta distribution with parameters i and $n - i + 1$, the same distribution will have the variable $1 - \left(\xi\underline{v} + 1\right)^{-1/a}$. Thus, from (4.62) it follows that:

$$F_v(v) = \frac{\mathrm{B}_{1-(\xi v+1)^{-1/a}}(i, n - i + 1)}{\mathrm{B}(i, n - i + 1)} \tag{5.62}$$

After tedious algebraic manipulations, which are omitted, the expected value of $\underline{v}$ is found to be:

$$\mathrm{E}\left[\underline{v}\right] = \frac{\Gamma(n + 1)\Gamma(n + 1 - i - \kappa)}{\kappa\Gamma(n + 1 - \kappa)\Gamma(n + 1 - i)} - \frac{1}{\kappa} \tag{5.63}$$

Hence, for $v = \mathrm{E}\left[\underline{v}\right]$, the estimate of the return period is

$$\frac{T\left(x_{(i:n)}\right)}{D} = \left(\frac{\Gamma(n+1)\Gamma(n+1-i-\xi)}{\Gamma(n+1-\xi)\Gamma(n+1-i)}\right)^{1/\xi} \tag{5.64}$$

This simplifies to the following for integer ξ; specifically, for $\xi = -1$ it becomes:

$$\frac{T\left(x_{(i:n)}\right)}{D} = \frac{n+1}{n+1-i} \tag{5.65}$$

for $\xi = 0$:

$$\frac{T\left(x_{(i:n)}\right)}{D} = \frac{\mathrm{e}^{H_n}}{\mathrm{e}^{H_{n-i}}} \tag{5.66}$$

where $H_n \coloneqq \sum_{i=1}^{n} 1/i$ the nth harmonic number, and for $\xi = 1$:

$$\frac{T\left(x_{(i:n)}\right)}{D} = \frac{n}{n-i} \tag{5.67}$$

Illustration of results for other values of ξ in the interval $[-1,1]$ is provided in Figure 5.10, where it is evident that assignment of return periods for the smallest half of the sample is indifferent to choice of ξ, but as we approach the largest value, $i = n$, the effect of ξ becomes dramatic. In particular, the option ξ = 1 (unbiased T) is unable to assign a return period to the largest value. Following this option practically means that we have to discard the largest value and assign a return period nD to the second largest. The reason is that equation (5.67) diverges for $i = n$. The divergence does not make this option an appropriate choice, while for reasons explained above, the option $\xi = -1$ is also inappropriate.

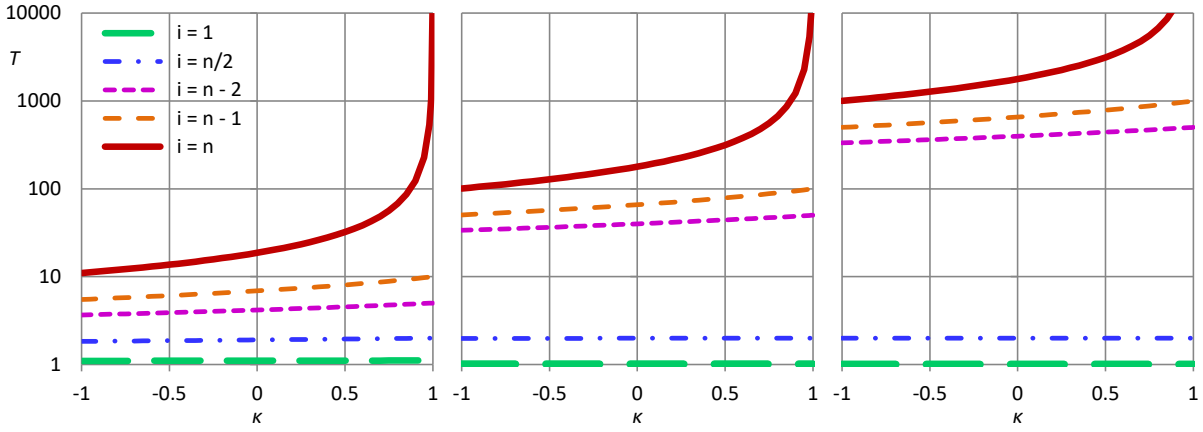

Figure 5.10 Illustration of the variation of the return period assigned by equation (5.64) to the ith smallest value in a sample of size n for the indicated values of i and for n = 10, 100 and 1000 for the left, middle and right panel, respectively.

In Figure 5.10 we may also see that the option $\xi = 0$, corresponding to an unbiased estimate of the logarithm of return period, is most promising and balanced and therefore we can proceed with this. Even though the equation (5.66) (for $\xi = 0$) is easy to evaluate, we can also make an approximation in the form (5.57), which is more practical, and accurate enough. To this aim, we proceed as follows. First we consider the case $i = n$ and

from (5.57) we find $T_{(n:n)}/D = (n+B)/A$ or $T_{(n:n)}/Dn = 1/A + B/An$. Taking the limit as $n \to \infty$ we obtain:

$$\Lambda_\infty \coloneqq \lim_{n\to\infty} \frac{1}{n\left(1 - F\left(x_{(n)}\right)\right)} = \frac{1}{A} \tag{5.68}$$

Next, we consider the other extreme case, $i = n = 1$, setting $x_{(1)} = \mu$ (the mean), and we similarly find:

$$\Lambda_1 \coloneqq \frac{1}{1 - F(\mu)} = \frac{1+B}{A} \tag{5.69}$$

The coefficients Λ_∞ and Λ_1 will be better explained in section 6.14. For now, it suffices to determine A and B by solving the system of the last two equations, i.e.,

$$A = \frac{1}{\Lambda_\infty}, \qquad B = \frac{\Lambda_1}{\Lambda_\infty} - 1 \tag{5.70}$$

From (5.57), $\Lambda_1 = \mathrm{e} = 2.718$ and $\Lambda_\infty = \mathrm{e}^{\gamma} = 1.781$, where $\gamma = 0.5772$ is the Euler constant. Thus, the approximation sought is:

$$\frac{T\left(x_{(i:n)}\right)}{D} \approx \frac{n + \mathrm{e}^{1-\gamma} - 1}{n - i + \mathrm{e}^{-\gamma}} = \frac{n + 0.526}{n - i + 0.561} \tag{5.71}$$

As $n \to \infty$, while (5.71) is exact for $T\left(x_{(n:n)}\right)$, it underestimates the return period of the second largest value, $T\left(x_{(n-1:n)}\right)$, by a factor:

$$\frac{\mathrm{e}}{1 + \mathrm{e}^{\gamma}} - 1 = -0.023 \tag{5.72}$$

Nonetheless, as *i* gets smaller, the error in determining $T\left(x_{(i:n)}\right)$ becomes zero.

We can also formulate a slightly better approximation, by distinguishing the equation for the maximum value from all others. That is:

$$\frac{T\left(x_{(i:n)}\right)}{D} \approx \begin{cases} \mathrm{e}^{\gamma}(n-1) + \mathrm{e} = 1.781n + 0.94, & i = n \\ \dfrac{n - 1 + \mathrm{e}^{1-\gamma}}{n - i - 1 + \mathrm{e}^{1-\gamma}} & i < n \end{cases} \tag{5.73}$$

where $\gamma = 0.5772$ is the Euler constant. This is asymptotically exact, as $n \to \infty$, for both the largest and the second largest sample values. Notice that neither the accurate equation (5.66) nor the approximations (5.71)-(5.73) are symmetric; namely for $n = 2m + 1$ and $i = m + 1$ they result in return period slightly higher than $2D$ (as demanded by symmetry), which tends to $2D$ as $n \to \infty$. Consequently, the equation of symmetry (5.58) does not hold.

With the same methodology, appropriately modified, we can determine unbiased estimates for other quantities, for example distribution quantiles. In fact, no adaptation is needed for the exponential and Pareto distribution as the variable *v* in equation (5.61) also represents the quantiles of those distributions. For the normal distribution we have $\Lambda_\infty = \mathrm{e}^{\gamma} = 1.781$ (same as above for $\xi = 1$) and $\Lambda_1 = 2$ (due to symmetry). The resulting formula is contained Table 5.5, along with a summary of all above results.

Table 5.5 Suggested formulae of sample estimation of return period (plotting position).

No.	$T_{(i:n)}/D$	$T_{(n)}/D$	Preserved quantity
I	$\frac{n+2\ln 2-1}{n-i+\ln 2}=\frac{n+0.386}{n-i+0.693}$	$(n-1)/\ln 2+2$ $=1.443n+0.56$	Median of $F\left(\underline{x}_{(i:n)}\right)$, $\bar{F}\left(\underline{x}_{(i:n)}\right)$, $\underline{T}_{(i:n)}$, of any distribution
II	$\frac{\mathrm{e}^{H_n}}{\mathrm{e}^{H_{n-i}}}$	e^{H_n}	Logarithm of $\underline{T}_{(i:n)}$ of any distribution; quantile $\underline{x}_{(i:n)}$ of exponential distribution (Exact)
III	$\frac{n+\mathrm{e}^{1-\gamma}-1}{n-i+\mathrm{e}^{-\gamma}}=\frac{n+0.526}{n-i+0.561}$	$\mathrm{e}^{\gamma}(n-1)+\mathrm{e}$ $=1.781n+0.94$	Approximation of II; also valid for the quantile $\underline{x}_{(i:n)}$ of positively skewed distributions of the EV1 domain of attraction
IV	$\frac{n+2\mathrm{e}^{-\gamma}-1}{n-i+\mathrm{e}^{-\gamma}}=\frac{n+0.123}{n-i+0.561}$	$\mathrm{e}^{\gamma}(n-1)+2$ $=1.781n+0.22$	Quantile $\underline{x}_{(i:n)}$ of normal distribution; also valid for symmetrical distributions of the EV1 domain of attraction
V	$\left(\frac{\Gamma(n+1)\Gamma(n+1-i-\xi)}{\Gamma(n+1-\xi)\Gamma(n+1-i)}\right)^{1/\xi}$	$\left(n\mathrm{B}(n,1-\xi)\right)^{1/\xi}$	Power of $\underline{T}_{(i:n)}$ to exponent ξ for any distribution; quantile $\underline{x}_{(i:n)}$ of Pareto distribution (Exact)
VI	$\frac{n+B}{n-i+A}$ $A=\left(\Gamma(1-\xi)\right)^{-1/\xi}$ $B=\left(\Gamma(2-\xi)\right)^{-1/\xi}-1$	$(n+B)/A$, e.g., $\xi=0.15$: $2.035n+0.92$ $\xi=0.5$: $3.142n+0.86$	Approximation of V; also valid for distributions of the EV2 domain of attraction with tail index ξ.

As a final suggestion after the above detailed analysis, equation (5.73) —case III (or if theoretical rigour is sought, (5.66)—case II) is the most appropriate for practical use for any distribution function. It represents unbiasedness in the estimation of the logarithm of return period which is a balance between unbiasedness in distribution function and in return period. It also provides unbiased quantile estimation for the exponential distribution. Furthermore, as seen in Table 5.5, for all distributions of the domain of attraction of EV1 the return period of the largest value is 1.781 times n (plus a constant), while in those of EV2 it is even higher. Having this in mind and observing the traditional formulae of Table 5.4 only the Hazen and Gringorten formulae satisfy this condition. Interestingly, the Gringorten formula is the virtually the same as case IV of Table 5.5. Note though that case IV was derived here for the normal distribution and not for the EV1 distribution, for which here we suggest the, slightly different, case III of Table 5.5 Thus, if one prefers to use the traditional formulae, one should prefer the Gringorten formula. We may also notice that the Hazen formula is not unjustified in light of the above analysis. Rather it appears equivalent to case VI of Table 5.5 which is for the Pareto distribution with $\xi = 0.15$.

The fact that case III, as an unbiased estimator of the logarithm of return period, is distribution free is perhaps its most significant advantage. Usually, in practice we need to

assign return periods to sample values prior to choosing a model and, in this respect, case III is an advantageous choice. Once we choose a model, we may reconsider the assignment of return periods using more accurate methods and formulae, which are based on K-moments that are discussed in Chapter 6 and particularly section 6.14. The latter analyses also consider other effects, such as the possible existence of autocorrelation, which influences the estimation of return periods.

Digression 5.D: Illustration of the range of sample estimates of return period

A stochastic simulation always helps to develop better intuition of theoretical concepts and spot possible errors. Here we perform a simulation exercise generating m = 10 000 random samples, each consisting of n = 10 values from the Pareto distribution with tail index ξ = 0.5. We intentionally choose a small n and a large ξ (note that for this ξ the variance of the variable diverges to infinity) for better illustration. For the chosen Pareto distribution of the parent stochastic variable $\underline{x}$ we have:

$$F(x) = 1 - (1+\xi x)^{-\frac{1}{\xi}}, \qquad \frac{T(x)}{D} = \frac{1}{1-F(x)} = (1+\xi x)^{\frac{1}{\xi}}$$

The theoretical return period of the ith order statistic $\underline{x}_{(i:n)}$ is determined from equation (4.62), i.e.,

$$\frac{T_{(i:n)}(x)}{D} = \frac{\mathrm{B}(i, n-i+1)}{\mathrm{B}(i, n-i+1) - \mathrm{B}_{F(x)}(i, n-i+1)}$$

For the special cases of the minimum and maximum order statistic we have, respectively,

$$T_{(1:n)}(x) = \left(\frac{T(x)}{D}\right)^n, \qquad 1 - \frac{D}{T_{(n:n)}(x)} = \left(1 - \frac{D}{T(x)}\right)^n$$

where in our case n = 10.

The theoretical curves $T(x)$ for the parent variable and for the largest and second largest order statistics, $\underline{x}_{(10)} \equiv \underline{x}_{(10:10)}$ and $\underline{x}_{(9:10)}$ have been plotted in the left panel of Figure 5.11. An interesting observation is that the curve $\underline{x}_{(n-1:n)}$ crosses that of $\underline{x}$ but not that of $\underline{x}_{(n:n)}$. Actually, this happens at all $i = 2, \ldots, n-1$ and for large n the intersection point corresponds to:

$$\frac{T_{(i:n)}(x)}{D} \approx \frac{n-1}{i-1}$$

For each of the 10 000 samples the following values were evaluated

$$x_{(10)}, \qquad F\left(x_{(10)}\right), \qquad T\left(x_{(10)}\right), \qquad \ln T\left(x_{(10)}\right)$$

and the same list for $x_{(9:10)}$ too. The empirical (sample) distributions of x, $x_{(10)}$ and $x_{(9:10)}$ are also plotted in the left panel of Figure 5.11 in the form of curves $T(x)$. They were estimated from equation (5.71) (option III of Table 5.5). Generally, the empirical plots show good agreement with the theoretical ones, thus indicating the consistency of the proposed framework.

Next the averages and the medians from all m = 10 000 values of the above list of variables were calculated. For $x_{(10)}$ the averages and the medians are plotted in the right panel of Figure 5.11, over the curve $T(x)$ of the parent variable. In all cases these empirical estimates fully agree (the points coincide) with the theoretical ones, with one exception: The theoretical average of $T\left(x_{(10)}\right)$ is ∞, while the sample estimate is necessarily a finite value, yet a very large one (>100D). Generally, the plot shows that the differences between the different options for assigning return period to the largest value (represented by the different points) can be substantial. The points corresponding to $\ln T\left(x_{(10)}\right)$ and $x_{(10)}$ look the more balanced choices, as already discussed.

We must note that the differences would be less dramatic if the sample size was greater or if the distribution had a smaller tail index.

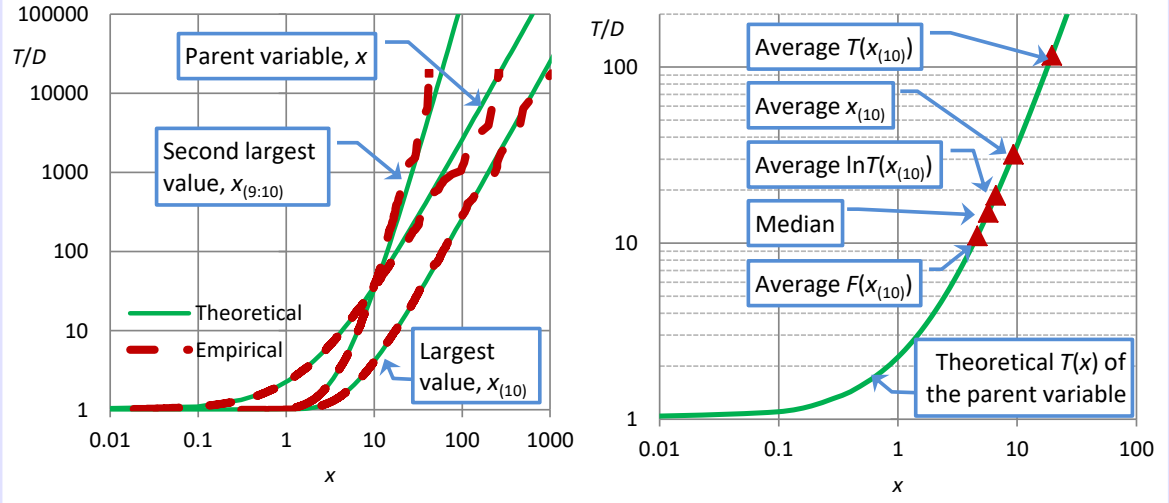

Figure 5.11 Simulation results from m = 10 000 random samples, each consisting of n = 10 values, from the Pareto distribution with tail index ξ = 0.5: **(left)** Theoretical and empirical (sample) distributions of the indicated variables, where the theoretical distribution were determined from equation (4.62) and the empirical ones from (5.71) (option III of Table 5.5). **(right)** Simulated averages from the m = 10 000 values of the indicated variables. The median, also plotted, is indifferent to the choice of the variable.

Digression 5.E: A funny way to calculate π through the properties of maxima

The curious reader may have noticed that the formula given in Table 5.5 (case VI) for the return period of the maximum value of the Pareto distribution with $\xi = 0.5$ contains the value of π. More rigorously, the formula in this case is

$$T_{(n)}/D = \pi(n-1) + 4$$

which for n = 2 becomes π + 4 (rather than 3, a value that would be assigned if we adopted the Weibull plotting position formula). Solving for π we find

$$\pi = \frac{T_{(n)}/D - 4}{(n-1)}$$

This enables a Monte Carlo technique to calculate π. Interestingly, ideas that could be classified as implementation of the Monte Carlo method to calculate π are much older than the formal Monte Carlo method. Georges Louis LeClerc (Comte de Buffon, French scientist; 1707-1788) became famous for "Buffon's needle," a method using needle tosses onto a lined background to estimate π (where, if the line distance is equal to needle length, π is found as twice the inverse of probability that the needle crosses a line). LeClerc's method became popular among scientists and his experiment was later repeated by many. However, his method and many other Monte Carlo algorithms to estimate π, including the one presented here, are good only for fun. Much faster and much more accurate deterministic algorithms exist to calculate π. Reitwiesner (1950) calculated by a deterministic algorithm, running on the ENIAC computer, the first 2035 decimal digits of π. Metropolis et al. (1950) examined their randomness, an exercise made thereafter many times showing that the digits of π have no apparent pattern and pass tests for statistical randomness. Dodge (1996) promoted an idea opposite to LeClerc's: that the digits of π form a "Natural Random Number Generator". Since January 2019, 31.4 trillion digits of π are known (found by the Chudnovsky algorithm); this information, equivalent to ~100 million books of 1000 pages each (note for comparison that the British Library has 25 million books), can serve as a basis for any simulation experiment. However, the simple random generators discussed in section 2.6 are more economic and convenient.

After this historical note, we return to our method to estimate π from the maxima of the Pareto distribution. The equation above, allows formulating the following algorithm.

1. We generate n random numbers from the Pareto distribution with $\xi = 0.5$ and take the maximum.
2. We repeat this procedure m times and calculate the average of all m maxima.
3. We find the return period of this average from the theoretical relationship of the Pareto distribution and calculate π from the above equation.

For step 1 we note that a random number from the Pareto distribution with $\xi = 0.5$ is generated from

$$x = \left(\frac{u^{-\xi} - 1}{\xi}\right) = 2\left(\frac{1}{\sqrt{u}} - 1\right)$$

where u is a random number from the uniform distribution. The maximum of n random numbers is thus

$$x_{(n)} = \max(x_1, \ldots, x_n) = 2\left(\frac{1}{\sqrt{\min(u_1, \ldots, u_n)}} - 1\right)$$

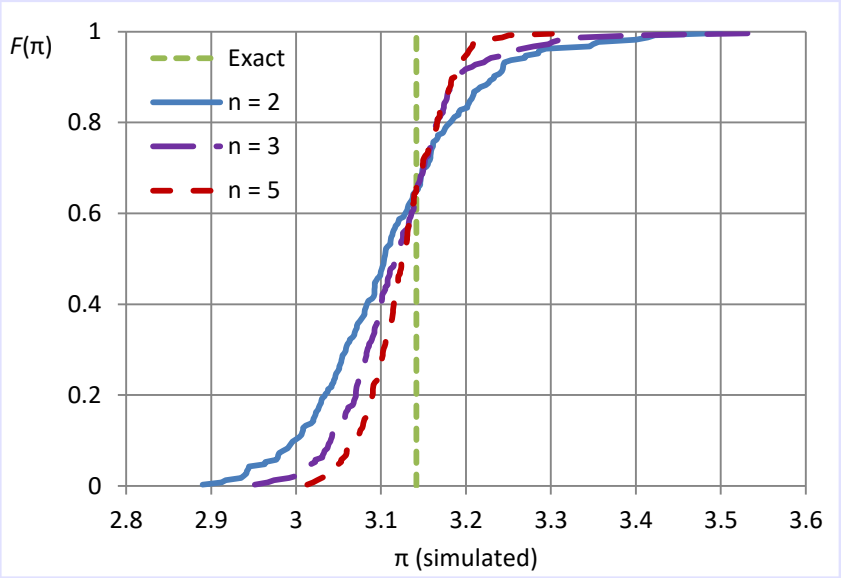

Figure 5.12 Simulation results to estimate π from m = 100 000 random samples, each consisting of n random numbers from the Pareto distribution with tail index ξ = 0.5. The curves depict the sampling distributions of π estimates for the indicated three values of n.

For step 2, the average of m simulated $x_{(n)}$, each denoted as $x_{(n)}^i$, $i = 1, \ldots, m$, is

$$\bar{x}_{(n)} = 2\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\sqrt{\min\left(u_1^i, \ldots, u_n^i\right)}} - 1\right)$$

For step 3, the return period of the average is

$$\frac{T_{(n)}}{d} = \left(1 + \xi\bar{x}_{(n)}\right)^{\frac{1}{\xi}} = \left(1 + \frac{\bar{x}_{(n)}}{2}\right)^2 = \left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\sqrt{\min\left(u_1^i, \ldots, u_n^i\right)}}\right)^2$$

and the simulated value of π is

$$\pi = \frac{1}{(n-1)}\left(\left(\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\sqrt{\min\left(u_1^i, \ldots., u_n^i\right)}}\right)^2 - 4\right)$$

Obviously, the larger the values of m and n, the better the estimate. Even for the minimum possible value, n = 2, the result is not bad, as shown in Figure 5.12. The three curves for n = 2, 3 and 5 shown in the figure intersect at π = 3.14.

Appendix 5-I: Approximation of the normal distribution for inferring the behaviour of its extremes

The density of the standard normal distribution can be written as:

$$f_{\mathrm{N}}(x) = \exp\left(-\ln\sqrt{2\pi} - x^2/2\right) \tag{5.74}$$

By numerical investigation it is seen that an approximation of its distribution function is:

$$F_{\mathrm{N}}(x) = 1 - \int_x^{\infty} f_{\mathrm{N}}(y)\,\mathrm{d}y \approx 1 - g(x; c_0, c_1, c_2) \tag{5.75}$$

where c_0, c_1, c_2 are numerical constants and the function $g(x; a_0, a_1, a_2)$, for any a_0, a_1, a_2, is defined as:

$$g(x; a_0, a_1, a_2) := \begin{cases} \exp\left(-(a_0 + a_1 x + a_2 x^2)\right), & x \geq 0 \\ 1 - g(-x; a_0, a_1, a_2), & x \leq 0 \end{cases} \tag{5.76}$$

An interesting property is that the function $g(\,)$ is preserved under multiplication for $x > 0$:

$$g(x; a_0, a_1, a_2)\, g(cx; b_0, b_1, b_2) = g(x; a_0 + b_0, a_1 + cb_1\,, a_2 + c^2 b_2) \tag{5.77}$$

Notice that the density function $f_{\mathrm{N}}(x)$ can itself be written as

$$f_{\mathrm{N}}(x) = g\left(x;\ \ln\sqrt{2\pi}\,, 0, 1/2\right) \tag{5.78}$$

Now for x = 0, $F_{\mathrm{N}}(x) = 1/2$, so that $g(0; c_0, c_1, c_2) = 1/2$ and in order for this to hold we must set $\exp(-c_0) = 1/2$ or $c_0 = \ln 2$. The constants c_1 and c_2 can be determined by minimizing the fitting error. Several combinations can provide a good fitting; here we preferred the following combination, easy to remember:

$$c_0 = \ln 2\,, \qquad c_1 = \frac{2}{3}, \qquad c_2 = \left(\frac{2}{3}\right)^2 = \frac{4}{9} \tag{5.79}$$

This yields equation (5.44).

We now determine integrals of the function $g(\,)$, which are useful in other calculations. We consider a linear transformation of x, $w = sx + m, w > 0$. Using calculus of probability, we find:

$$F_{\mathrm{N}}(sx + m) = 1 - \int_x^{\infty} s\, f_{\mathrm{N}}(sw + m)\,\mathrm{d}w \tag{5.80}$$

$$\begin{aligned} s\, f_{\mathrm{N}}(sw+m) &= \exp\left(\ln s - \frac{1}{2}\ln(2\pi) - \frac{(sw+m)^2}{2}\right) \\ &= \exp\left(\ln s - \frac{1}{2}\ln(2\pi) - \frac{m^2}{2} - msw - \frac{s^2w^2}{2}\right) \end{aligned} \tag{5.81}$$

For appropriate a_0, a_1, a_2 and c, namely those satisfying:

$$s = \sqrt{2a_2}, \qquad m = a_1/\sqrt{2a_2}, \qquad \ln c = \ln\sqrt{\frac{a_2}{\pi}} + a_0 - \frac{a_1^2}{4a_2} \tag{5.82}$$

equation (5.81) can be written as

$$sf_{\mathrm{N}}(sw+m) = c\; g(w; a_0, a_1, a_2) = \exp\big(\ln c - (a_0 + a_1 w + a_2 w^2)\big) \tag{5.83}$$

Combining the above equations, we have:

$$F_{\mathrm{N}}(sx+m) = 1 - c\int_x^{\infty} g(w; a_0, a_1, a_2)\,\mathrm{d}w \tag{5.84}$$

and solving for the integral we find

$$\int_x^{\infty} g(w; a_0, a_1, a_2)\,\mathrm{d}w = \frac{1 - F_{\mathrm{N}}\big(\sqrt{2a_2}\,x + a_1/\sqrt{2a_2}\big)}{\sqrt{a_2/\pi}\exp(a_0 - a_1^2/4a_2)} \tag{5.85}$$

which is exact and valid for $x > 0$. If we approximate F_{N} using $g(\,)$ as in (5.75), then after algebraic manipulations we find:

$$\int_x^{\infty} g(w; a_0, a_1, a_2)\,\mathrm{d}w \approx g\left(x; \ln\left(2\sqrt{\frac{a_2}{\pi}}\right) + a_0 + \frac{\sqrt{2}a_1}{3\sqrt{a_2}} - \frac{a_1^2}{36a_2}, \frac{8a_1}{9} + \frac{2\sqrt{2a_2}}{3}, \frac{8a_2}{9}\right) \tag{5.86}$$

It can be verified that if $(a_0, a_1, a_2) = (\ln\sqrt{2\pi}, 0, 1/2)$ as in (5.78), then the corresponding a coefficients in the right-hand side of (5.86) become (ln2, 2/3, 4/9) as in (5.79).

Appendix 5-II: Approximation of the distribution of extremes of two correlated normal variables

We assume that $(\underline{x}_1, \underline{x}_2)$ have standard normal distribution and are dependent with correlation coefficient r. We define:

$$\underline{y} \coloneqq \max(\underline{x}_1, \underline{x}_2), \qquad \underline{z} \coloneqq \min(\underline{x}_1, \underline{x}_2) \tag{5.87}$$

The exact probability densities are (Nadarajah and Katz, 2008):

$$f_y(y) = 2f_{\mathrm{N}}(y)F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\,y\right), \qquad f_z(z) = 2f_{\mathrm{N}}(z)F_{\mathrm{N}}\left(-\sqrt{\frac{1-r}{1+r}}\,z\right) = 2f_{\mathrm{N}}(z) - f_y(z) \tag{5.88}$$

Using calculus of probability, we find:

$$\mathrm{E}\left[\underline{y}\right] = \sqrt{\frac{1-r}{\pi}}, \qquad \mathrm{E}\left[\underline{z}\right] = -\sqrt{\frac{1-r}{\pi}} \tag{5.89}$$

However, there is no analytical solution for the exact distribution function (the integral of $f_z(z)$) and we seek an approximation. For $y > 0$ the distribution function is:

$$\begin{aligned} F_y(y) &= 1 - \int_y^\infty 2 f_{\mathrm{N}}(w) F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\, w\right) \mathrm{d}w \\ &= 1 - 2\int_y^\infty f_{\mathrm{N}}(w)\,\mathrm{d}w + 2\int_y^\infty f_{\mathrm{N}}(w)\left(1 - F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\, w\right)\right)\mathrm{d}w \\ &= 2F_{\mathrm{N}}(y) - 1 + 2\int_y^\infty f_{\mathrm{N}}(w)\left(1 - F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\, w\right)\right)\mathrm{d}w \end{aligned} \tag{5.90}$$

Using the approximation (5.75)-(5.76) and the property (5.77), the last integral in (5.90) becomes:

$$\begin{aligned} &\int_y^\infty f_{\mathrm{N}}(w)\left(1 - F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\, w\right)\right)\mathrm{d}w \\ &\quad\approx \int_y^\infty g\left(w;\ \ln\sqrt{2\pi}, 0, 1/2\right) g\left(\sqrt{\frac{1-r}{1+r}}\, w;\ \ln 2, \frac{2}{3}, \frac{4}{9}\right)\mathrm{d}w \\ &\quad= \int_y^\infty g\left(w;\ \ln\sqrt{8\pi}, \frac{2}{3}\sqrt{\frac{1-r}{1+r}}, \frac{4}{9}\frac{1-r}{1+r} + \frac{1}{2}\right)\mathrm{d}w \\ &\quad= \frac{1 - F_{\mathrm{N}}\left(\frac{1}{3}\sqrt{\frac{17+r}{1+r}}\, y + 2\sqrt{\frac{1-r}{17+r}}\right)}{\frac{2}{3}\sqrt{\frac{17+r}{1+r}}\exp\left(-2\frac{1-r}{17+r}\right)} \end{aligned} \tag{5.91}$$

Consequently, combining (5.90) and (5.91), we find

$$F_y(y) = 2F_{\mathrm{N}}(y) - 1 + \frac{1 - F_{\mathrm{N}}\left(\frac{1}{3}\sqrt{\frac{17+r}{1+r}}\, y + 2\sqrt{\frac{1-r}{17+r}}\right)}{\frac{1}{3}\sqrt{\frac{17+r}{1+r}}\exp\left(-2\frac{1-r}{17+r}\right)} \tag{5.92}$$

By setting

$$m \coloneqq 2\sqrt{\frac{1-r}{17+r}}, \qquad s \coloneqq \frac{1}{3}\sqrt{\frac{17+r}{1+r}} \tag{5.93}$$

equation (5.92) can be written as

$$F_y(y) = 2F_{\mathrm{N}}(y) - 1 + \exp\left(\frac{m^2}{2}\right)\frac{1 - F_{\mathrm{N}}(sy+m)}{s} = 2F_{\mathrm{N}}(y) - 1 + \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(-sy-m)}{s} \tag{5.94}$$

Note that $0 \le m \le 1/\sqrt{2}, s \ge 1$.

Likewise, for $y < 0$:

$$F_y(y) = \int_{-\infty}^{y} 2f_{\mathrm{N}}(w)F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\,w\right)\mathrm{d}w \tag{5.95}$$

The integral now is:

$$\int_{-\infty}^{y} f_{\mathrm{N}}(w)F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\,w\right)\mathrm{d}w \approx \int_{-\infty}^{y} g\left(-w;\ \ln\sqrt{2\pi}, 0, 1/2\right)g\left(-\sqrt{\frac{1-r}{1+r}}w;\ \ln 2, \frac{2}{3}, \frac{4}{9}\right)\mathrm{d}w \tag{5.96}$$

and hence:

$$\begin{aligned}
\int_{-\infty}^{y} f_{\mathrm{N}}(w)F_{\mathrm{N}}\left(\sqrt{\frac{1-r}{1+r}}\,w\right)\mathrm{d}w &\approx \int_{-\infty}^{y} g\left(w;\ \ln\sqrt{8\pi}, -\frac{2}{3}\sqrt{\frac{1-r}{1+r}}, \frac{4}{9}\frac{1-r}{1+r}+\frac{1}{2}\right)\mathrm{d}w \\
&= \int_{-y}^{\infty} g\left(w;\ \ln\sqrt{8\pi}, \frac{2}{3}\sqrt{\frac{1-r}{1+r}}, \frac{4}{9}\frac{1-r}{1+r}+\frac{1}{2}\right)\mathrm{d}w \\
&= \frac{1-F_{\mathrm{N}}\left(-\frac{1}{3}\sqrt{\frac{17+r}{1+r}}\,y + 2\sqrt{\frac{1-r}{17+r}}\right)}{\frac{2}{3}\sqrt{\frac{17+r}{1+r}}\exp\left(-2\frac{1-r}{17+r}\right)} = \frac{F_{\mathrm{N}}\left(\frac{1}{3}\sqrt{\frac{17+r}{1+r}}\,y - 2\sqrt{\frac{1-r}{17+r}}\right)}{\frac{2}{3}\sqrt{\frac{17+r}{1+r}}\exp\left(-2\frac{1-r}{17+r}\right)} \\
&= \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(sy-m)}{s}
\end{aligned} \tag{5.97}$$

Summarizing, the distribution function of $\underline{y} \coloneqq \max\left(\underline{x}_1, \underline{x}_2\right)$ is:

$$F_y(y) \approx \begin{cases} 2F_{\mathrm{N}}(y) - 1 + \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(-sy-m)}{s}, & y \geq 0 \\ \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(sy-m)}{s}, & y \leq 0 \end{cases} \tag{5.98}$$

For $r = 0$, for which $m = 2/\sqrt{17}, s = \sqrt{17}/3$, the result is:

$$F_y(y) \approx \begin{cases} 2F_{\mathrm{N}}(y) - 1 + \frac{\sqrt{17}}{3}\exp\left(\frac{2}{17}\right)\frac{F_{\mathrm{N}}\left(-\sqrt{17}y/3 - 2/\sqrt{17}\right)}{s}, & y \geq 0 \\ \frac{\sqrt{17}}{3}\exp\left(\frac{2}{17}\right)\frac{F_{\mathrm{N}}\left(\sqrt{17}y/3 - 2/\sqrt{17}\right)}{s}, & y \leq 0 \end{cases} \tag{5.99}$$

which is very close to $\left(F_{\mathrm{N}}(y)\right)^2$.

The difference $F_{\mathrm{N}}(y) - F_y(y)$ is symmetric about $y = 0$, given by:

$$F_{\mathrm{N}}(y) - F_y(y) \approx F_{\mathrm{N}}(-|y|) - \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(-s|y| - m)}{s} \tag{5.100}$$

and at $y = 0$ it takes its maximum value, which is:

$$F_{\mathrm{N}}(0) - F_y(0) = \frac{1}{2} - \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(-m)}{s} \tag{5.101}$$

For completeness, the distribution function of $\underline{z} \coloneqq \min\left(\underline{x}_1, \underline{x}_2\right)$ is calculated in a similar manner and is found to be:

$$F_z(z) = 2F_{\mathrm{N}}(z) - F_y(z) \approx \begin{cases} 1 - \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(-sz-m)}{s} & y \geq 0 \\ 2F_{\mathrm{N}}(z) - \exp\left(\frac{m^2}{2}\right)\frac{F_{\mathrm{N}}(sz-m)}{s} & y \leq 0 \end{cases} \tag{5.102}$$

Chapter 6. Knowable moments and their relationship to extremes

6.1 Rationale and definitions

In Chapter 4 we have explained (and illustrated in Digression 4.B) that classical moments beyond order 2 or 3 are *unknowable* and its estimation from data is not feasible. In Chapter 6 we study a new type of moments, the knowable moments (K-moments) which can be reliably estimated for high orders and are particularly useful in analyses of extremes.

Let $\underline{x}$ be a stochastic variable and $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ be IID copies of it, forming a sample. The maximum of all, which is the largest (pth) order statistic, is by definition:

$$\underline{x}_{(p)} \coloneqq \max\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right) \tag{6.1}$$

It is readily obtained that if $F(x)$ is the distribution function of $\underline{x}$ and $f(x)$ its probability density function, then those of $\underline{x}_{(p)}$ are:

$$F^{(p)}(x) = \big(F(x)\big)^p, \qquad f^{(p)}(x) = pf(x)\big(F(x)\big)^{p-1} \tag{6.2}$$

where the former is the product of p instances of $F(x)$ (justified by the IID assumption) while the latter is none other than the derivative of $F^{(p)}(x)$ with respect to x. The *expected maximum of order p* of $\underline{x}$, i.e., the expected value of $\underline{x}_{(p)}$, is therefore

$$\mathrm{E}[\underline{x}_{(p)}] = \mathrm{E}\big[\max\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right)\big] = p\mathrm{E}\left[\left(F\left(\underline{x}\right)\right)^{p-1} \underline{x}\right] \tag{6.3}$$

Likewise, the expected minimum of the p variables is:

$$\mathrm{E}\big[\min\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right)\big] = p\mathrm{E}\left[\left(1 - F\left(\underline{x}\right)\right)^{p-1} \underline{x}\right] = p\mathrm{E}\left[\left(\overline{F}\left(\underline{x}\right)\right)^{p-1} \underline{x}\right] \tag{6.4}$$

We must stress that the variables $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ we consider here are not meant in succession in time and in this respect do not form a stochastic process, but are regarded as an *ensemble* of copies of $\underline{x}$. In other words, the possible dependence in time in a stochastic process is not considered up to now (but will be considered later starting from its impacts on estimation; see section 6.9).

The expected value in (6.3) defines a statistical moment, which, following Koutsoyiannis (2019a), we call *noncentral knowable moment of order p*:

$$K'_p \coloneqq p\mathrm{E}\left[\left(F\left(\underline{x}\right)\right)^{p-1} \underline{x}\right], \qquad p \geq 1 \tag{6.5}$$

The meaning of the term *knowable* will be discussed later, in section 6.4. By generalizing (6.5), we define the following variants of *knowable moment of orders* (p, q), where all definitions are valid for $p \geq q$.

- *Noncentral*:

$$K'_{pq} \coloneqq (p - q + 1)\mathrm{E}\left[\left(F\left(\underline{x}\right)\right)^{p-q} \underline{x}^q\right] \tag{6.6}$$

- *Central*:

$$K_{pq} \coloneqq (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}(\underline{x}-\mu)^{q}\right] \tag{6.7}$$

where μ is the mean of $\underline{x}$, i.e. $\mu \coloneqq \mathrm{E}[\underline{x}_{(p)}] \equiv K'_{11}$.

- *Tail-based (noncentral)*:

$$\overline{K}'_{pq} \coloneqq (p-q+1)\mathrm{E}\left[\left(1-F(\underline{x})\right)^{p-q}\underline{x}^{q}\right] = (p-q+1)\mathrm{E}\left[\left(\overline{F}(\underline{x})\right)^{p-q}\underline{x}^{q}\right] \tag{6.8}$$

- *Hypercentral*:

$$K^{+}_{pq} \coloneqq (p-q+1)\mathrm{E}\left[\left(2F(\underline{x})-1\right)^{p-q}(\underline{x}-\mu)^{q}\right], \qquad p \geq q \tag{6.9}$$

For brevity, will also refer to the knowable moments as *K-moments*. As evident from the above notation for $q = 1, K_p \equiv K_{p1}$ (and likewise for the other variants). The K-moments were introduced in Koutsoyiannis (2019a), even though, according to the current notation, the central and the tail-based moments were not included and the hypercentral moments were called central. The current version of central moments' definition has some advantages with respect to the transformations by shift of origin of the variable (like $\underline{x} - c$; see below). Nonetheless, all K-moment categories are related to each other and the relationships are given in Appendix 6-II. Some characteristic relationships or values for specific values of q or p are summarized in Table 6.1.

Table 6.1 Characteristic relationships or values of different model categories. General equations for any p and q are given in Appendix 6-II.

p, q	Characteristic relationships or values	Eqn. no.
$q = 0$	$K'_{p0} = \overline{K}'_{p0} = K_{p0} = K^{+}_{2p,0} = \mu_0 = \mu'_0 = 1, \qquad K^{+}_{2p+1,0} = 0$	(6.10)
$p = q = 1$	$\mu_1 = K_1 = K^{+}_1 = 0\,, \qquad \mu'_1 = K'_1 = \overline{K}'_1 = \mu$	(6.11)
$q = 1$	$K_p = K'_p - \mu, \qquad K'_p = K_p + \mu$	(6.12)
$q = 2$	$K_{p2} = K'_{p2} - 2\mu K'_{p-1,1} + \mu^2, \qquad K'_{p2} = K_{p2} + 2\mu K_{p-1,1} + \mu^2$	(6.13)
$p = q$	$K_{pp} = K^{+}_{pp} = \mu_p, \qquad K'_{pp} = \overline{K}'_{qq} = \mu'_p$	(6.14)
$p = q + 1$	$\overline{K}'_{q+1,q} = 2K'_{qq} - K'_{q+1,q}, \qquad K'_{q+1,q} = 2\overline{K}'_{qq} - \overline{K}'_{q+1,q}$ $K^{+}_{q+1,q} = 2K_{q+1,q} - 2K_{qq}, \qquad K_{q+1,q} = \frac{1}{2}K^{+}_{q+1,q} + K^{+}_{qq}$	(6.15)
$p = q + 2$	$\overline{K}'_{q+2,q} = 3K'_{qq} - 3K'_{q+1,q} + K'_{q+2,q}, \qquad K'_{q+2,q} = 3\overline{K}'_{qq} - 3\overline{K}'_{q+1,q} + \overline{K}'_{q+2,q}$ $K^{+}_{q+2,q} = 3K_{qq} - 6K_{q+1,q} + 4K_{q+2,q}, \qquad K_{q+2,q} = \frac{3}{4}K^{+}_{qq} + \frac{3}{4}K^{+}_{q+1,q} + \frac{1}{4}K^{+}_{q+2,q}$	(6.16)
$p = q + 3$	$\overline{K}_{q+3,q} = 4K'_{qq} - 6K'_{q+1,q} + 4K'_{q+2,q} - K'_{q+3,q}$ $K_{q+3,q} = 4\overline{K}'_{qq} - 6\overline{K}'_{q+1,q} + 4\overline{K}'_{q+2,q} - \overline{K}'_{q+3,q}$ $K^{+}_{q+3,q} = -4K_{qq} + 12K_{q+1,q} - 16K_{q+2,q} + 8K_{q+3,q},$ $K^{+}_{q+3,q} = -\frac{1}{2}K^{+}_{qq} + \frac{3}{4}K^{+}_{q+1,q} - \frac{1}{2}K^{+}_{q+2,q} + \frac{1}{8}K^{+}_{q+3,q}$	(6.17)

6.2 Theoretical calculation of K-moments

Applying the definition (6.6), we can determine the noncentral K-moments from

$$K'_{pq} = (p-q+1)\int_{-\infty}^{\infty} \big(F(x)\big)^{p-q} x^q\, f(x)\mathrm{d}x \tag{6.18}$$

The calculation may be facilitated if the distribution function is explicitly invertible, through the inverse function $x(F) = F^{-1}\big(F(x)\big)$. In this case

$$K'_{pq} = (p-q+1)\int_{0}^{1} x(F)^q\, F^{p-q}\mathrm{d}F \tag{6.19}$$

or, equivalently,

$$K'_{pq} = \int_{0}^{1} \left(x\left(F^{\frac{1}{p-q+1}}\right)\right)^q \mathrm{d}F \tag{6.20}$$

Likewise, we calculate the central K-moments from:

$$K_{pq} = (p-q+1)\int_{-\infty}^{\infty} \big(F(x)\big)^{p-q} (x-\mu)^q\, f(x)\mathrm{d}x \tag{6.21}$$

or

$$K_{pq} = (p-q+1)\int_{0}^{1} (x(F)-\mu)^q\, F^{p-q}\mathrm{d}F = \int_{0}^{1} \left(x\left(F^{\frac{1}{p-q+1}}\right)-\mu\right)^q \mathrm{d}F \tag{6.22}$$

When analytical calculation is infeasible, numerical calculation of theoretical moments involves no difficulty; thus, the existence of an analytical solution of theoretical moments of a certain distribution should not be regarded as an important criterion for choosing that distribution. The important issue for model fitting is whether the moments are *knowable* or not, in the sense of their estimation from a sample; their theoretical values are always knowable once the distribution parameters have been specified.

6.3 Specific cases of explicit expressions of K-moments

Many customary distribution functions allow convenient theoretical calculation of the tail-based K-moments, while noncentral or central moments may not have an analytical expression. In these cases, we can evaluate the noncentral K-moments at a second step, exploiting the fact that the sequences of noncentral and tail-based K-moments are related through a binomial transform, whose details are contained in Appendix 6-I. As shown in Appendix 6-II, these relationships are:

$$\overline{K}'_{pq} = -\sum_{i=1}^{p-q+1} (-1)^i \binom{p-q+1}{i} K'_{i+q-1,q},$$
$$K'_{pq} = -\sum_{i=1}^{p-q+1} (-1)^i \binom{p-q+1}{i} \overline{K}'_{i+q-1,q} \tag{6.23}$$

and for the particular case $q = 1$:

$$\overline{K}'_p = -\sum_{i=1}^{p} (-1)^i \binom{p}{i} K'_i, \qquad K'_p = -\sum_{i=1}^{p} (-1)^i \binom{p}{i} \overline{K}'_i \tag{6.24}$$

This method is especially useful for a class of distribution functions, among those of Table 2.4, which are most useful in the study of extremes of nonnegative variables. This class comprises the PBF distribution and its special and limiting cases, Pareto, Weibull and exponential, for which general analytical formulae for $\overline{K}'_{pq}$ are possible. These are contained in Table 6.2, which also covers the case of a mixed distribution with discontinuity at the origin, with $P_0 \coloneqq P\{\underline{x} = 0\} = 1 - P_1$ $(0 < P_1 \leq 1)$.

Hypercentral moments can also be determined by analogous relationships based on the binomial transform, which are contained in Appendix 6-II.

We should stress, however, that numerical evaluation of the binomial transform works well for p of the order of several tens, but not of hundreds or thousands. The reason is that the binomial transform of a sequence for order p is equivalent to differencing the sequence p times and it is well known that differencing many times causes numerical errors, which may lead to runaway if p is large. Therefore, analytical relationships or numerical approximations of other types are desirable for large p.

Table 6.2 Analytical results for the tail-based moments of the PBF distribution and its special cases. The distributions are defined on $[0, \infty)$ with a possible discontinuity at zero, with $P_0 \coloneqq P\{x = 0\} = 1 - P_1$ $(0 < P_1 \leq 1)$. All parameters are positive: λ is a scale parameter, ξ is the tail index and ζ is a second shape parameter (lower-tail index).

Name	Tail function, $\overline{F}(x)$	Tail-based moment, $\overline{K}'_{pq}$	Eqn. no.
Pareto-Burr-Feller (PBF)	$P_1\left(1 + \zeta\xi\left(\frac{x}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\zeta\xi}}$	$\lambda^q P_1^{p-q+1} \frac{q}{\zeta} \frac{1}{(\zeta\xi)^{q/\zeta}} \mathrm{B}\left(\frac{p-q+1}{\zeta\xi} - \frac{q}{\zeta}, \frac{q}{\zeta}\right)$	(6.25)
Weibull ($\xi = 0$)	$P_1 \exp\left(-\left(\frac{x}{\lambda}\right)^{\zeta}\right)$	$\lambda^q P_1^{p-q+1} (p-q+1)^{-q/\zeta} \frac{q}{\zeta} \Gamma\left(\frac{q}{\zeta}\right)$	(6.26)
Pareto* ($\zeta = 1$)	$P_1\left(1 + \xi\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}$	$\lambda^q P_1^{p-q+1} \frac{q}{\xi^q} \mathrm{B}\left(\frac{p-q+1}{\xi} - q, q\right)$	(6.27)
Exponential ($\xi = 0, \zeta = 1$)	$P_1 \exp\left(-\frac{x}{\lambda}\right)$	$\frac{\lambda^q P_1^{p-q+1} q!}{(p-q+1)^q}$	(6.28)

Table 6.3 Analytical results for the noncentral K-moments of the Pareto, exponential, Dagum, EV1 and EV2 distributions. The distributions are defined on $[0, \infty)$ with a possible discontinuity at zero, with $P_0 \coloneqq P\{x = 0\} = 1 - P_1$ $(0 < P_1 \leq 1)$. An exception is the EV2 distribution, whose support is the entire real line. All parameters are positive: λ is a scale parameter, ξ is the tail index and ζ is a second shape parameter (lower-tail index).

Case, Tail function $\bar{F}(x)$	K'_{pq}	Eqn. no.
Pareto[1], $P_1 = 1$ $\left(1 + \xi\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}$	$K'_p = \frac{\lambda}{\xi}(p\mathrm{B}(1-\xi, p) - 1), \qquad K'_1 = \frac{\lambda}{1-\xi},$ $K'_2 = \frac{2\lambda}{1-\xi} - \frac{\lambda}{2-\xi}$ $K'_{p2} = \left(\frac{\lambda}{\xi}\right)^2 ((p-1)(\mathrm{B}(1-2\xi, p-1) - 2\,\mathrm{B}(1-\xi, p-1)) + 1)$ $K'_{22} = \frac{2\lambda^2}{(1-2\xi)(1-\xi)}$	(6.29)
Pareto with discontinuity, $P_1 < 1$ $P_1\left(1 + \xi\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}$	$K'_p = \frac{\lambda}{\xi}\left(pP_1^{\xi}\mathrm{B}_{P_1}(1-\xi, p) - 1 + (1-P_1)^p\right), \qquad K'_1 = \frac{\lambda P_1}{1-\xi}$ $K'_{p2} = \left(\frac{\lambda}{\xi}\right)^2 \Big((p-1)P_1^{\xi}\left(P_1^{\xi}\,\mathrm{B}_{P_1}(1-2\xi, p-1) - 2\,\mathrm{B}_{P_1}(1-\xi, p-1)\right) - (1-P_1)^{p-1} + 1\Big)$ $K'_{22} = \frac{2\lambda^2 P_1}{(1-2\xi)(1-\xi)}$	(6.30)
Exponential[1], $P_1 = 1$ $\exp\left(-\frac{x}{\lambda}\right)$	$K'_p = \lambda\, H_p, \qquad K'_1 = \lambda, \qquad K'_2 = \frac{3\lambda}{2}$ $K_{p2} = \left(\left(H_{p-1}\right)^2 + H^{(2)}_{p-1}\right)\lambda^2, \qquad K'_{22} = 2\lambda^2$	(6.31)
Exponential with discontinuity[2], $P_1 < 1$ $P_1 \exp\left(-\frac{x}{\lambda}\right)$	$K'_p = \lambda\, p\, P_1\ {}_3F_2\,(1,1,1-p; 2,2; P_1)$ $K'_{pq} = \lambda\,(p-q+1)\,P_1\left({}_3F_2\,(1,1,q-p; 2,2; P_1)\right.$ $\left. - (q-1)\ {}_4F_3\,(1,1,q-p; 2,2; P_1)\right)$ $K'_1 = \lambda\, P_1, \qquad K'_2 = \lambda P_1(2 - P_1/2), \qquad K'_{22} = \lambda^2 P_1$	(6.32)
Dagum, $P_1 = 1$ $1 - \left(1 + \frac{\xi}{\zeta}\left(\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)^{-\zeta}$	$K'_p = \lambda(\xi/\zeta)^{\xi} p\zeta\ \mathrm{B}(1-\xi, p\zeta + \xi)$ $K'_{pq} = \lambda^q(\xi/\zeta)^{q\xi}(p-q+1)\zeta\ \mathrm{B}(1-q\xi, (p-q+1)\zeta + q\xi)$	(6.33)
EV2[3], $1 - \exp\left(-\xi\left(\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)$	$K'_p = \lambda(\xi p)^{\xi}\Gamma(1-\xi)$ $K'_{pq} = \lambda^q(\xi(p-q+1))^{q\xi}\Gamma(1-q\xi)$	(6.34)
EV1[4], $1 - \exp\left(-\mathrm{e}^{-\frac{x}{\lambda}}\right)$	$K'_p = \lambda(\ln p + \gamma)$ $K'_{p2} = \lambda^2\left((\ln(p-1) + \gamma)^2 + \frac{\pi^2}{6}\right)$	(6.35)

[1] K'_{p2} of the exponential and the Pareto case can serve as a basis for the calculation of the K'_p of the special case $\zeta = 1/2$ of the Weibull and PBF distributions, respectively.

[2] ${}_aF_b$ is the generalized hypergeometric function.

[3] The EV2 distribution is derived from the Dagum distribution for $\zeta \to \infty$.

[4] The EV1 distribution is derived from the EV2 distribution by replacing x with $x - 1/\xi$, and letting $\xi \to 0$.

For small values of q, the exponential and Pareto distributions admit explicit expressions also for the noncentral (and central) K-moments, which become simple if $P_1 = 1$. These are contained in Table 6.3. Another distribution, which admits analytical expressions of noncentral K-moments for $P_1 = 1$ is the Dagum distribution and its special cases, the Extreme Value type I and II distributions. These are also contained in Table 6.3. Note that the parameterization of the mathematical expressions of the distributions in Table 6.2 and Table 6.3 may differ from those in Table 2.4 as here we have given emphasis on showing the connection of distributions to each other (e.g. the derivation of the exponential distribution as the limit of the Pareto distribution for $\xi \to 0$.

It is noted that the mixed distribution with a discontinuity at the lower bound (typically $x = 0$), is useful to examine, and is thus included in in Table 6.2 and Table 6.3, for two reasons. First, there are natural processes, such as rainfall and occasionally streamflow, in which the probability dry ($P_0 \coloneqq P\{\underline{x} = 0\} = 1 - P_1$) is nonzero (and thus $0 < P_1 \leq 1$). Second, in several analyses we are interested about values of $\underline{x}$ above a threshold x_0 and in this case the distribution becomes discontinuous at x_0.

If the discontinuity at the origin is quantified by $P_1 \coloneqq P\{\underline{x} > 0\}$ and we denote the K-moments of the discontinuous and continuous distribution with and without a superscript '*', respectively, then it is easy to see that the tail-based K-moments in the two cases are related by:

$$\overline{K}_p'^* = P_1^p \overline{K}_p' \tag{6.36}$$

Based on this, we find in Appendix 6-III that the noncentral K-moments in the two cases are related though a Bernoulli transform, whose properties are given in Appendix 6-I. Namely, $K_p'^*$ is the Bernoulli transform of K_p' with parameter P_1:

$$K_p'^* = \sum_{l=1}^{p} \binom{p}{l} K_l' P_1^l (1-P_1)^{p-l} = (1-P_1)^p \sum_{l=1}^{p} \binom{p}{l} K_l' \left(\frac{P_1}{1-P_1}\right)^l \tag{6.37}$$

Unlike the binomial transform, the Bernoulli transform does not entail numerical problems as it reflects summation of positive quantities rather than differences thereof. Yet a simple approximation is proposed, which can be useful in some cases:

$$K_p'^* = \begin{cases} P_1 K_1' p^b & p \leq \dfrac{2}{P_1} \\ K_{p'}', & p > \dfrac{2}{P_1} \end{cases}, \qquad p' = P_1 p, \qquad b = \frac{\ln(K_2'/K_1' P_1)}{\ln(2/P_1)} \tag{6.38}$$

The rationale of this approximation is given in Appendix 6-III.

While the PBF and the Dagum distributions, as well as their special and limiting cases discussed in this section, are quite general and convenient in their use, other customary distributions such as normal, lognormal and gamma, are in common use for several hydrometeorological processes. These distributions do not admit analytical relationships of K-moments, but good approximations are discussed in section 6.14.

6.4 Relationship of knowable and classical moments

The classical moments can be recovered as special cases of K-moments:

$$K'_{pp} \equiv \mu'_p, \qquad K_{pp} \equiv \mu_p \tag{6.39}$$

While the classical moments are extremely useful as theoretical concepts, and their values can be derived rather easily if the distribution function is specified, their estimation from samples is problematic if we go to moment order higher than 2 to 3. This has been stressed in the title of the article by Lombardo et al. (2014): "*Just two moments*", while for the same reason in Koutsoyiannis (2019a) classical moments beyond that order have been termed *unknowable* (see Digression 4.B).

As we will see in section 6.9, the contrary happens with K-moments. They can be estimated reliably even for a very high order p (hence their name *knowable*), provided that the order q remains low.

Not only can K-moments be estimated for high orders p, but they can also predict the value of the estimates of the classical moments. Next, we will derive this prediction for the noncentral classical moments in detail. We note that this prediction does not coincide with the true value of the classical moment. This may sound paradoxical as it is known that, for order p however large, $\hat{\underline{\mu}}'_p$,is an unbiased estimator of μ'_p. In practice, however, the convergence of $\hat{\underline{\mu}}'_p$ to μ'_p is very slow, while the K-moments can give us an indication of what we can anticipate for the value of $\hat{\mu}'_p$, which does not coincide with μ'_p. In this respect, by examining the moment estimators, we will establish relationships between K- and classical moments broader and more essential than (6.39).

For large p the classical moment estimator will give:

$$\hat{\underline{\mu}}'_p = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i^p \approx \frac{1}{n}\underline{x}_{(n)}^p \tag{6.40}$$

This is related to the well-known mathematical fact that the maximum norm is the limit of the p-norm as $p \to \infty$ as explained in Digression 4.B. Taking expected values in (6.40), we find:

$$\mathrm{E}[\hat{\underline{\mu}}'_p] \approx \frac{1}{n}\mathrm{E}\left[\underline{x}_{(n)}^p\right] = \frac{K'_{n+p-1,p}}{n} \tag{6.41}$$

and since $\mathrm{E}[\hat{\underline{\mu}}'_p] = \mu'_p$, for large p:

$$K'_{n+p-1,p} = n\mu'_p \tag{6.42}$$

or, equivalently, for large q and for $p = n - q + 1$:

$$K'_{pq} \approx (p - q + 1)\mu'_q \tag{6.43}$$

from which for $p = q$, we recover (6.39), in this case holding precisely.

However, if $\hat{\mu}'_p$ is estimated from a sample and p is large, we do not anticipate that $\hat{\mu}'_p$ would be close to the true classical moment μ'_p. Rather, because of (6.40), we can anticipate that:

$$\hat{\mu}'_p \approx \frac{\left(x_{(n)}\right)^p}{n} \approx \frac{(K'_{n1})^p}{n} \tag{6.44}$$

Likewise, if the classical moment is estimated as the average of m different (and independent) samples, each of size n, then:

$$\frac{1}{m}\sum_{j=1}^{m} \hat{\mu}'_{p_j} \approx \frac{1}{m}\hat{\mu}'_{p_{(m)}} \approx \frac{1}{mn} x^p_{(n)_{(m)}} \approx \frac{1}{mn} x^p_{(nm)} \tag{6.45}$$

and thus, we can anticipate that our so produced estimate will be:

$$\hat{\mu}'_p \approx \frac{\left(x_{(mn)}\right)^p}{mn} \approx \frac{\left(K'_{mn,1}\right)^p}{mn} \tag{6.46}$$

To establish a more general formula that will give us the anticipated value for our estimate $\hat{\mu}'_p$, applicable for small and large p, we observe that, because of (6.41), the ratio $mn\ \mu'_p\ /K'_{mn+p-1,p}$ is 1 for large p, and thus multiplying the rightmost part of (6.44) by it will not have any effect if p is large, while having the desired effect if p is small and particularly if p = 1. Thus, our formula becomes

$$\hat{\mu}'_p \approx \frac{\left(K'_{mn,1}\right)^p}{K'_{mn+p-1,p}} \mu'_p \tag{6.47}$$

and is valid either for the average estimate from many samples (m > 1) or for just one sample of size n (m = 1). Indeed, for p = 1 and for any m and n, our formula (6.47), yields:

$$\hat{\mu}'_1 = \frac{\left(K'_{mn,1}\right)^1}{K'_{mn,1}} \mu'_1 = \mu'_1 \tag{6.48}$$

while for large p we recover (6.44) as already explained.

Digression 6.A: Example on the relationship of knowable and classical moments

We illustrate the relationship of K- and classical moments using synthetic samples with size n = 100 from the exponential distribution with lower bound zero and scale parameter 1. This distribution has simple expressions of its moments, i.e.,

$$\mu'_p = K'_{pp} = p!, \qquad K'_p = H_p \tag{6.49}$$

(see Table 6.2 and Table 6.3). For q > 2 the K-moment K'_{pq} does not have a closed analytical expression but its calculation can be easily made by numerical integration.

Figure 6.1 (left panel) shows comparison of the theoretical (true) moments, of orders 1 to 100, of the distribution with the empirical estimations from a single sample as well as from a set of 1000 samples (by averaging 1000 estimates). The single sample estimates deviate from (are lower than) the true moments for $p \geq 3$ and the deviation becomes one order of magnitude as p approaches n = 100. We will refer to this deviation as *slow convergence bias*, because theoretically speaking there is no bias per se, according to the bias definition in section 4.3. From a practical point of view, the moment estimates are not the same thing as the true moments, despite the theoretical guarantee that the estimates are unbiased. Even the average of 1000 estimates from different synthetic samples deviates substantially from the true moments for order p > 10. On the other hand, equation (6.47) captures very well the behaviour of the estimates for both m = 1 and m = 1000.

Figure 6.1 (right panel) provides additional information on the same exercise and in particular on the terms appearing in equation (6.47) for the case of a single sample. First it shows that the true and estimated $(K'_{n1})^p$ agree, which indicates that even for $p = n$ the estimate of K'_{n1} is close to the true value (where the estimator will be discussed later, in section 6.9). Second, the graphs show the variation of the term $K'_{n+p-1,p}$ which appears in the denominator of (6.47). For small p, $K'_{n+p-1,p} \approx (K'_{n1})^p$, while for large p, $K'_{n+p-1,p} \approx \mu'_p$. As a consequence, by virtue of (6.47), for large p, $\hat{\mu}'_p \approx (K'_{n1})^p$, or $(\hat{\mu}'_p)^{1/p} \approx K'_{n1}$. This is seemingly different from equation (6.44), which gives $(\hat{\mu}'_p)^{1/p} \approx K'_{n1}/n^{1/p}$; however for large p (approaching $n = 100$), $n^{1/p} \approx 1$ and therefore the two equations become consistent.

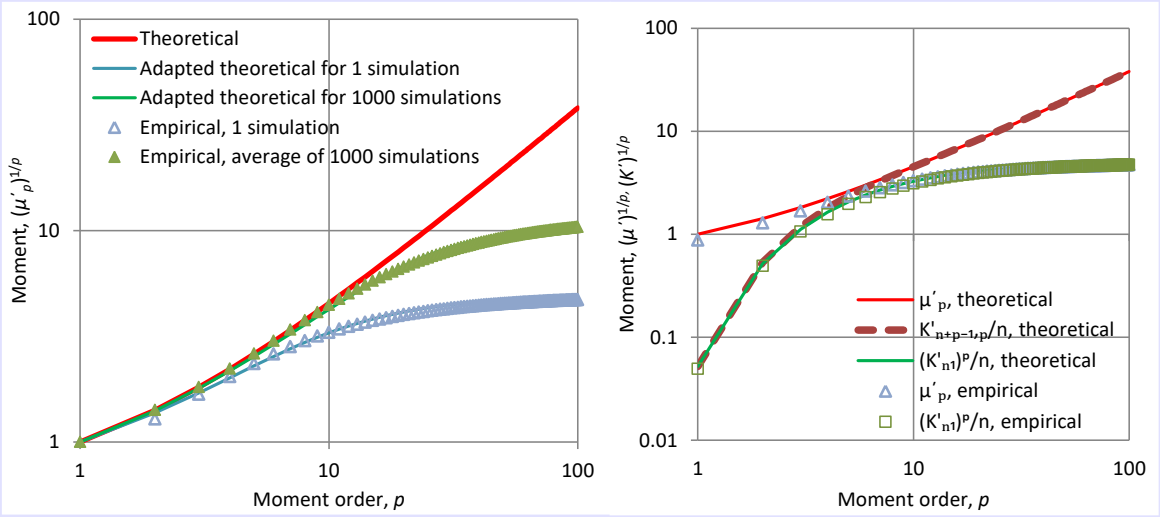

Figure 6.1 (left) Comparison of the estimates of classical noncentral moments from 1 and 1000 independent samples from the exponential distribution to (a) the true (theoretical) moments and (b) to the values determined by equation (6.47) (adapted theoretical). **(right)** Additional information of the terms appearing in equation (6.47).

As a general conclusion, the classical moment estimators do not practically estimate classical moments but hybrid quantities involving both classical and K-moments, as implied by equation (6.47). The question arises then, when high-order moments are of interest, whether it is useful to involve classical moment estimates in statistical calculations, now knowing, from equation (6.47), what they really represent or it is better to use merely K-moments. To answer this question, we need to examine not the averages of estimates but the ranges in which they vary.

This information is provided by Figure 6.2 which depicts moments from 100 simulated samples with length $n = 2000$ from lognormal distribution LN(0,1). K-moments, specifically K'_{p1} in the left panel and K'_{p2} in the right panel, are compared to classical moments $\mu'_{p'} \approx K'_{p',p'}$. For facilitating visual comparison, the order p' of the classical moments was determined so that the estimate $\hat{\mu}'_{p'}$ coincide with that of K'_{p1} and K'_{p2} in the left and right panel, respectively. This entails a specific relationship between p' and p, which was determined numerically and is shown in the figure caption.

A fist observation in Figure 6.2 is that true and estimated K-moments coincide, while estimates of classical moments differ substantially from the true ones. A second observation is that the 95% prediction limits, determined from the simulation, are much wider in the classical moments than in the K moments. The two prediction intervals approach each other only when p approaches $n = 2000$ (note though that the value of p for the classical moments when $p' = 2000$ is much less than 2000, close to 160. All in all, the plots show that we do not gain anything by using high order classical moments as K-moments are much more reliably estimated.

The breadths of the prediction intervals are depicted in Figure 6.3 in terms of the ratios of maximum to minimum prediction limit. Even without considering the slow convergence bias, which, as illustrated in Figure 6.2, is very high for the classical moments, the broad prediction

intervals disfavour again the use of the classical moments. We observe in Figure 6.3 that for the third classical moment, the prediction limits have a ratio of 1.77 (higher to lower, without considering the slow convergence bias). The same ratio appears in K-moments at a moment of order as high as 250 (for q = 1) or 100 (for q = 2). Therefore, instead of estimating the third classical moment, it is safer to estimate and use K-moments of much higher order and specifically up to $p = n/10$ for K'_{p1} and up to $p = n/20$ for K'_{p2}.

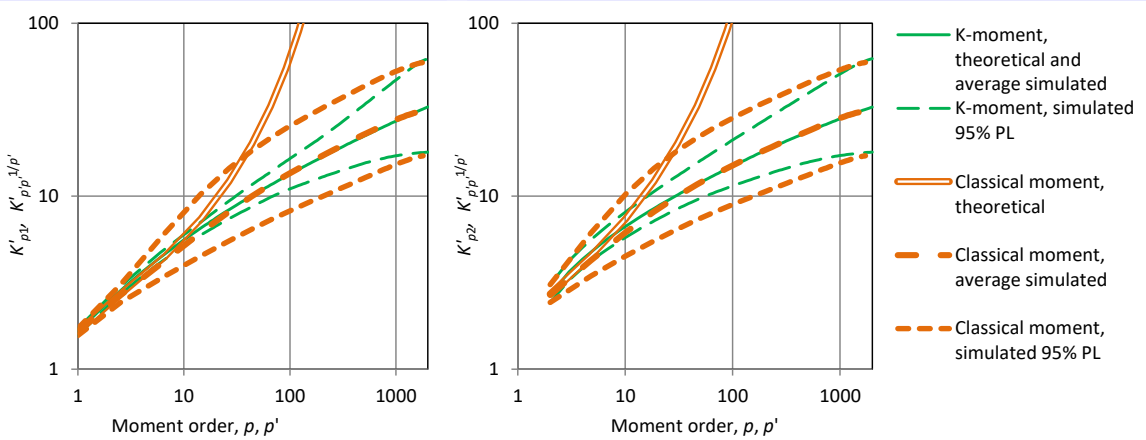

Figure 6.2. Comparison of K- and classical moment estimates for the lognormal distribution LN(0,1) from 100 simulations, each with n = 2000. Order q of K-moments is q = 1 for the left panel and q = 2 for the right panel. The order p' of the classical moments (see explanation in text) is determined from $p' = q + 220 \ln(1 + (p^3 - q^3) / (500 (q + 1))$. PL stands for prediction limits.

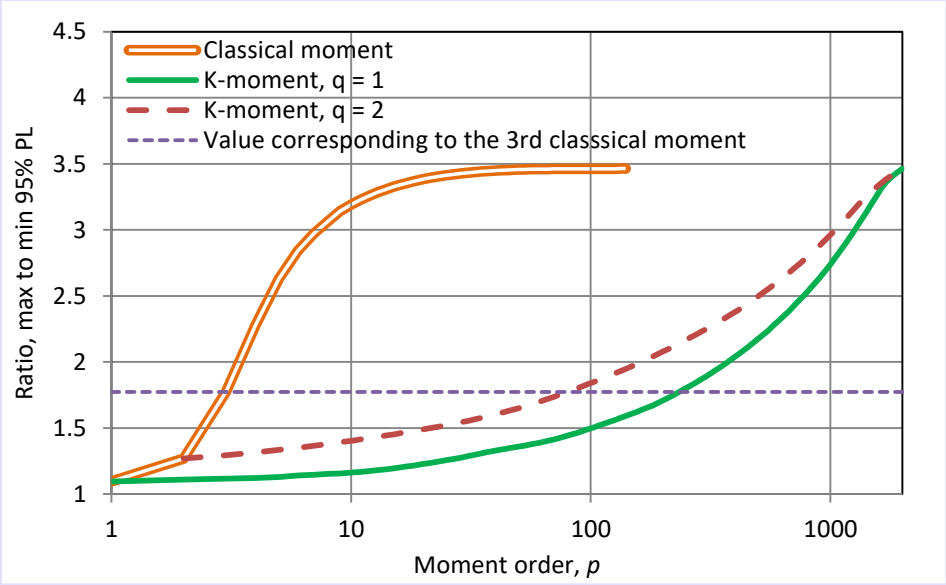

Figure 6.3 Ratios of maximum to minimum prediction limits, as a function of moment order p, for the simulation experiment of Figure 6.2.

6.5 Relationship of K-moments with order statistics and maxima

From equation (6.3) it is evident that the noncentral K-moment of orders (p, 1) is none other than the expected value of the largest order statistic of a sample of $\underline{x}$ of size p. More generally, excepting tail-based K-moments, which represent expected values of minima, K-moments of all other categories represent expected values of maxima. To see this, we first consider the power transformation $\underline{z}' = \underline{x}^q$ and further assume that it is monotonic,

which is the case if q is an odd number or if $\underline{x}$ is nonnegative. Then, it is readily seen that $F_{z'}(z') = P\{\underline{z}' \le z'\} = P\{\underline{x} \le z'^{1/q}\} = F(z'^{1/q})$. Consequently,

$$\begin{aligned} K'_{pq} &= (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}\underline{x}^q\right] = (p-q+1)\mathrm{E}\left[\left(F(\underline{z}'^{1/q})\right)^{p-q}\underline{z}'\right] \\ &= (p-q+1)\mathrm{E}\left[\left(F_{z'}(\underline{z}')\right)^{p-q}\underline{z}'\right] = K'_{z'_{p-q+1,1}} \\ &= \mathrm{E}\left[\max\left(\underline{x}_1^q, \underline{x}_2^q, \dots, \underline{x}_{p-q+1}^q\right)\right] = \mathrm{E}\left[\underline{x}^q_{(p-q+1)}\right] \end{aligned} \tag{6.50}$$

This means that the noncentral K-moment K'_{pq} of $\underline{x}$ is identical to the expected maximum of order $p-q+1$ of $\underline{z}' = \underline{x}^q$.

Furthermore, we consider the transformation $\underline{z} = \left(\underline{x}-\mu\right)^q$, assuming that q is an odd number, so that it be monotonic and thus $F_z(z) = P\{\underline{z} \le z\} = P\{\underline{x} \le z^{1/q}+\mu\} = F(z^{1/q}+\mu)$. Consequently,

$$\begin{aligned} K_{pq} &= (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}\left(\underline{x}-\mu\right)^q\right] = (p-q+1)\mathrm{E}\left[\left(F(\underline{z}^{1/q}+\mu)\right)^{p-q}\underline{z}\right] \\ &= (p-q+1)\mathrm{E}\left[\left(F_z(\underline{z})\right)^{p-q}\underline{z}\right] = K'_{z_{p-q+1,1}} \\ &= \mathrm{E}\left[\max\left(\left(\underline{x}_1-\mu\right)^q, \left(\underline{x}_2-\mu\right)^q, \dots, \left(\underline{x}_{p-q+1}-\mu\right)^q\right)\right] \\ &= \mathrm{E}\left[\left(\max\left(\underline{x}_1, \dots, \underline{x}_{p-q+1}\right)-\mu\right)^q\right] \end{aligned} \tag{6.51}$$

which means that the central K-moment K_{pq} of $\underline{x}$ is identical to the expected maximum of order $p-q+1$ of $\underline{z} = \left(\underline{x}-\mu\right)^q$.

The above properties should also hold asymptotically, for large p, even if the transformations z or z' are not monotonic (e.g. for even q) in any case of positively skewed distribution. Interestingly, for a symmetric distribution, a property analogous to (6.51) holds for the hypercentral moments with even q. Indeed, in this case and for $z \ge 0$, $F_z(z) = P\{\underline{z} \le z\} = P\{-z^{1/q}+\mu \le \underline{x} \le z^{1/q}+\mu\} = F\left(z^{1/q}+\mu\right) - F(-z^{1/q}+\mu)$ and due to symmetry $F_z(z) = 2F\left(z^{1/q}+\mu\right) - 1$. Thus,

$$\begin{aligned} K^+_{pq} &\coloneqq (p-q+1)\mathrm{E}\left[\left(2F\left(\underline{x}\right)-1\right)^{p-q}\left(\underline{x}-\mu\right)^q\right] \\ &= (p-q+1)\mathrm{E}\left[\left(2F\left(\underline{z}^{1/q}+\mu\right)-1\right)^{p-q}\underline{z}\right] = \\ &= (p-q+1)\mathrm{E}\left[\left(F_z(\underline{z})\right)^{p-q}\underline{z}\right] = K'_{z_{p-q+1,1}} \\ &= \mathrm{E}\left[\max\left(\left(\underline{x}_1-\mu\right)^q, \left(\underline{x}_2-\mu\right)^q, \dots, \left(\underline{x}_{p-q+1}-\mu\right)^q\right)\right] \end{aligned} \tag{6.52}$$

which means that the hypercentral K-moment K^+_{pq} of a stochastic variable $\underline{x}$ with symmetrical distribution for q even is identical to the expected maximum of order $p-q+1$ of $\underline{z}$. In contrast, for q odd, the hypercentral K-moment K^+_{pq} will obviously be zero.

Now coming to the tail-based moments, assuming that the power transformation $\underline{z}' = \underline{x}^q$ is monotonic (i.e., q is an odd number or $\underline{x}$ is nonnegative), we will have $\overline{F}_{z'}(z') = P\{\underline{z}' > z'\} = P\{\underline{x} > z'^{1/q}\} = \overline{F}(z'^{1/q})$; hence:

$$\begin{aligned}\overline{K}'_{pq} &= (p-q+1)\mathrm{E}\left[\left(\overline{F}(\underline{x})\right)^{p-q}\underline{x}^q\right] = (p-q+1)\mathrm{E}\left[\left(\overline{F}(\underline{z}'^{1/q})\right)^{p-q}\underline{z}'\right] \\ &= (p-q+1)\mathrm{E}\left[\left(\overline{F}_{z'}(\underline{z}')\right)^{p-q}\underline{z}'\right] = \overline{K}'_{z'_{p-q+1,1}} \\ &= \mathrm{E}\left[\min\left(\underline{x}_1^q, \underline{x}_2^q, \dots, \underline{x}_{p-q+1}^q\right)\right]\end{aligned} \tag{6.53}$$

In section 6.9 we will discuss a further connection of K-moments and order statistics which enables estimation of K-moments from an ordered sample.

6.6 Relationship of K-moments with L-moments and probability weighted moments

L-moments (Hosking et al.,1985a,b; Hosking, 1990) represent a very useful and popular moment category as, contrary to classical moments, have unbiased estimators for high orders. According to their definition, L-moments are linear combinations of order statistics. Naturally then, L-moments are connected by linear relationships with K-moments for the specific case of q = 1. The relationships for the first four orders p are given in Koutsoyiannis (2019a).

L-moments are also related to probability weighted moments (PWM), which had been introduced earlier than the former (Greenwood et al., 1979). Actually, the latter are more directly defined than the former and therefore are preferable—and in fact the estimation of L-moments is made from that of PWM. In particular, the definition of the PWM involves three different orders, p, s and q and is $\beta_{p,s,q} \coloneqq \mathrm{E}\left[\left(F(\underline{x})\right)^p\left(1-F(\underline{x})\right)^s\underline{x}^q\right]$. However, only the case q = 1 has been studied and the most common form, which is also used for estimation of L-moments, $\beta_p \coloneqq \beta_{p,0,1} = \mathrm{E}\left[\left(F(\underline{x})\right)^p\underline{x}\right]$. This is proportional to the noncentral K- moment for q = 1:

$$K'_p = p\beta_{p-1} \tag{6.54}$$

The small difference (multiplication by p) makes the K-moments more intuitive. First, it makes K-moments increasing functions of p, which is consistent to the behaviour of classical moments. Also, it links them directly to the largest order statistic as discussed above (see also Koutsoyiannis, 2019a).

6.7 Basic statistical characteristics based on K-moments

It is useful to note that the central moment K_{pq} of $\underline{x}$ is identical to the noncentral K-moment of $\underline{y} \coloneqq \underline{x} - \mu$ and, in particular, for q = 1, K_p is the expected maximum of order p of $\underline{y} = \underline{x} - \mu$. Observing that the distribution of $\underline{y}$ is $F_y(y) = F(y+\mu)$ and the density is $f_y(y) = f(y+\mu)$, we can easily verify the above property, which we can denote by the following notation:

$$K_{pq}[\underline{x}] = K'_{pq}[\underline{x} - \mu] \tag{6.55}$$

Useful (and easy to demonstrate) properties of central K-moments are: (a) their invariance to a shift of origin and (b) their homogeneity in terms of multiplication by a scalar. Both these properties are reflected in the relationship:

$$K_{pq}[a + b\underline{x}] = b^q K_{pq}[\underline{x}] \tag{6.56}$$

It is thus evident that for any p, p', q, the ratio

$$\frac{K_{pq}[a + b\underline{x}]}{K_{p'q}[a + b\underline{x}]} = \frac{K_{pq}[\underline{x}]}{K_{p'q}[\underline{x}]} \tag{6.57}$$

does not depend on a and b. In other words, this ratio is invariant under linear transformation of the variable. This makes meaningful the standardization of the variable $\underline{x}$ either by K_2 if $q = 1$ or by $K_{22}^{1/2}$ if $q = 2$. The standardized variables:

$$\underline{w}_1 := \frac{\underline{x} - \mu}{K_2[\underline{x}]}, \qquad \underline{w}_2 := \frac{\underline{x} - \mu}{\left(K_{22}[\underline{x}]\right)^{1/2}} \tag{6.58}$$

will have $K_2[\underline{w}_1] = K_{22}[\underline{w}_2] = 1$ and thus

$$K_p[\underline{w}_1] = K'_p[\underline{w}_1] = \frac{K_p[\underline{x}]}{K_2[\underline{x}]}, \qquad K_{p2}[\underline{w}_2] = K'_{p2}[\underline{w}_2] = \frac{K_{p2}[\underline{x}]}{K_{22}[\underline{x}]} \tag{6.59}$$

Table 6.4 Typical marginal statistical characteristics of distributions using different moment categories. Characteristics based on classical moments are also given as Option 3, while in Option 1 those based on L-moments are shown, with λ_p denoting the L-moment of order p.

Characteristic	Order p	Option 1	Option 2	Option 3*
Location	1	$K'_{11} = \mu$ (the classical mean)		
Variability	2	$K_{21}^+ = 2K_{21} = 2(K'_{21} - \mu) = 2\lambda_2$	$K_{22}^+ = K_{22} = \mu_2 = \sigma^2$ (the classical variance)	
Skewness (dimensionless)	3	$\frac{K_{31}^+}{K_{21}^+} = 2\frac{K_{31}}{K_{21}} - 3 = \frac{\lambda_3}{\lambda_2}$	$\frac{K_{32}^+}{K_{22}^+} = 2\frac{K_{32}}{K_{22}} - 2$	$\frac{K_{33}}{K_{22}^{3/2}} = \frac{\mu_3}{\sigma^3}$
Kurtosis (dimensionless)	4	$\frac{K_{41}^+}{K_{21}^+} = 4\frac{K_{41}}{K_{21}} - 8\frac{K_{31}}{K_{21}} + 6 = \frac{4}{5}\frac{\lambda_4}{\lambda_2} + \frac{6}{5}$	$\frac{K_{42}^+}{K_{22}^+} = 4\frac{K_{42}}{K_{22}} - 6\frac{K_{32}}{K_{22}} + 3$	$\frac{K_{44}}{K_{22}^2} = \frac{\mu_4}{\sigma^4}$

Within the framework of K-moments and according to the rule of thumb "Just two [classical] moments" we may assume that the power of $\underline{x}$, i.e. q, should be taken $q = 1$ or 2, while p can be however large. In this manner, for $p > 1$ we have two alternative options to define statistical characteristics related to moments of the distribution. As discussed in Digression 2.E, the most customary characteristics with respect to classical moments are those for orders p from 1 to 4, which indicate the location, variability, skewness and kurtosis of a distribution. Similar is the meaning if we replace classical moments with K-moments. These characteristics are shown in Table 6.4 for both moment categories and for three options. Options 1 and 2 correspond to K-moments for $q = 1$ and 2, respectively,

while Option 3 corresponds to the classical moments. For the definitions of the characteristics of Options 1 and 2, the hypercentral variant of K-moments is used as it is more suitable for this purpose (see Koutsoyiannis, 2019a). The synoptic characteristics of Table 6.4 are useful quantified indications of the general behaviour of a stochastic variable and occasionally can also be used for the fitting of a distribution. However, as we will see in section 6.15, other fitting methods involving higher order K-moments may be preferable if we are particularly interested about extremes.

6.8 The K-moments as the Laplace transform of the quantile function

The K-moment definition in general implies that the order q should be an integer in order to ensure that x^q is a real number. Also, so far, the order p was assumed to be an integer. However, if we specify q = 1 we can generalize that p can be a real number. Then we can write the K-moment as a function of the real variable p. Adapting (6.19) we write:

$$K'(p) \coloneqq K'_p = p \int_0^1 x(F)\, F^{p-1} \mathrm{d}F \tag{6.60}$$

Now, given that F takes on values in the interval [0, 1] and in order to obtain a variable spanning all nonnegative real numbers, we use the transformations :

$$F = \mathrm{e}^{-s}, \qquad g(s) = x(\mathrm{e}^{-s}), \qquad G(p) = \frac{K'(p)}{p} \tag{6.61}$$

Then we can write:

$$G(p) = \int_0^\infty g(s)\, e^{-sp} \mathrm{d}s = \mathcal{L}\{g\}(p) \tag{6.62}$$

where $\mathcal{L}\{g\}(p)$ is the Laplace transform of $g(s)$. This can help in determining F by inverting the transform:

$$g(s) = \mathcal{L}^{-1}\{G\}(s), \qquad x(F) = g(-\ln F), \qquad F(x) = \mathrm{e}^{-g^{-1}(x)} \tag{6.63}$$

6.9 Unbiased estimators of noncentral and tail knowable moments

The quantities $\left(F\left(\underline{x}\right)\right)^{p-q} \underline{x}^q$, whose expectations define K-moments, are estimated from a sample without using powers of $\underline{x}$ higher than q (which could be conveniently assumed to be 1 or 2), thus making the estimation reliable. The literature provides unbiased estimators on probability weighted moments and L-moments, both corresponding to q = 1, based on the analysis by Landwehr et al. (1979). Here we adapt those results for the K-moments.

As a general remark, the estimation of K-moments relies on that of quantities $\left(F\left(\underline{x}\right)\right)^{p-q}$. This becomes possible if we involve order statistics, i.e., if we arrange the sample in ascending order. From the definition of the different variants of K-moments we can think of estimators of the form:

$$\underline{\widehat{K}}'_{pq} = \frac{p-q+1}{n}\sum_{i=1}^{n}\left(F\left(\underline{x}_{(i:n)}\right)\right)^{p-q}\underline{x}^{q}_{(i:n)}$$

$$\underline{\widehat{K}}_{pq} = \frac{p-q+1}{n}\sum_{i=1}^{n}\left(F\left(\underline{x}_{(i:n)}\right)\right)^{p-q}\left(\underline{x}_{(i:n)}-\hat{\mu}\right)^{q} \tag{6.64}$$

$$\underline{\widehat{K}}^{+}_{pq} = \frac{p-q+1}{n}\sum_{i=1}^{n}\left(2F\left(\underline{x}_{(i:n)}\right)-1\right)^{p-q}\left(\underline{x}_{(i:n)}-\hat{\mu}\right)^{q}$$

where $\underline{x}_{(i:n)}$ is the ith element of a sample of $\underline{x}$ of size n, sorted in ascending order; the maximum of all $\underline{x}_i$ is denoted as $\underline{x}_{(n)} \equiv \underline{x}_{(n:n)}$. It is stressed that the ordering of the sample is meant in terms of $\underline{x}$ and not $\underline{x}^q$. More precisely, $\underline{x}^q_{(i:n)} \coloneqq \left(\underline{x}_{(i:n)}\right)^q$, which can be different from $\left(\underline{x}^q\right)_{(i:n)}$, except if q is an odd number or if $\underline{x}$ is strictly nonnegative.

While $F\left(\underline{x}_{(i:n)}\right)$ has a simple estimator, which depends on i and the sample size n (and not on the specific form of $F(x)$; see section 5.6), this estimator is not necessarily optimal if $F\left(\underline{x}_{(i:n)}\right)$ (or $2F\left(\underline{x}_{(i:n)}\right)-1$) is raised to a power and is multiplied by $\underline{x}_{(i:n)}$ (or $\underline{x}_{(i:n)}-\hat{\mu}$) raised to another power. In the subsections that follow we will examine better estimators.

Based on the properties of order statistics (section 4.12) and the above discourse, it is reasonable to construct estimators in which $F\left(\underline{x}_{(i:n)}\right)$ do not depend on $\underline{x}_{(i:n)}$ but only on i and n. In this case the estimator of $(p/n)\left(F\left(\underline{x}_{(i:n)}\right)\right)^{p-1}$ is no longer a stochastic variable but a regular variable depending on i, n and p. If we denote it as b_{inp} then an estimator of the noncentral moment K'_{pq} will be:

$$\underline{\widehat{K}}'_{pq} = \sum_{i=1}^{n} b_{i,n,p-q+1}\,\underline{x}^{q}_{(i:n)} \tag{6.65}$$

which for $q = 1$ takes the form:

$$\underline{\widehat{K}}'_{p} = \sum_{i=1}^{n} b_{inp}\,\underline{x}_{(i:n)} \tag{6.66}$$

We prove in Appendix 6-IV that for a random sample and $q = 1$ the estimator in (6.65) is unbiased if we choose:

$$b_{inp} = \begin{cases} 0, & i < p \\ \dfrac{p}{n}\,\dfrac{\Gamma(n-p+1)}{\Gamma(n)}\,\dfrac{\Gamma(i)}{\Gamma(i-p+1)}, & i \geq p \geq 0 \end{cases} \tag{6.67}$$

where p can be any positive number $\leq n$(usually, but not necessarily, integer). It can be easily verified that:

$$\sum_{i=1}^{n} b_{inp} = 1 \tag{6.68}$$

which is a necessary condition for unbiasedness. Furthermore, for $p = 1$, $b_{in1} = 1/n$, while for $p = 2$, the quantity $(n/2)b_{in2}$ can be regarded as the estimate of $F\left(x_{(i:n)}\right)$, i.e.:

$$\hat{F}\left(x_{(i:n)}\right) = \frac{i-1}{n-1} \tag{6.69}$$

This has the interesting property of symmetry:

$$\hat{F}\left(x_{(n+1-i:n)}\right) = 1 - \hat{F}\left(x_{(i:n)}\right), \qquad \left(2\hat{F}\left(x_{(n+1-i:n)}\right) - 1\right) = -\left(2\hat{F}\left(x_{(i:n)}\right) - 1\right) \tag{6.70}$$

Other special cases of K-moment estimator coefficients b_{inp} are shown in Table 6.5. The fact that $b_{inp} = 0$ for $i < p$ suggests that, as the moment order increases, progressively, fewer data values determine the moment estimate, until it remains only one, the maximum, when $p = n$, with $b_{nnn} = 1$. Furthermore, if $p > n$ then $b_{inp} = 0$ for all i, $1 \le i \le n$, and therefore estimation becomes impossible.

An illustration of the performance of the unbiased estimator of equations (6.65)- (6.67) is given in Figure 6.4 for two distributions, lognormal and Pareto, where the estimates are indistinguishable from the true values, even for the highest possible order, $p = n$ (= 2000 in our example).

Table 6.5 Special cases of K-moment estimator coefficients.

Case	b_{inp}	Case	b_{inp}
$p = 0$	$b_{in0} = 0$	$p = n - 1$	$b_{n-1,n,n-1} = \frac{1}{n}, b_{n,n,n-1} = 1 - \frac{1}{n}$
$p = 1$	$b_{in1} = \frac{1}{n}$	$p = n$	$b_{nnn} = 1$
$p = 2$	$b_{in2} = \frac{2}{n}\frac{i-1}{n-1}$	$i = n$	$b_{nnp} = \frac{p}{n}$
$p = 3$	$b_{in2} = \frac{3}{n}\frac{i-1}{n-1}\frac{i-2}{n-2}$	$i = p$	$b_{pnp} = p\mathrm{B}(p, n-p+1)$ symmetry: $b_{pnp} = b_{n-p,n,n-p}$; minimum at $p = n/2$
$p = 4$	$b_{in4} = \frac{3}{n}\frac{i-1}{n-1}\frac{i-2}{n-2}\frac{i-3}{n-3}$		

Coming now to the tail K-moments, to get an unbiased estimator thereof, it suffices to revere the order of the sample, i.e. to replace $\underline{x}_{(i:n)}$ with $\underline{x}_{(n-i+1:n)}$ in (6.66), which becomes:

$$\hat{\overline{K}}'_p = \sum_{i=1}^{n} b_{inp}\, \underline{x}_{(n-i+1:n)} \tag{6.71}$$

where b_{inp} is again given by (6.67).

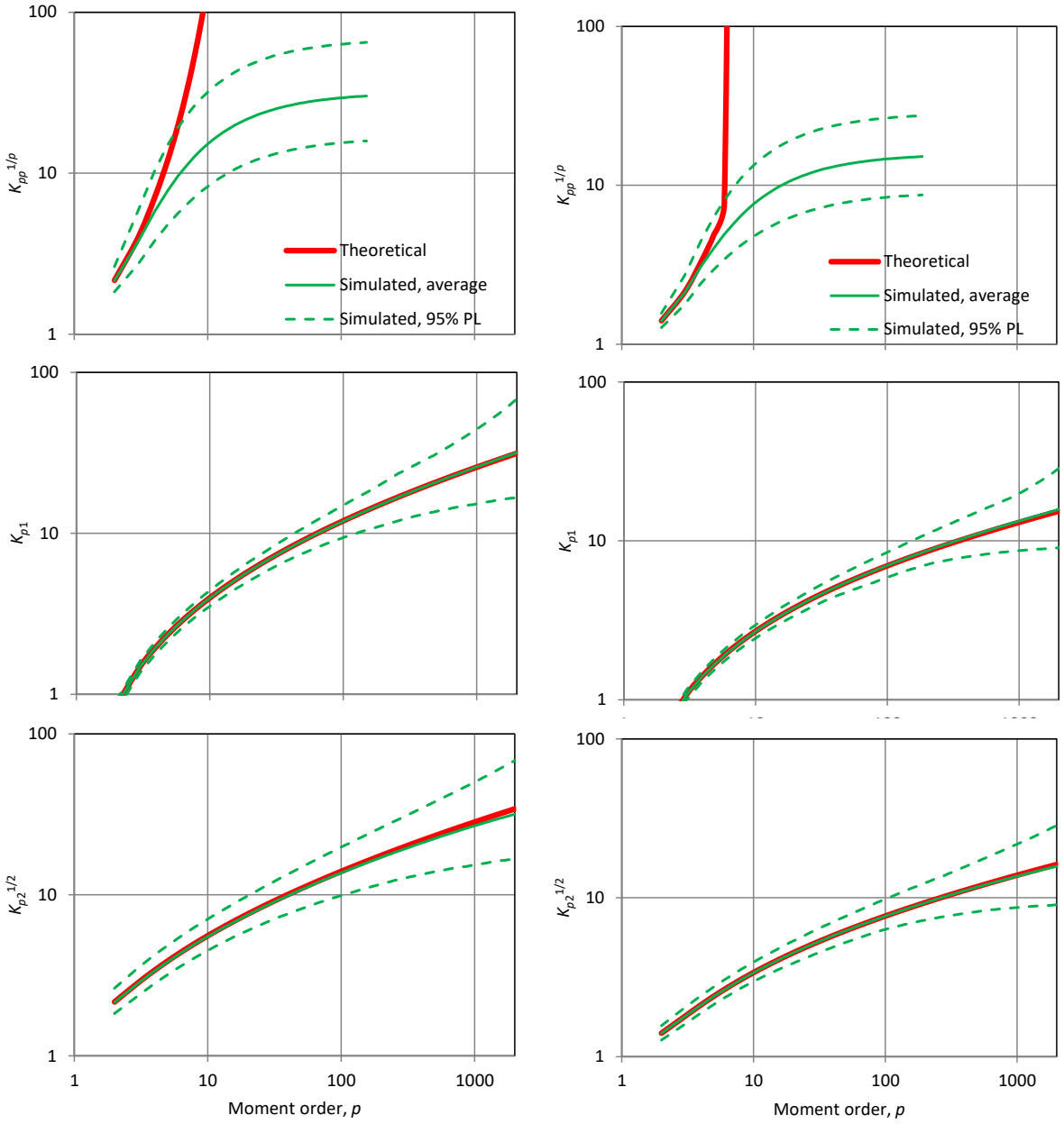

Figure 6.4 Illustration of the performance of the K-moment estimator of equations (6.65)-(6.67) for the lognormal distribution (**left column**; LN(0,1)) and the Pareto distribution (**right column**; tail index ξ = 0.15, scale parameter β = 1, lower bound ε = 0) for q = 1 **(middle row)** and q = 2 **(bottom row)**. For comparison the performance of the estimators of classical moments are also shown (**upper row**; notice that in the Pareto distribution the true moments are ∞ for $p > 1/0.15 = 6.87$). The true moments were determined by numerical integration for the lognormal distribution and from the equations in Table 6.2 for the Pareto distribution. The estimates are averages of 100 simulations each with n = 2000 and are indistinguishable from the true (theoretical) values. The 95% prediction limits (PL) are also shown.

6.10 Simplified estimators for independent samples

As can be readily verified from (6.65)-(6.67), the unbiased estimator of $\left(F\left(\underline{x}\right)\right)^{p-1}$ is not precisely the $(p-1)$ power of $\hat{F}\left(x_{(i)}\right)$. However, we can find simplified and approximately

unbiased estimators with this property. We generalize the estimator $\hat{F}\left(x_{(i:n)}\right)$ of (6.69) in the following form:

$$\hat{F}\left(x_{(i:n)}\right) = \frac{i-a}{n-b} \tag{6.72}$$

where a and b are constants or, more generally, functions of n. Then we form the estimator of K_p in one of the following two forms,

$$\underline{\hat{K}}'_p = \frac{p}{n}\sum_{i=1}^{n}\left(\frac{i-a}{n-b}\right)^{p-1}\underline{x}_{(i:n)}\,, \qquad \underline{\hat{K}}'_p = \frac{p}{n}\sum_{i=p}^{n}\left(\frac{i-a}{n-b}\right)^{p-1}\underline{x}_{(i:n)} \tag{6.73}$$

where the difference is in the lower limit of the sum; the second form assumes that the weight of $\underline{x}_{(i:n)}$ for $i < p$ is zero, as in the unbiased estimator (6.67).

It can be easily seen that the condition

$$a \geq b \tag{6.74}$$

ensures that $\hat{F}\left(x_{(i:n)}\right)$ will take values from 0 to 1, irrespective of the values $x_{(i)}$. Furthermore, the condition

$$2a - b - 1 = 0 \tag{6.75}$$

gives the estimator the symmetric property (6.70).

It is well known (see section 5.6) that the values

$$a = 0, \qquad b = -1 \tag{6.76}$$

make a precisely unbiased estimator of $F\left(x_{(i:n)}\right)$. However, here our aim is to find unbiased estimators of K-moments. As shown above, the values

$$a = b = 1 \tag{6.77}$$

are consistent with both the above requirements (6.68) and (6.70) and indeed have been used by Koutsoyiannis (2019a) for any order p, noting though that they result in some bias. Another common choice, proposed by Hosking et al. (1985a, b) for a similar case (see also Stedinger et al., 1993), is:

$$a = 0.35, \qquad b = 0 \tag{6.78}$$

which notably does not have the symmetric property (6.70).

Here, we suggest for the first form of the estimator (6.73) the parameters:

$$a = \frac{1 + n - \sqrt{n^2 - 1}}{2}, \qquad b = n - \sqrt{n^2 - 1} \tag{6.79}$$

with which the estimate in (6.72) becomes

$$\hat{F}\left(x_{(i:n)}\right) = \frac{2i - 1 - n + \sqrt{n^2 - 1}}{2\sqrt{n^2 - 1}} \tag{6.80}$$

By construction, this has the property of symmetry of equation (6.70) and satisfies precisely the necessary condition for unbiasedness of equation (6.68) for $p \leq 4$. For larger

p, the latter condition is slightly violated, but this can be fixed as described in Appendix 6-V.

For the second form of the estimator (6.73), after a systematic numerical investigation, the following parameter values have been found to be optimal:

$$a = \frac{1}{2} + \frac{1}{4(n-4)}, \qquad b = \frac{1}{2(n-4)} \tag{6.81}$$

With these parameters, the the estimate in (6.72) becomes:

$$\hat{F}\left(x_{(i:n)}\right) = \frac{4i(n-4) - 2n + 7}{4n(n-4) - 2} \tag{6.82}$$

Again, this has the property of symmetry of equation (6.70) but it does not satisfy precisely the necessary condition for unbiasedness of equation (6.68) for $p > 1$. However, the error is very small, as detailed in Appendix 6-V.

The above two estimators are simpler in their application than the precisely unbiased estimator (6.65)-(6.67) and could be used as alternatives for relatively low values of the order p. Deviations are negligible for $p < n/10$.

Another, even simpler estimation option, a quick-and-dirty estimator using just one data point per K-moment, will be discussed in Digression 6.D.

6.11 Approximately unbiased estimators of central K-moments

For $q = 1$ the estimator of the central K-moments, based on b_{ipn}, i.e.:

$$\underline{\widehat{K}}_p = \sum_{i=\mathrm{p}}^{n} b_{inp}\left(\underline{x}_{(i:n)} - \underline{\hat{\mu}}\right) \tag{6.83}$$

is again unbiased. Here we try to find an approximately unbiased estimator for $q = 2$. We start from the estimator that is given by (6.65) but replacing $\underline{x}_{(i:n)}$ with $\underline{x}_{(i:n)} - \underline{\hat{\mu}}$:

$$\underline{\widehat{K}}_{p2} = \sum_{i=\mathrm{p}}^{n} b_{i,n,p-1}\left(\underline{x}_{(i)} - \underline{\hat{\mu}}\right)^2 \tag{6.84}$$

From classical statistics of the second-order moments ($p = q = 2$) it is known that:

$$\frac{\mathrm{E}\left[\underline{\widehat{K}}_{22}\right] - K_{p2}}{K_{22}} = -\frac{1}{n} \tag{6.85}$$

It is not easy to derive theoretically a similar relationship for any $p \geq 2 = q$. However, a systematic simulation study shows that a good approximation is provided by generalizing the latter equation, i.e.:

$$\frac{\mathrm{E}\left[\underline{\widehat{K}}_{p2}\right] - K_{p2}}{K_{p2}} \approx -\frac{1}{n} \tag{6.86}$$

This results in:

$$\mathrm{E}\left[\underline{\hat{K}}_{p2}\right] \approx \frac{n-1}{n} K_{p2} \tag{6.87}$$

and thus, an approximately unbiased estimator is obtained by multiplying $\underline{\hat{K}}_{p2}$ by $n/(n-1)$. Obviously, the latter term becomes negligible if n is large. However, for small n, the bias should be taken into account. Simulation results for sample size as small as $n = 10$ are shown in Figure 6.5, which indicate the good performance of approximation (6.87).

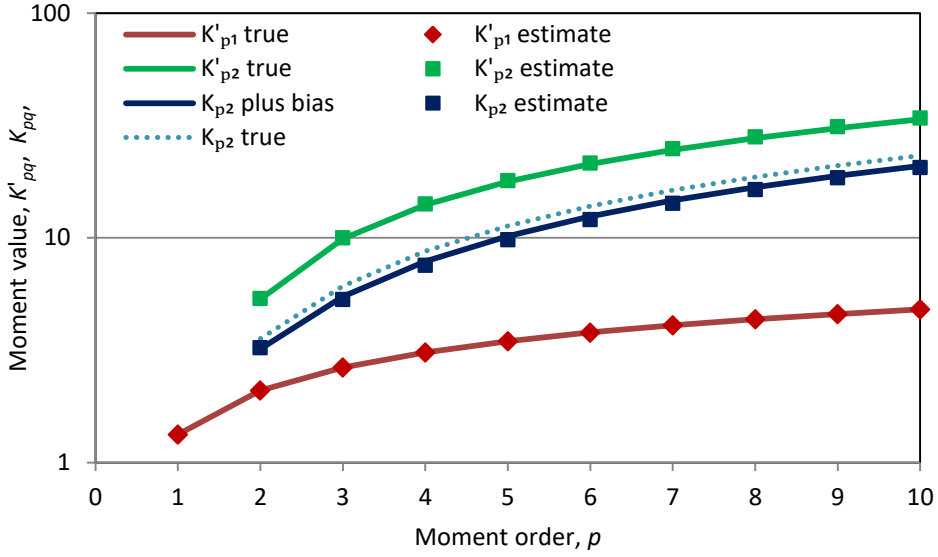

Figure 6.5 Simulation results of estimated noncentral and central moments of a Pareto distribution from a sample with size $n = 10$ for order $q = 1$ and 2. The Pareto distribution has tail index $\xi = 0.25$, scale parameter $\beta = 1$, and lower bound zero. The true moments are calculated from the equations in Table 6.2. The estimated moments are average sample estimates from 10 000 simulations. The curve "K_{p2} plus bias" corresponds to the right-hand side of equation (6.87).

6.12 Effect of autocorrelation on estimation

A K-moment is a characteristic of the marginal, first order, distribution of the process and therefore it is not affected by the dependence structure. However, the estimator is: time dependence induces bias to estimators of K-moments. Thus, the unbiasedness or approximate unbiasedness claimed in the previous sections ceases to hold in stochastic processes.

We quantify the effect of dependence by adapting the moment order for which the estimation is made. Specifically, the estimator of the noncentral K-moment for $q = 1$, i.e.:

$$\underline{\hat{K}}'_p = \sum_{i=1}^{n} b_{inp} \underline{x}_{(i:n)} \tag{6.88}$$

does not actually correspond to the theoretical K-moment for the same p, but to that for a smaller $p' \neq p$, i.e.:

$$K'_{p'} \coloneqq p' \mathrm{E}\left[\left(F\left(\underline{x}\right)\right)^{p'-1} \underline{x}\right] \tag{6.89}$$

Likewise, for central K-moments and for $q = 2$. We wish to find p'.

Noting that in $\widehat{\underline{K}}_1' \equiv \underline{\hat{\mu}}$ there is no bias, we begin studying K_2'. We recall from (6.3) that $K_2' = \mathrm{E}[\underline{x}_{(2)}] = \mathrm{E}\big[\max\big(\underline{x}_i, \underline{x}_j\big)\big]$, where $\big(\underline{x}_i, \underline{x}_j\big)$ are independent. In case of dependence the quantity:

$$K_2'^{\mathrm{d}} \coloneqq \mathrm{E}\big[\max\big(\underline{x}_i, \underline{x}_j\big)\big] \tag{6.90}$$

where the superscript 'd' stands for dependence, will be different from K_2'. We may assume that in the case of a time series with dependence, rather than of a random sample, what we actually estimate is $K_2'^{\mathrm{d}}$, rather than K_2'. In case of positive dependence, $K_2'^{\mathrm{d}} < K_2'$.

In Appendix 6-VI we determine that for a process with standard normal distribution the following adjustment is required:

$$\Theta^{\mathrm{d}}(n; r(\tau)) \coloneqq \frac{K_2^{\mathrm{d}} - K_2}{K_2} \approx -\frac{1}{n(n-1)} \sum_{\tau=1}^{n-1} r(\tau)\,(n-\tau) \tag{6.91}$$

where $r(\tau)$ is the process autocorrelation for lag τ. For a Markov process, in which $r(\tau) = r^{\tau}$,

$$\Theta^{\mathrm{M}}(n, r) \approx -\frac{2r}{(1-r)(n-1)} \tag{6.92}$$

where the superscript 'M' stands for Markov. Equation (6.92) clearly shows that, unless n is very low and r very high (e.g. > 0.90), $\Theta^{\mathrm{M}}(n) \approx 0$ and thus we can neglect the effect of autocorrelation. However, for an HK process, where $r(\tau) \approx H(2H-1)\tau^{2H-2}$, as shown in Appendix 6-VI:

$$\Theta^{\mathrm{HK}}(n, H) \approx \frac{2H(1-H)}{n-1} - \frac{1}{2(n-1)^{2-2H}} \tag{6.93}$$

For $H \gg 1/2$, the last term of the right-hand part is typically non-negligible and therefore the bias should be taken into account. A graphical depiction of adjustment coefficient $\Theta^{\mathrm{HK}}(n, H)$ is provided in Figure 6.6.

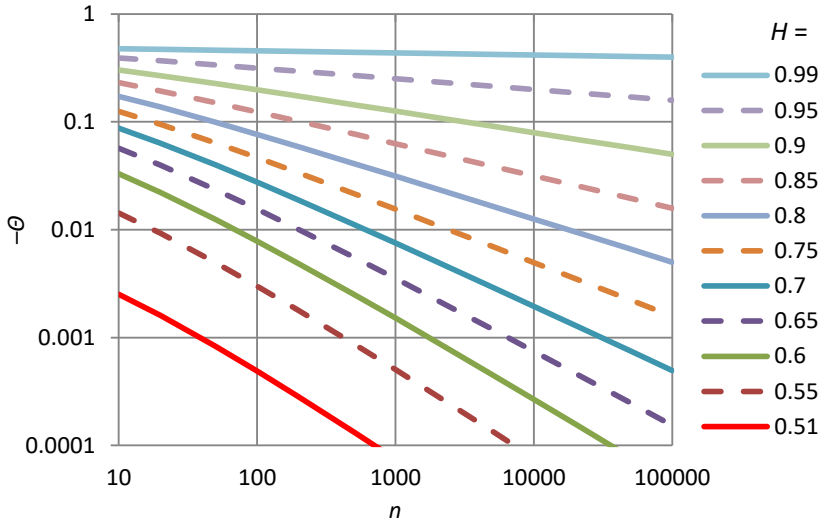

Figure 6.6 Adjustment coefficient $\Theta\ \big(\equiv \Theta^{\mathrm{HK}}(n, H)\big)$ of the central K-moment for $q = 1$ for a Hurst-Kolmogorov process.

Now, we assume, based on simulation results, that the same adjustment applies approximately on all orders p, i.e.,

$$\Theta^{\mathrm{HK}}(n,H) = \frac{K_p^{\mathrm{d}} - K_p}{K_p} \tag{6.94}$$

Under this assumption we find in Appendix 6-VI, the following approximation of a modified order p' for which:

$$p' \approx 2\Theta + (1-2\Theta)p^{\left((1+\Theta)^2\right)}, \qquad K_{p'} = K_p^{\mathrm{d}} = (1+\Theta)K_p \tag{6.95}$$

where for notational simplification we have shortened $\Theta^{\mathrm{HK}}(n,H)$ to Θ.

The same approximation for p' will hold for the noncentral K-moments. Further, we may use the same coefficients and results for q = 2. Specifically, for the noncentral K'_{p2} the framework could be kept unchanged. However, for the central K_{p2}, while we can keep p' as derived above, we must have in mind that in the presence of HK behaviour $\underline{\widehat{K}}_{22}$ is a biased estimator of K_{22}. Applying the results of section 4.6 (in particular, equation (4.24)) for the HK process we find that the bias is:

$$\frac{\mathrm{E}\left[\underline{\widehat{K}}_{22}\right] - K_{22}}{K_{22}} = -\frac{1}{n^{2-2H}} \tag{6.96}$$

This is similar to (6.85) expect that n is raised to the power $2-2H$ (for $H = 1/2$ we recover (6.85)); also this bias is roughly 2Θ. Generalizing this for any p, as we did in section 6.11, and expressing it for the adapted order p' we write:

$$\frac{\mathrm{E}\left[\underline{\widehat{K}}_{p'2}\right] - K_{p'2}}{K_{p'2}} = -\frac{1}{n^{2-2H}} \tag{6.97}$$

Hence what we estimate in this case is the K-moment plus (negative) bias.

While the derivation of these relationships in Appendix 6-VI was based on the normal distribution, the final equation (6.212) can be used as an approximation regardless of the distribution. Figure 6.7 provides an empirical confirmation that the approximation is satisfactory, based on simulations from the lognormal distribution, which differs substantially from the normal, and a high Hurst parameter, H = 0.9. Interestingly, what would be assigned as K-moment for p =2000, without taking into account the effect of dependence, actually corresponds to the true K-moment of $p' \approx 500$—and as we will see in section 6.14 this has dramatic consequences in the assignment of return periods to events. Another important observation from comparing Figure 6.7t with Figure 6.4 (left), which are for the same marginal distribution, is the dramatic broadening of the prediction intervals. This illustrates how dependence amplifies uncertainty.

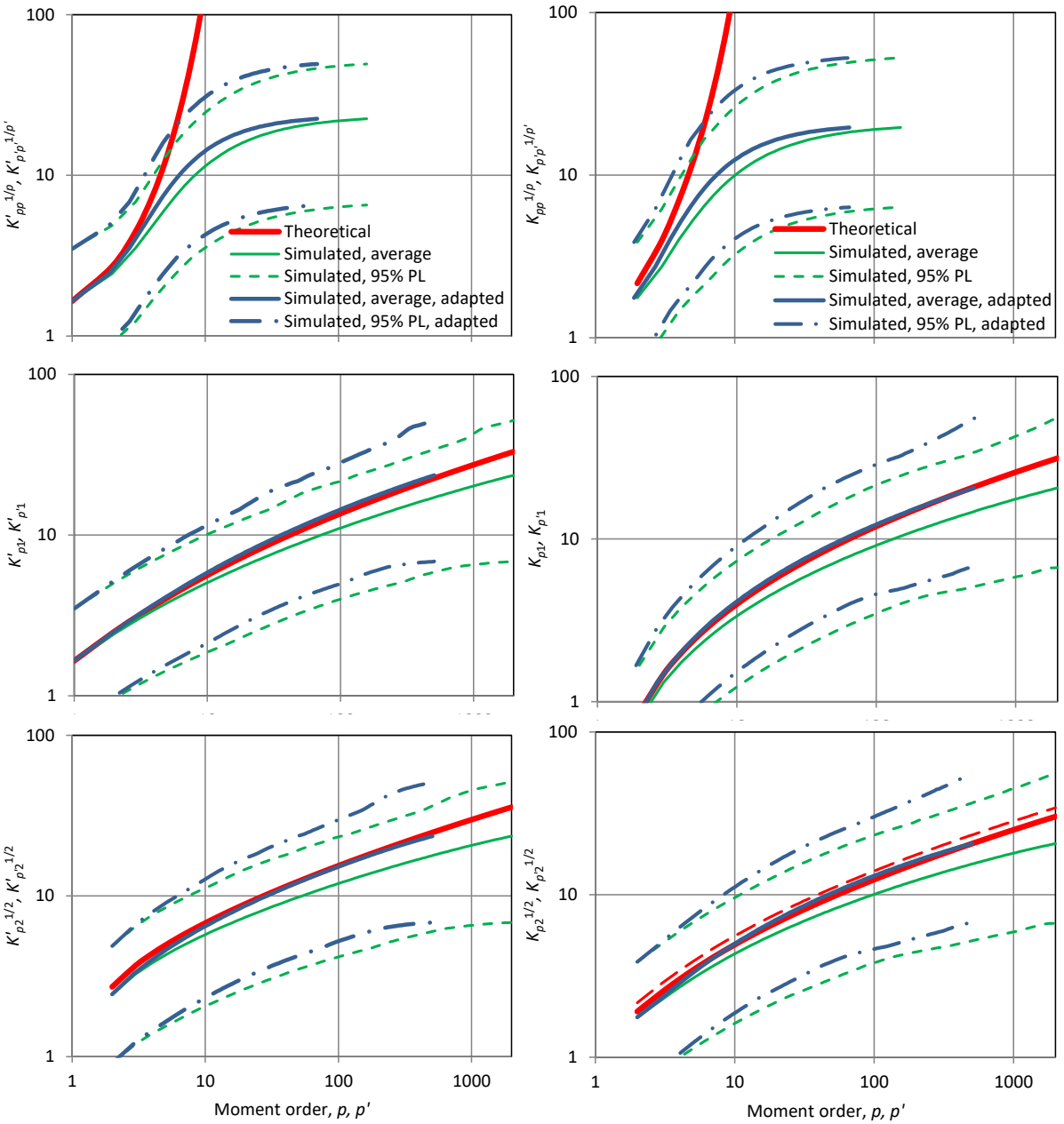

Figure 6.7 Illustration of the performance of the adaptation of K-moment estimation for an HK process with Hurst parameter 0.9 and lognormal marginal distribution (LN(0,1); same as in Figure 6.4, left): **(left column)** noncentral moments; **(right column)** central moments; **(middle row)** q = 1 **(bottom row)** q = 2. For comparison classical moments are also shown **(upper row)**. The estimates are averages of 200 simulations each with n = 2000 and are almost indistinguishable from the theoretical values. For K_{p2} (lower right) the true K-moments are represented by the dashed curve while the curve marked "theoretical" corresponds to the true plus bias, where bias is determined from equation (6.97). The 95% prediction limits (PL) are also shown.

Digression 6.B: Does periodicity affect estimation of K-moments?

Hydroclimatic processes are influenced by regular seasonal or diurnal changes, or else periodicities. Often, in their stochastic treatment we use cyclostationary models, whose statistical properties are deterministic periodic functions of time. In other cases, we are interested in the overall behaviour of a process. For example, when studying rainfall extremes, often we are interested only in the magnitude of the extremes, rather than in the season in which an extreme has occurred. In such cases we can model the natural process as a stationary stochastic process,

treating the periodicity indirectly through a periodic autocovariance function (Koutsoyiannis, 2017).

As an example, we assume a fully deterministic, strictly periodic process composed of a single harmonic with period a, described by:

$$v(t) = \cos(2\pi\,(t+b)/a)\,, \qquad 0 \le b \le a$$

where b is the phase. This is as a deterministic process but, when it is superimposed to a stochastic process, the resulting process is a stochastic process, too. Therefore, we need to know the stochastic properties of $v(t)$ as if it was a stochastic process. In particular, its autocorrelation function is (Koutsoyiannis, 2017):

$$r_v(h) = \cos(2\pi\, h/a)$$

This does not vanish off for large lags; particularly for lags h that are multiples of the period a keeps a constant value 1. Therefore, as autocorrelation is occasionally high, irrespective of lag, one may suspect that this behaviour may influence estimation. However, this is not the case because, in fact, the autocorrelation alternates between positive and negative values and the average is zero.

We will illustrate this by a simulation experiment. To make it more interesting we superimpose a harmonic component to a Markov process and we will thus show that neither the Markovian behaviour nor the periodicity affect estimation. For the Markov process we use the simple AR(1) model in discrete time $\tau \coloneqq t/D$:

$$\underline{u}_\tau = r\underline{u}_{\tau-1} + \sqrt{1-r^2}\,\underline{w}_\tau$$

where $\underline{w}_\tau$ is standard Gaussian white noise. It is easily seen that $\underline{u}_\tau$ is Gaussian:

$$f_u(u) = \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}$$

with autocorrelation for discrete lag $\eta \coloneqq h/D$:

$$r_u(\eta) = r^\eta$$

The marginal distribution and density of the harmonic function are (cf. Markonis and Koutsoyiannis, 2013):

$$F_u(v) = 1 - \frac{\arccos(v)}{\pi}, \qquad f_v(v) = \frac{1}{\pi\sqrt{1-v^2}}, \qquad -1 \le v \le 1$$

and its variance is $\mathrm{var}[\,\underline{v}] = 1/2$. Now we form the composite process

$$\underline{z} \coloneqq \sqrt{2/3}\,(\underline{u} + \underline{v})$$

which has variance $\mathrm{var}[\,\underline{z}] = (2/3)(1+1/2) = 1$ and autocorrelation (identical to autocovariance):

$$r_z(\eta) = (2/3)(r^\eta + \cos(2\pi\,\eta D/a))$$

The marginal density function of $\underline{z}$, obtained from the convolution equation (2.102), is:

$$f_z(z) = \sqrt{\frac{3}{4\pi^3}} \int_{-1}^{1} \frac{\exp\left(-\frac{1}{2}\left(\sqrt{\frac{3}{2}}z - u\right)^2\right)}{\sqrt{1-u^2}}\,\mathrm{d}u$$

This integral cannot be evaluated analytically, but as shown in Figure 6.8(a), its numerical evaluation suggests that it is close to the Gaussian density with some noticeable deviations at large values of the stochastic variable. If we make the transformation

$$\underline{y} = \underline{z}\,|\underline{z}|^{0.05}$$

and determine its density using equation (2.11), then that density becomes almost indistinguishable from that of the normal distribution, specifically that of $\underline{u}$. This is depicted in Figure 6.8(a). Therefore, if we make the final transformation

$$\underline{x} = e^{\underline{y}}$$

we will get a variable $\underline{x}$ with lognormal marginal distribution LN(0,1), as that shown in Figure 6.4 (left), but now with Markov dependence and periodicity, as described above by the autocorrelation $r_z(\eta)$. As shown in Figure 6.8(b), the autocorrelation for $\underline{y}$ and $\underline{x}$ do not differ substantially from that of $\underline{z}$, $r_z(\eta)$.The latter are calculated by stochastic simulation to avoid other type of numerical integrations which are more tedious. A feeling about how the generated time series look is obtained from the plot of Figure 6.8(c).

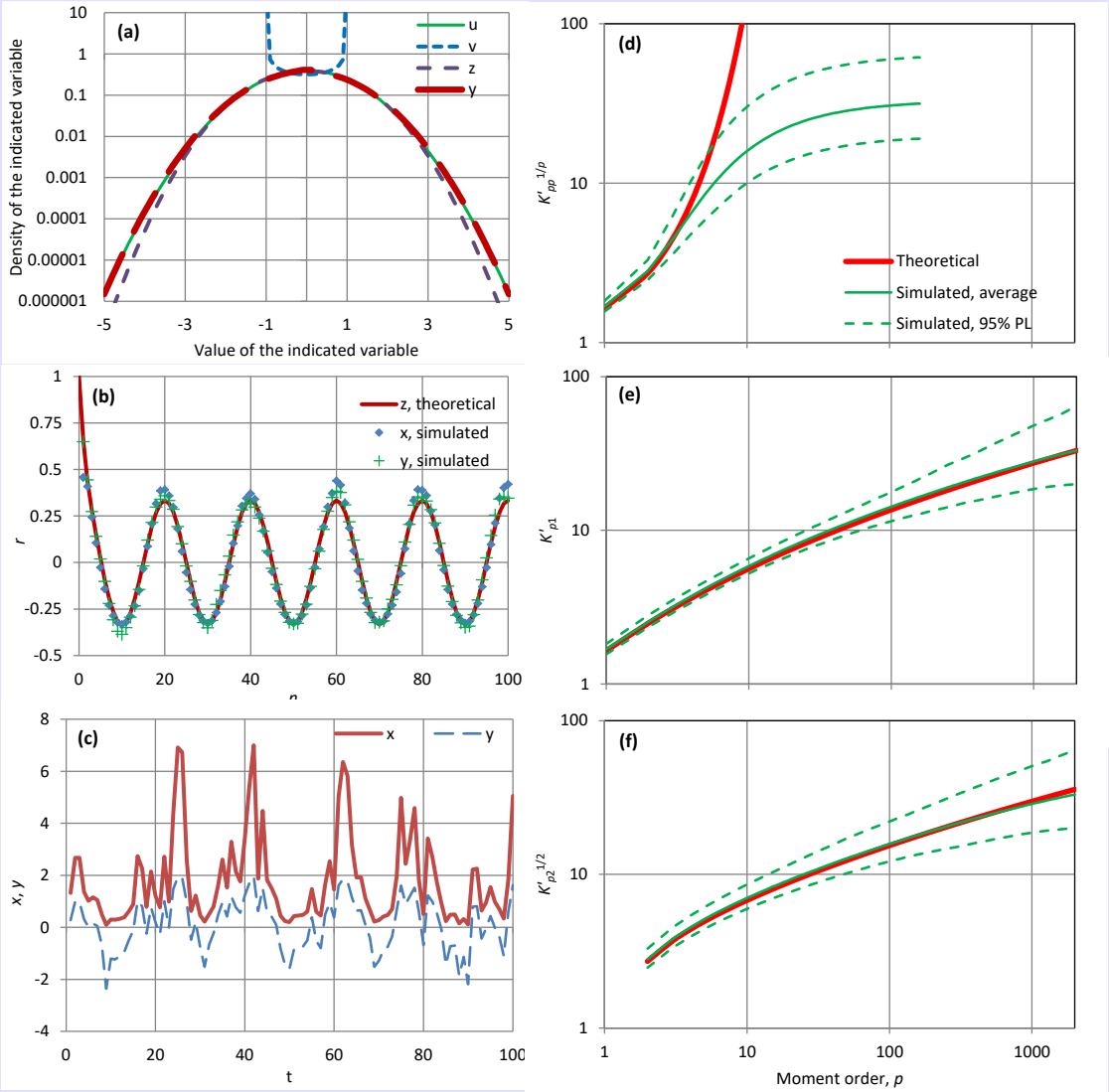

Figure 6.8 Illustration of theoretical and simulated results for the example of Digression 6.B, where the final process $\underline{x}$ has lognormal marginal distribution LN(0,1), the lag-1 autocorrelation of the Markov process is r = 0.5 and the period of the periodic component is a = 20: **(a)** probability densities of variables $\underline{u}, \underline{v}, \underline{z}$ and $\underline{y}$; **(b)** autocorrelation of processes $\underline{z}, \underline{x}$ and $\underline{y}$; (c) 100 terms of time series from processes $\underline{x}$ and $\underline{y}$; **(d)** classical moments; **(e)-(f)** K-moments for q = 1 and (2). Panels (d)-(f) are identical to those of Figure 6.4 (left), which is for the same marginal distribution. The estimates are averages of 100 simulations each with n = 2000 and in the case of K-moments are indistinguishable from the true (theoretical) values.

Now we perform a simulation experiment for the K-moments as that presented in section 6.9 and depicted in Figure 6.4 (left), which is also for the lognormal distribution LN(0,1). To estimate

K-moments we use the unbiased estimator of equations (6.65)-(6.67) without any type of correction at all. Panels (d)-(f) in the figure show the classical and K-moments, theoretical and estimated. Comparing them with the respective plots of Figure 6.4 (left), we clearly see that they are identical. This confirms our claim that periodicity and Markov dependence to not affect K-moment estimates.

6.13 Estimation by merging information from dependent records

Assuming that we have several observation records, representing the same stochastic process, we can use them simultaneously to improve our estimations. If the records can be regarded as random samples and each one is independent of the others, we can merge them in a single sample and use the merged sample for estimation. Assuming that there are m samples, each with length n_1, we will then form a merged sample of size $n = m\,n_1$ and we can apply the usual estimation procedures for that sample, while estimation uncertainty will depend on the size n only—and obviously will be smaller than that from a sample with size n_1.

But what happens if the different records represent correlated processes? This is a frequent case we met in practice; for example, if we study rainfall observed at several adjacent stations. Again, merging the records improves the estimation reducing uncertainty, but not as much as if there was a single random sample of size n.

We denote $\underline{x}_{ij}, i = 1, \dots, m, j = 1, \dots n_1$, the stochastic variable representing the jth item of sample i and $\hat{\underline{\mu}}_i = \left(\underline{x}_{i1} + \cdots + \underline{x}_{in_1}\right)/n_1$ the average of the ith sample. We assume that the samples are cross-correlated with same correlation $r \geq 0$, i.e., $\operatorname{corr}\left[\underline{x}_{ij}, \underline{x}_{i'j}\right] = r, i \neq i'$, while $\operatorname{corr}\left[\underline{x}_{ij}, \underline{x}_{i'j'}\right] = 0, j \neq j'$. It is easy to show that, if $\gamma \coloneqq \operatorname{var}\left[\underline{x}_{ij}\right]$, then the variance of $\hat{\underline{\mu}}_i$ is γ/n_1 and the overall average is:

$$\gamma^{(n)} \coloneqq \operatorname{var}\left[\frac{\hat{\underline{\mu}}_1 + \cdots + \hat{\underline{\mu}}_m}{m}\right] = \frac{\gamma}{n}\left(1 + r(m-1)\right) \tag{6.98}$$

In a sample of size n the variance would be γ/n. If we introduce an *equivalent sample size* n' so that the variance in (6.98) be γ/n', then it is readily found that

$$n' = \frac{m\,n_1}{1 + r(m-1)} \tag{6.99}$$

It can be noted that as m increases (tends to infinity) the equivalent sample size tends to an upper limit, which is $n'_{\max} = n_1/r$.

In the above analysis, whose results have been well known for many decades (Yule, 1945; Matalas and Langbein, 1962; Castellarin et al., 2005), the determination of the equivalent sample size has been based on the variance of the sample average. We can repeat the analysis for higher moments, although this is more complicated. Stedinger (1983) has shown (for normal distribution) that if we use the variance of the sample variance as a basis, then an equivalent sample size n'' is derived by a relationship analogous to (6.99):

$$n'' = \frac{n}{1 + r^2(m-1)} \tag{6.100}$$

Obviously, $n'' \geq n'$ and the upper limit as $m \to \infty$ becomes $n''_{\max} = n_1/r^2 \geq n'_{\max}$. This suggests that as the moment order increases, the information gain from merging records increases too. Mimiyianni (2010) has confirmed this property in a systematic simulation study. This is very encouraging if we study extremes, as extremes are related to high-order moments.

The K-moments estimation is thus expected to improve substantially by merging records, particularly for high orders. The analysis we have made in section 6.12 could also be used in this case as by merging cross-correlated records we induce autocorrelation to the merged sample. However, the problem is more complicated now, as it depends on three parameters, r, m and n_1. What is required is again to estimate an adapted moment order p' for each estimated K'_{pq}. Stochastic simulation is obviously a generic method that can easily handle this problem. However, a quick technique for practical use is this.

- For $p \leq m$ we set $p' = p$.
- For $p > m$ we use the HK framework of section 6.12 for H corresponding to r, i.e.:

$$H = \frac{1}{2} + \frac{\ln(1+r)}{2\ln 2} \tag{6.101}$$

  Specifically, we apply (6.93) to find $\Theta^{\mathrm{HK}}(n, H)$, and then modify (6.95) to estimate p' as:

$$p' \approx 2\Theta + (1-2\Theta)(p-m+1)^{((1+\Theta)^2)} + m - 1 \tag{6.102}$$

It can be readily verified that for $p = m$, $p' = p$.

Figure 6.9 presents two examples with application of the above technique for m = 100 subsamples of size $n_1 = 20$ each and for a low and a high value of correlation coefficient. The simulated subsamples were constructed by the equations $\underline{x}_{ij} = \exp \underline{y}_{ij}$, $\underline{y}_{ij} = a\underline{v}_{0j} + \sqrt{1-a^2}\underline{v}_{ij}$ where $\underline{v}_{ij}, i = 0, \ldots, m, j = 1, \ldots, n_1$ are independent stochastic variables with distribution N(0,1). The parameter a was chosen 0.5 and 0.8 for the low and high correlation, respectively, resulting in correlation coefficients for $\underline{y}$, r = 0.25 and r = 0.64, respectively, which were used in calculations. (The correlations for $\underline{x}$ are r = 0.16 and r = 0.52, respectively.)

The figure shows clearly that merging of samples enables estimation of K-moments of much higher order, of about 1000 or more, while a single sample allows estimation up to order 20. The importance of estimating high order moments will become clear in section 6.14, where it will be seen that estimates of high-order K-moments are equivalent to quantile estimates of high return periods. In addition, the estimation uncertainty, quantified by the prediction intervals, is substantially reduced in the case of the merged sample, in comparison to that of an individual sub-sample. However, comparison with Figure 6.4 (left), which is for the same marginal distribution, shows that the uncertainly, is still higher than in a purely random sample of equal length.

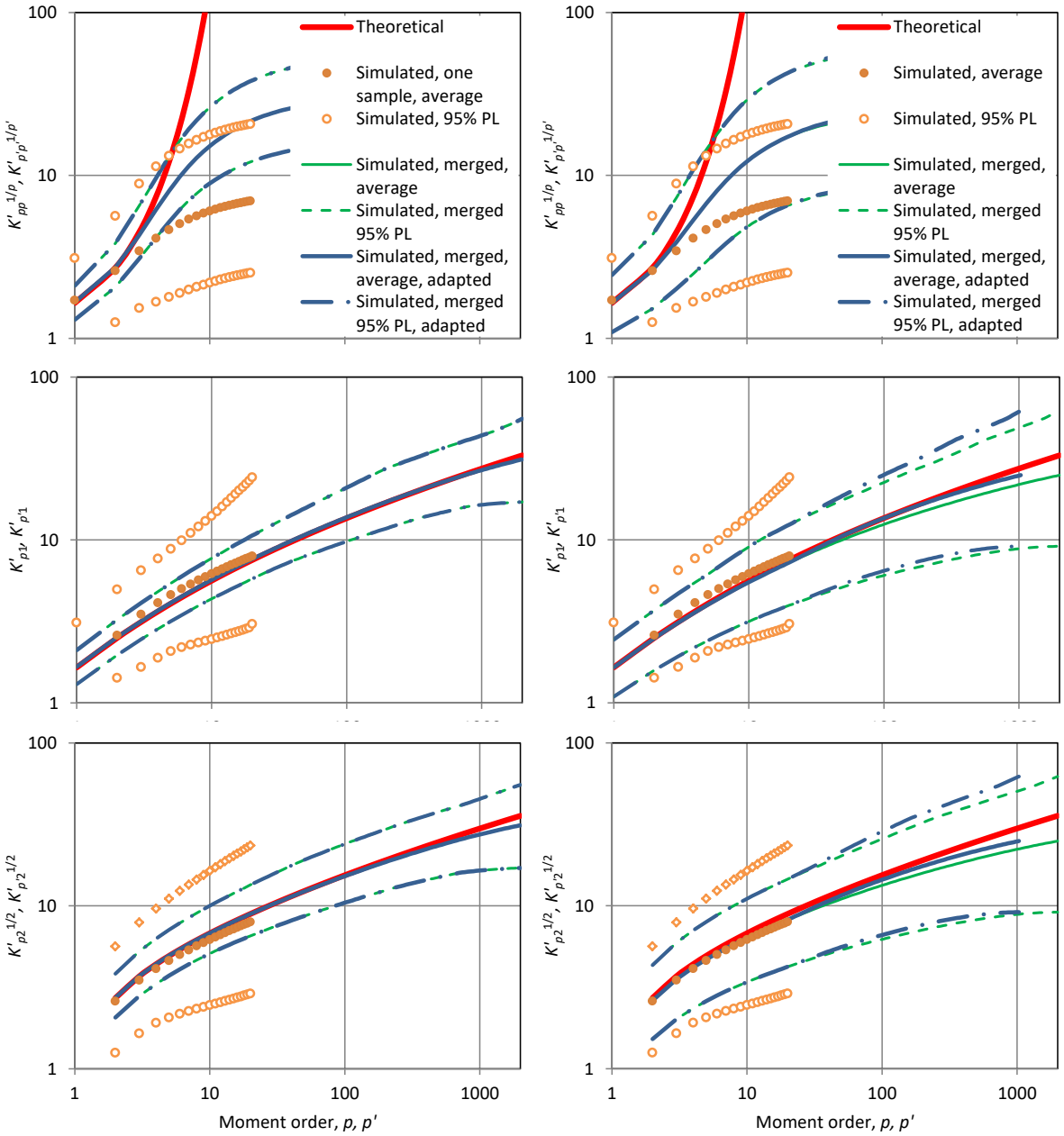

Figure 6.9 Illustration of the performance of the adaptation of K-moment estimation for a merged sample composed of m = 100 correlated sub-samples of size n_1 = 20 each. Each sub-sample is random with lognormal marginal distribution (LN(0,1); same as in Figure 6.4, left). The correlation of the logarithms of the variables (which have normal distribution) are r = 0.25 **(left column)** and r = 0.64 **(right column)**. Noncentral K-moments are shown for q = 1 **(middle row)** and q = 2 **(bottom row)**. Classical noncentral moments are also shown **(upper row)**. For comparison, estimates from one sample with size n_1 = 20 are also shown. All estimates are averages of 100 simulations, while 95% prediction limits (PL) are also shown. In the left column (r = 0.25) the adaptation turns out to be negligible and the adapted curves and are indistinguishable from the non- adapted.

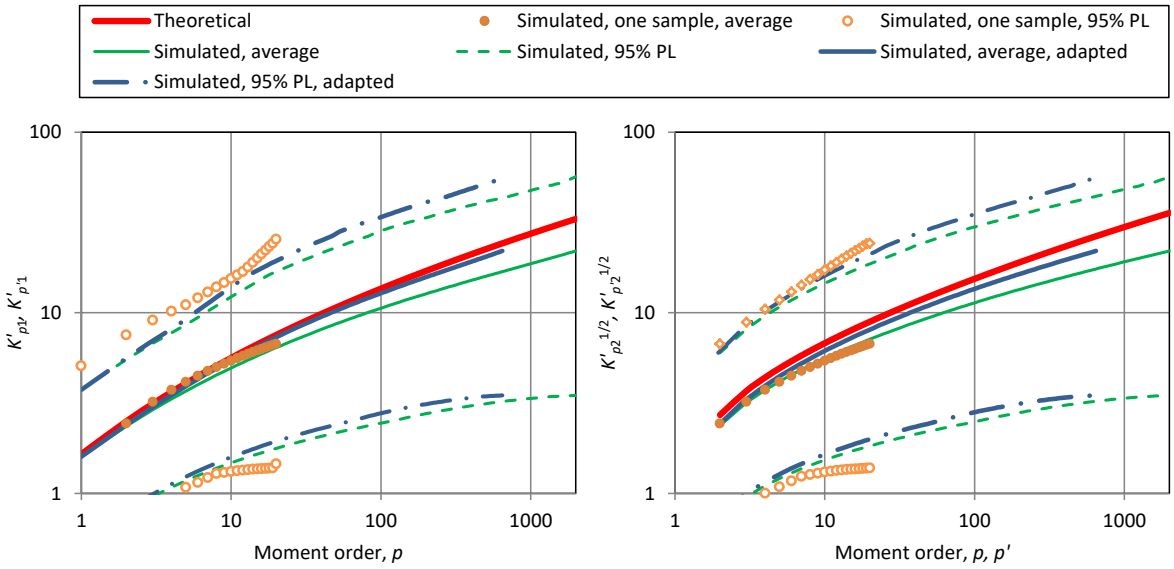

Figure 6.10 Illustration of the performance of the adaptation of K-moment estimation for a merged time series composed of m = 100 correlated time series of size n_1 = 20 each, generated from the exponentiated HK process for H = 0.8, having lognormal marginal distribution (LN(0,1)). The correlation of the logarithms of the variables of the different time series (which have normal distribution) is r = 0.64. Noncentral K-moments are shown for q = 1 **(left)** and q = 2 **(right)**. For comparison estimates from one time series with size n_1 = 20 are also shown. All estimates are averages of 100 simulations, while 95% prediction limits (PL) are also shown.

A final question to discuss is how to deal in situations where the subsamples are not IID but time series from a process with HK behaviour. Again, stochastic simulation provides the means for proper adaptation of the estimates. A quick-and-dirty solution is to use the HK framework of section 6.12 for the entire range of moment orders and for a "bulk" Hurst parameter. In this case the variance of $\underline{\hat{\mu}}_i$ is γ/n_1^{2-2H} and equation (6.98), which gives the variance of the overall average, is modified to

$$\gamma^{(n)} = \frac{\gamma}{n_1^{2-2H} m}\big(1 + r(m-1)\big) \tag{6.103}$$

The bulk Hurst parameter could be specified according to the following equation, which fits an HK climacogram to the time scales 1 and n:

$$H_\mathrm{b} = 1 + \frac{\ln\big(\gamma^{(n)}/\gamma\big)}{2\ln(n)} \tag{6.104}$$

This results in:

$$H_\mathrm{b} = \left(1 - \frac{\ln m}{\ln n}\right) H + \frac{\ln\Big(\big(1 + r(m-1)\big)\Big) + \ln m}{2\ln(n)} \tag{6.105}$$

Illustration of this technique is provided in Figure 6.10 for 100 time series of size 20 each. Each of them was constructed with the same method as in the examples of Figure 6.9 except that the variables $\underline{v}_{ij}$ for each i follow the HK process with H = 0.8, while they are independent for different i. The parameter α was chosen 0.8, resulting in correlation coefficients for $\underline{y}$, r = 0.64. The resulting H_b is 0.892 and the overall performance of the

method, compared to accurate simulation results is satisfactory (even though with slightly higher H = 0.91, not shown in the figure, the results would perfectly correspond to the simulation results). Figure 6.10 shows that even in this case merging of time series enables estimation K-moments of much higher order, of $p' \approx 650$, while a single time series allows estimation up to order $p' = 16$ only. However, the estimation uncertainty, quantified by the prediction intervals, is not reduced substantially in the case of the merged time series.

6.14 Return periods of K-moment values and Λ-coefficients

As we have seen in section 5.6, order statistics have an important advantage over other statistics, as to each of them we can assign a value of the distribution function and hence of return period. This turns out to be the case also with K-moments as they are closely related to order statistics. Intuitively we can expect that the return period corresponding to the noncentral K-moment of orders (p,1), i.e. the value $x = K'_p$, will correspond to a return period of about $2pD$ (where D is the time step or, more generally, a time reference for the specification of return period). This is precise for a symmetric distribution and for p = 1, as K'_1 is the mean value which has return period $2D$ and, as we will see below, it cannot be much lower than $2pD$ for any p and for any distribution.

Generally, we can express the return period by the relationship:

$$\frac{T\left(K'_p\right)}{D} = \Lambda_p p \tag{6.106}$$

where Λ_p is a coefficient generally depending on the distribution function and the order p. As we will see, the range of variation of Λ_p is not wide. As a first rough approximation, the rule of thumb:

$$\Lambda_p = 2, \qquad \frac{T\left(K'_p\right)}{D} = 2p \tag{6.107}$$

helps intuition. However, the precise definition of Λ_p is:

$$\Lambda_p := \frac{1}{p\left(1 - F\left(K'_p\right)\right)} \tag{6.108}$$

For given p and distribution function $F(x)$, K'_p is analytically or numerically determined from its definition. Then $T(K'_p)$ and Λ_p are determined from their definitions.

In absence of an analytical solution, we can establish an exact relationship between p and T by doing numerical calculations for several p. The small variation of Λ_p with p makes possible a very good approximation if we first accurately determine the specific values Λ_1 and Λ_∞. The value Λ_1 is very easy to determine, as it refers to the return period of the mean:

$$\Lambda_1 = \frac{1}{1 - F(\mu)} = \frac{T(\mu)}{D} \tag{6.109}$$

Furthermore, in a number of customary distributions, specifically those belonging to the domain of attraction of the Extreme Value Type I distribution, Λ_∞ has a constant value, independent of the distribution. As shown in Appendix 6-VII (see also section 5.6) this is:

$$\Lambda_\infty = \mathrm{e}^{\gamma} = 1.781 \tag{6.110}$$

where γ is the Euler constant.

For the approximation of Λ_p we can use the following simple relationship, which is satisfactory for several distributions:

$$\Lambda_p \approx \Lambda_\infty + \frac{(\Lambda_1 - \Lambda_\infty)}{p} \tag{6.111}$$

This yields a linear relationship between the return period T and p:

$$\frac{T(K'_p)}{D} = p\Lambda_p \approx \Lambda_\infty p + (\Lambda_1 - \Lambda_\infty) \tag{6.112}$$

However, in some distributions like the lognormal and Weibull, the decay of Λ_p with increasing p is very slow. Furthermore, in some cases Λ_p in not always a decreasing function of p, as implied by equation (6.111). To account for such cases, we may enrich (6.111) adding two more parameters β and B, according to the expression:

$$\begin{gathered}\Lambda_p \approx \Lambda_\infty + \frac{A}{p} - B\ln\left(\beta'\left(1 + \frac{\beta}{(p+1)^\beta - 1}\right)\right), \\ A = \Lambda_1 - \Lambda_\infty + B\ln\left(\beta'\left(1 + \frac{\beta}{2^\beta - 1}\right)\right), \qquad \beta' := \begin{cases} 1, & \beta \geq 0 \\ \dfrac{1}{1-\beta}, & \beta \leq 0 \end{cases}\end{gathered} \tag{6.113}$$

where the expression for A ensures exact recovering of Λ_1. Most interesting are the cases:

(a) $\beta = 0$, in which:

$$\Lambda_p \approx \Lambda_\infty + \frac{A}{p} - B\ln\left(1 + \frac{1}{\ln(p+1)}\right), \quad A = \Lambda_1 - \Lambda_\infty + B\ln\left(1 + \frac{1}{\ln(2)}\right) \tag{6.114}$$

(b) $\beta = 1$, in which:

$$\Lambda_p \approx \Lambda_\infty + \frac{A}{p} - B\ln\left(1 + \frac{1}{p}\right), \qquad A = \Lambda_1 - \Lambda_\infty + B\ln(2) \tag{6.115}$$

(c) $\beta = -1$, in which:

$$\Lambda_p \approx \Lambda_\infty + \frac{A}{p} - B\ln\left(1 + \frac{1}{2p}\right), \qquad A = \Lambda_1 - \Lambda_\infty + B\ln(3/2) \tag{6.116}$$

The rationale and details of this approximation are given in Appendix 6-VIII. The resulting return period is

$$\frac{T(K'_p)}{D} = p\Lambda_p \approx p\Lambda_\infty + \mathrm{A} - Bp\ln\left(\beta'\left(1 + \frac{\beta}{(p+1)^\beta - 1}\right)\right) \tag{6.117}$$

We refer to the approximations (6.107), (6.112) and (6.117) as the *rule-of-thump*, *linear* and *nonlinear approximations*, respectively.

Table 6.6 gathers the equations for a number of customary distributions giving all quantities that support the approximation of the entire series of Λ_p. Furthermore, Figure 6.11 illustrates the very satisfactory approximation achieved by the above method for all distributions. A prominent characteristic of the lognormal distribution is the very slow decay of the Λ-coefficient. If compared to the Pareto distribution (also shown in the figure), the lognormal distribution has the limit $\Lambda_\infty = 1.781$ against 2.255 of the Pareto case. However, even for moment order as high as 100 000, the former distribution retains a Λ value much higher than the latter.

It is interesting that, according to equations (6.111) and (6.113), only two to four numbers, namely the coefficients Λ_1 and Λ_∞, and occasionally β and B, can approximate the complete series of Λ_p, practically for any distribution function. In turn, as the Λ-coefficients are independent of the scale and location parameters (see below), by adding the latter (in the form of $\mu = K_1'$ and any of K_2, K_{22} or their noncentral variants) to the three Λ-coefficients we obtain an efficient parameterization of any distribution function. This is justified by the fact that the Λ-coefficients, along with a location and scale parameter, determine the complete series of K-moments K_p, while the latter are related to the quantile function via the Laplace transform (section 6.8).

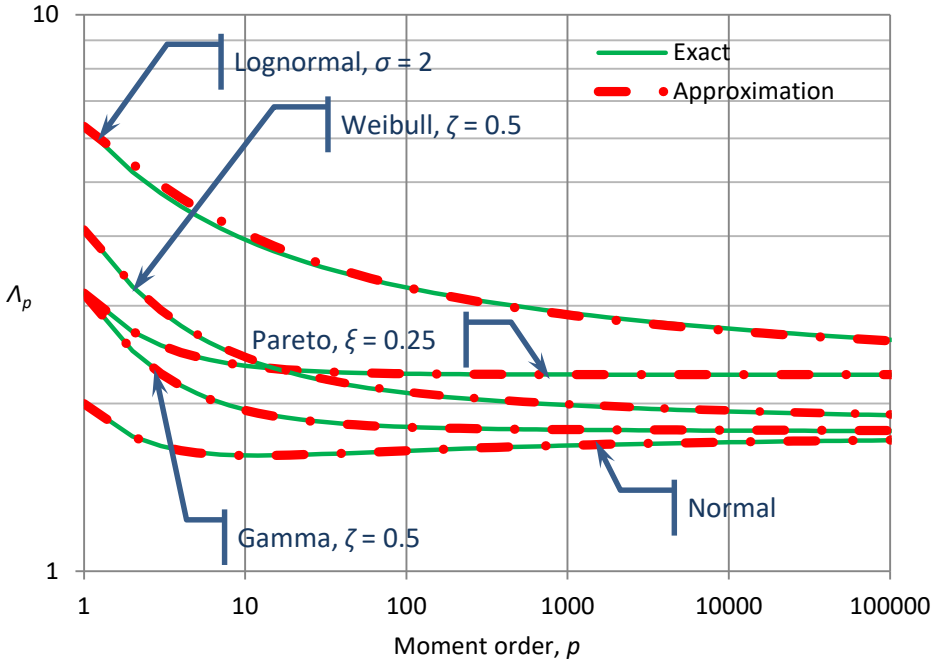

Figure 6.11 Illustration of the approximation of Λ-coefficients achieved by equation and (6.113) for the indicated distribution functions.

Table 6.6 Characteristic parameters for accurate approximations of Λ-coefficients and K-moments for several customary distributions.

Distribution[1] and tail function, $\overline{F}(x)$	Λ_1	Λ_∞	β	B
Normal, $\int_x^\infty \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma}\mathrm{d}u$	2	e^γ	0	0.73
Exponential[2], $\mathrm{e}^{-x/\lambda}$	e	e^γ	1	$\frac{2\Lambda_1 - 3\Lambda_\infty}{2(1-\ln 2)} = 0.152$
Gamma[3], $\frac{\Gamma_{x/\lambda}(\zeta)}{\Gamma(\zeta)}$	$\frac{\Gamma(\zeta)}{\Gamma_\zeta(\zeta)}$	e^γ	0	$-0.3\ln(\Lambda_1 - 1.93) - 0.05$
Weibull, $\exp\left(-\left(\frac{x}{\lambda}\right)^\zeta\right)$	$\mathrm{e}^{\left(\Gamma\left(1+\frac{1}{\zeta}\right)\right)^\zeta}$	e^γ	$\frac{0.02}{\zeta} - 0.003$	$1.16 + 1.2(\Lambda_\infty - \Lambda_1)$
Lognormal, $\int_x^\infty \frac{\exp\left(-\frac{(\ln(u/\lambda))^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma\,u}\mathrm{d}u$	$\frac{2}{\mathrm{erfc}(\sigma/2^{3/2})}$	e^γ	$-\frac{0.18}{\ln\sigma + 1}$	$0.73 - 2\sigma - 0.4\sigma^3$
Pareto[4], $\left(1+\xi\left(\frac{x}{\lambda}\right)\right)^{-\frac{1}{\xi}}$	$(1-\xi)^{-\frac{1}{\xi}}$	$\Gamma(1-\xi)^{\frac{1}{\xi}}$	1	$\frac{2\Lambda_1 + (\xi-3)\Lambda_\infty}{2(1-\ln 2)}$
PBF, $\left(1+\zeta\xi\left(\frac{x}{\lambda}\right)^\zeta\right)^{-\frac{1}{\xi\zeta}}$	$\left(1+\left(\frac{\mathrm{B}\left(\frac{1}{\zeta\xi}-\frac{1}{\zeta},\frac{1}{\zeta}\right)}{\zeta}\right)^\zeta\right)^{\frac{1}{\zeta\xi}}$	$\Gamma(1-\xi)^{\frac{1}{\xi}}$	$\zeta\xi + \frac{0.02}{\zeta} - 0.14\xi - 0.003$	$1.16 + 1.2(\Lambda_\infty - \Lambda_1)$
Dagum,[5] $1-\left(1+\frac{\xi}{\zeta}\left(\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)^{-\zeta}$	$\frac{1}{1-\left(\left(\zeta\mathrm{B}(1-\xi,\zeta+\xi)\right)^{-\frac{1}{\xi}}+1\right)^{-\zeta}}$	$\Gamma(1-\xi)^{\frac{1}{\xi}}$	1	$\frac{2\zeta\Lambda_1 - (2\zeta+\xi-1)\Lambda_\infty - \zeta - 1}{2\zeta(1-\ln 2)}$
EV2[6], $1-\exp\left(-\xi\left(\frac{x}{\lambda}\right)^{-\frac{1}{\xi}}\right)$	$\frac{1}{1-\exp\left(-\left(\Gamma(1-\xi)\right)^{-\frac{1}{\xi}}\right)}$	$\Gamma(1-\xi)^{\frac{1}{\xi}}$	1	$\frac{2(\Lambda_1 - \Lambda_\infty) - 1}{2(1-\ln 2)}$
EV1[7], $1-\exp\left(-\mathrm{e}^{-\frac{x}{\lambda}}\right)$	$1/(1-\exp(-\mathrm{e}^{-\gamma}))$	e^γ	1	$\frac{2(\Lambda_1 - \Lambda_\infty) - 1}{2(1-\ln 2)}$

[1] For all distributions the domain is $x \geq 0$, except for the normal and EV1, in which $x \in \mathbb{R}$. Linear transformations of x that change the lower bound have no effect on any of the characteristics given in the table.

[2] The exponential distribution can also be approximated as a special case of the Gamma or the Weibull distribution for $\zeta = 1$, and even by the simple approximation of equation (6.111), but the specific approximation with $\beta = 1$ is the best. Furthermore, it admits an exact equation, based on the relationships of Table 6.3, which is $\Lambda_p = \mathrm{e}^{H_p}/p$.

[3] The linear approximation (equation (6.111)) is also good for the gamma distribution but only for $\zeta < 1$.

[4] The Pareto distribution can also be approximated as a special case of the PBF distribution for $\zeta = 1$, and even by the linear approximation (equation (6.111)). However, the specific approximation given here with $\beta = 1$ is the best. Furthermore, it admits an exact equation, based on the relationships of Table 6.3, which is $\Lambda_p = \left((p+1-\xi)\,\mathrm{B}(1-\xi,p+1)\right)^{1/\xi}/p$.

[5] The Dagum distribution also admits an exact equation, based on the relationships of Table 6.3, which is $\Lambda_p = 1/\left(1-\left(\left(p\zeta B(1-\xi,p\zeta+\xi)\right)^{-1/\xi}+1\right)^{-\zeta}\right)p$.

[6] The EV2 distribution and all its expressions can be derived from the Dagum one for $\zeta \to \infty$; its exact equation, based on the relationships of Table 6.3, is $\Lambda_p = 1/\left(1-\exp\left(-\left(\Gamma(1-\xi)\right)^{-\frac{1}{\xi}}/p\right)\right)p$.

[7] The EV1 distribution and all its expressions can be derived from the EV1 one by substituting $x/\lambda \leftarrow x/\lambda + 1/\xi$ and letting $\xi \to 0$; its exact equation, based on the relationships of Table 6.3, is $\Lambda_p = 1/(1-\exp(-\mathrm{e}^{-\gamma}/p))\,p$.

Likewise, to evaluate, precisely or approximately, the tail moments for any distribution we introduce the tail-based Λ-coefficient of order p as:

$$\overline{\Lambda}_p := \frac{1}{p\,F\left(\overline{K}_p'\right)} \tag{6.118}$$

where $\overline{\Lambda}_p$ has similar properties with Λ_p and in particular varies only slightly with p. For $p = 1$ it is readily seen that

$$\overline{\Lambda}_1 = 1/F(\mu) = \Lambda_1/(\Lambda_1 - 1) \tag{6.119}$$

The limiting value $\overline{\Lambda}_\infty$ depends only on the lower-tail index ζ of the distribution:

$$\overline{\Lambda}_\infty = \Gamma(1+1/\zeta)^{-\zeta} \tag{6.120}$$

and its own limits are:

$$\lim_{\zeta\to 0}\overline{\Lambda}_\infty = 0, \qquad \lim_{\zeta\to\infty}\overline{\Lambda}_\infty = \mathrm{e}^{\gamma} \tag{6.121}$$

For the normal distribution, which is symmetric, $\overline{\Lambda}_\infty = \Lambda_\infty = \mathrm{e}^{\gamma}$, while for the exponential distribution $\overline{\Lambda}_\infty = 1$.

Like in the case of Λ-coefficients for noncentral moments, here again we give three different approximations. Proceeding from the least to most accurate, these are:

1. The *rule-of-thumb approximation* (zero-parameter):

$$\overline{\Lambda}_p = \begin{cases} 1, & \text{positively skewed distributions} \\ 2, & \text{negatively skewed or symmetric distributions} \end{cases} \tag{6.122}$$

2. The *linear approximation* (two-parameter):

$$\overline{\Lambda}_p \approx \overline{\Lambda}_\infty + \frac{\overline{\Lambda}_1 - \overline{\Lambda}_\infty}{p}, \qquad \frac{T\left(\overline{K}_p'\right)}{D} \approx \overline{\Lambda}_\infty p + \left(\overline{\Lambda}_1 - \overline{\Lambda}_\infty\right) \tag{6.123}$$

3. The *nonlinear approximation* (four-parameter, where in addition to $\overline{\Lambda}_1$ and $\overline{\Lambda}_\infty$ the parameters $\overline{\beta}$ and $\overline{B}$). Assuming $\overline{\beta} \geq 0$ we have:

$$\overline{\Lambda}_p \approx \overline{\Lambda}_\infty + \frac{\overline{A}}{p} + \overline{B}\ln\left(1+\frac{\overline{\beta}}{(p+1)^{\overline{\beta}}-1}\right), \quad \overline{A} = \overline{\Lambda}_1 - \overline{\Lambda}_\infty - \overline{B}\ln\left(1+\frac{\overline{\beta}}{2^{\overline{\beta}}-1}\right) \tag{6.124}$$

The most interesting cases are:

(a) $\beta = 0$, in which:

$$\overline{\Lambda}_p \approx \overline{\Lambda}_\infty + \frac{\overline{A}}{p} + \overline{B}\ln\left(1+\frac{1}{\ln(p+1)}\right),\ \ \overline{A} = \overline{\Lambda}_1 - \overline{\Lambda}_\infty - \overline{B}\ln\left(1+\frac{1}{\ln(2)}\right) \quad (6.125)$$

(b) $\beta = 1$, in which:

$$\overline{\Lambda}_p \approx \overline{\Lambda}_\infty + \frac{\overline{A}}{p} + \overline{B}\ln\left(1+\frac{1}{p}\right), \qquad \overline{A} = \overline{\Lambda}_1 - \overline{\Lambda}_\infty - \overline{B}\ln(2) \quad (6.126)$$

Table 6.7 gathers the equations for a number of customary distributions giving all quantities that support the approximation of the entire series of the tail Λ-coefficients $\overline{\Lambda}_p$. We stress that the equations of both Table 6.6 and Table 6.7 refer to continuous distributions. In the case of mixed distributions with a discontinuity $P_0 = 1 - P_1$ we should have in mind that the Λ_∞ retains its value as in the continuous case, while $\overline{\Lambda}_\infty = 0$. In practical applications, it is advisable to work with the continuous part of the distribution and treat the discontinuity separately.

Recapitulating the above discourse, the Λ-coefficients have the following important properties:

- They vary in a narrow range (close to 2 for noncentral K-moments or close to 1 for tail K-moments of positively skewed distributions) and this facilitates the determination of the complete series by only a couple of them (namely, Λ_1 and Λ_∞) and, if higher accuracy is required, by the additional parameters β and B (and likewise for the tail K-moments).
- They are well approximated by generic functions, irrespective of the particular distribution function.
- Their definition in terms of return period renders them suitable to study extremes.
- Also, their definition, in connection to their generic approximators, supports the indirect but quick determination of theoretical (true) values of K-moments of any order in absence of analytical relationships.

The last point suggests that we could follow a similar approach for the calculation of K-moments for higher q. Specifically, we can generalize (6.108) and define Λ-coefficients of orders (p, q) as:

$$\Lambda'_{pq} \coloneqq \frac{1}{(p-q+1)\left(1-F\left(K'^{1/q}_{pq}\right)\right)} \quad (6.127)$$

for noncentral moments and

$$\Lambda_{pq} \coloneqq \frac{1}{(p-q+1)\left(1-F\left(K^{1/q}_{pq}+\mu\right)\right)} \quad (6.128)$$

for central ones (and similarly for tail K-moments). We readily observe that for q = 1 we recover (6.108) and specifically:

$$\Lambda'_{p1} = \Lambda_{p1} = \Lambda_p \quad (6.129)$$

as $K'_p = K_{pq} + \mu$. However, for $q > 1, \Lambda'_{pq} \neq \Lambda_{pq}$ except in the limit as $q \to \infty$. Like the Λ-coefficients of q = 1, Λ_{pq} will also vary in a narrow range. In particular, as the tail index of the distribution of $\underline{x}^q$ will be $q\xi$, the limit $\Lambda'_{\infty q} = \Lambda_{\infty q}$ can be readily determined. Thus, with reference to the distributions in Table 6.6 and for q = 2, those that have tail index zero and $\Lambda_\infty = \mathrm{e}^\gamma$, will also have $\Lambda'_{\infty 2} = \Lambda_{\infty 2} = \mathrm{e}^\gamma = 1.781$. Furthermore, those that have $\Lambda_\infty = \Gamma(1-\xi)^{\frac{1}{\xi}}$ will have $\Lambda'_{\infty 2} = \Lambda_{\infty 2} = (\Gamma(1-2\xi))^{1/2\xi}$ and more generally, for $q < 1/\xi$, $\Lambda'_{\infty q} = \Lambda_{\infty q} = (\Gamma(1-q\xi))^{1/q\xi}$.

Table 6.7 Characteristic parameters for accurate approximations of tail Λ-coefficients and tail K-moments for several customary distributions.

Distribution[1] and tail function, $\overline{F}(x)$	$\overline{\Lambda}_1$	$\overline{\Lambda}_\infty$	$\overline{\beta}$	$\overline{B}$
Normal, $\int_x^\infty \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma}\mathrm{d}u$	2	e^γ	0	0.73
Exponential[2], $\mathrm{e}^{-x/\lambda}$	$\frac{\mathrm{e}}{\mathrm{e}-1}$	1	1	$\frac{2\overline{\Lambda}_1 - 3}{2(1-\ln 2)} = 0.267$
Gamma[3], $\frac{\Gamma_{x/\lambda}(\zeta)}{\Gamma(\zeta)}$	$\frac{\Gamma(\zeta)}{\Gamma(\zeta) - \Gamma_\zeta(\zeta)}$	$\Gamma(1+1/\zeta)^{-\zeta}$	0	$2.27 - 2\ln(2.2 - \overline{\Lambda}_1) - 2\overline{\Lambda}_1$
Weibull, $\exp\left(-\left(\frac{x}{\lambda}\right)^\zeta\right)$	$\frac{\mathrm{e}^{\left(\Gamma\left(1+\frac{1}{\zeta}\right)\right)^\zeta}}{\mathrm{e}^{\left(\Gamma\left(1+\frac{1}{\zeta}\right)\right)^\zeta} - 1}$	$\Gamma(1+1/\zeta)^{-\zeta}$	1	$\frac{2\overline{\Lambda}_1 - 2\overline{\Lambda}_\infty - 1}{2(1-\ln 2)}$
Lognormal, $\int_x^\infty \frac{\exp\left(-\frac{(\ln(u/\lambda))^2}{2\sigma^2}\right)}{\sqrt{2\pi}\,\sigma\,u}\mathrm{d}u$	$\frac{2}{2 - \mathrm{erfc}(\sigma/2^{3/2})}$	e^γ	0	$0.73 + 2.4\sigma - 0.7\sigma^{3/2}$
Pareto[4], $\left(1+\xi\left(\frac{x}{\lambda}\right)\right)^{-\frac{1}{\xi}}$	$\frac{1}{1-(1-\xi)^{\frac{1}{\xi}}}$	1	1	$\frac{2\overline{\Lambda}_1 + \xi - 3}{2(1-\ln 2)}$
PBF[5], $\left(1+\zeta\xi\left(\frac{x}{\lambda}\right)^\zeta\right)^{-\frac{1}{\xi\zeta}}$	$\frac{1}{1-\left(1+\left(\frac{\mathrm{B}\left(\frac{1}{\zeta\xi}-\frac{1}{\zeta},\frac{1}{\zeta}\right)}{\zeta}\right)^\zeta\right)^{-\frac{1}{\zeta\xi}}}$	$\Gamma(1+1/\zeta)^{-\zeta}$	1	$\frac{2\overline{\Lambda}_1 + \overline{\Lambda}_\infty(\zeta\xi+\xi-2) - \zeta\xi - 1}{2(1-\ln 2)}$

[1] For all distributions it is assumed that $x \geq 0$, except for the normal, in which $x \in \mathbb{R}$. The exponential, Weibull, Pareto and PBF distributions admit exact equations, based on the relationships of Table 6.2,

[2] The equations for the exponential distribution are identical to those of Weibull for $\zeta = 1$ and can also be approximated as a special case of the Gamma or the distribution for $\zeta = 1$, and even by the linear approximation (equation (6.123)), but the specific approximation with $\beta = 1$ is the best. Furthermore, its exact equation, based on the relationships of Table 6.3, is $\Lambda_p = 1/\left(1 - e^{-1/p}\right)p$.

[3] The linear approximation (equation (6.123)) is also good for the gamma distribution but only for $\zeta < 1$.

[4] The Pareto distribution can also be derived as a special case of the PBF distribution for $\zeta = 1$.

[5] The approximation given for PBF setting $\overline{\beta} = 1$ is good for $\zeta\xi \leq 1$ and needs further study for $\zeta\xi > 1$ ($\overline{\beta} > 1$).

Now if we consider a linear transformation of the variable $\underline{x}$, i.e., $\underline{w} = b\underline{x} + c$, then the following equations, whose proof is contained in Appendix 6-IX, hold:

$$K_{pq}[\underline{w}] = b^q K_{pq}[\underline{x}], \qquad \Lambda_{pq}[\underline{w}] = \Lambda_{pq}[\underline{x}] \tag{6.130}$$

In other words, the central Λ-coefficients are invariant under linear transformations of the variables. This enables the theoretical calculation of Λ_{pq} by linearly transforming the original variable $\underline{x}$ to one with a simpler expression of distribution (e.g., with lower bound or mean zero and unit scale parameter) and performing the calculations on $\underline{w}$. Then K_{pq} can be determined from (6.130) through the central K-moments of the simpler distribution of $\underline{w}$. The noncentral K-moments can be determined from the central ones; the relationships are contained in Appendix 6-II.

Digression 6.C: The behaviour of the normal distribution

Here we will find approximations of the K-moments and the Λ-coefficients of the normal distribution. Using the approximation of the normal distribution by (5.44) and its quantile function by (5.45), we find in Appendix 6-X that:

$$K'_p = B_p + \sum_{k=1}^{p} \binom{p}{k} (-1)^k B_k, \qquad \Lambda_p \approx \frac{2}{p \exp\left(-\frac{2}{3}K'_p\left(1+\frac{2}{3}K'_p\right)\right)} \tag{6.131}$$

where

$$B_p := p\left(\int_0^{1/2} x(F)\, F^{p-1}\, \mathrm{d}F\right) \approx -\frac{3\sqrt{\pi}\mathrm{e}^{p/4}\,\mathrm{erfc}\left(\sqrt{p}/2\right)}{2^{p+2}} \tag{6.132}$$

These approximations, depicted in Figure 6.12 in comparison to the exact values, are rather satisfactory. However, evaluation of (6.131) beyond p > 100 is problematic because of the numerical instability of the binomial transform contained in it.

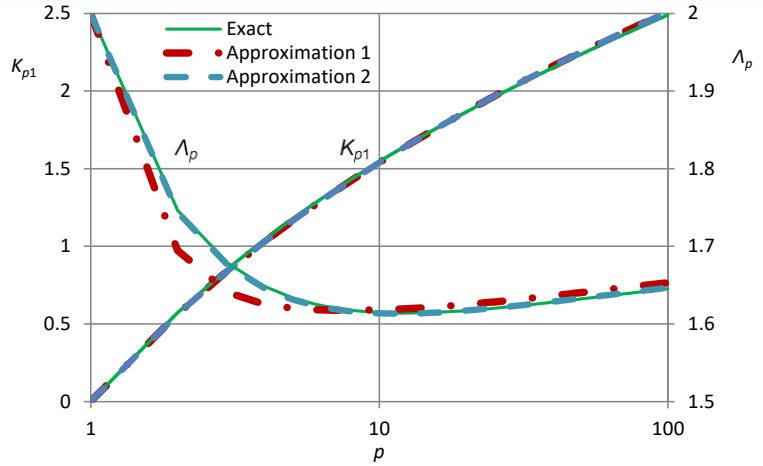

Figure 6.12 Comparison of approximate and exact values of K-moments and Λ-coefficients of order q = 1 for the normal distribution. Approximations 1 and 2 are calculated by equations (6.131) and (6.133), respectively, and the exact values are calculated by numerical integration.

A better approximation can be obtained in a simpler manner from equation (6.114) with coefficients Table 6.6.This yields:

$$\Lambda_p \approx 1.781 + \frac{0.871}{p} - 0.73 \ln\left(1 + \frac{1}{\ln(p+1)}\right) \tag{6.133}$$

As seen in Figure 6.12 for p up to 100 and Figure 6.11 for much higher p, this latter approximation is very accurate and thus preferable over that of equation (6.131).

Digression 6.D: Do return periods assigned by order statistics and K-moments differ?

Both frameworks of order statistics and K-moments provide means to assign empirical return periods from an observed sample. In order statistics we use equation:

$$\frac{T_{(i:n)}}{D} = \frac{n+B}{n-i+A} \tag{5.57}$$

for specified A and B, while for the K-moment of order (p,1) we use equation:

$$G(p) \coloneqq \frac{T\left(K_p'\right)}{D} = p\Lambda_p \tag{6.108}$$

where specific forms of Λ_p and hence of the function $G(p)$ have been extensively discussed in section 6.14. Here we use the linear approximation:

$$G(p) = \Lambda_\infty p + (\Lambda_1 - \Lambda_\infty) \tag{6.134}$$

The two approaches for assigning return periods are comparable at certain return periods, namely those corresponding to integer i between 1 and n in equation (5.57), for which the return period $T_{(i:n)}$ is defined. Given $T_{(i:n)}$ there is a specific p such as $T\left(K_p'\right) = T_{(i:n)} = T$. This is given as $p = G^{-1}\left(T_{(i:n)}/D\right)$.

Thus, to each of the n values $T_{(i:n)}$ in a sample we can assign values x in two different ways:

1. $x_{(i:n)}$, i.e. the ith smallest of the n observations;
2. $x_{(i,n)}^{\mathrm{K}} \coloneqq K_p'$, where $p = G^{-1}\left(T_{(i:n)}/D\right)$ and K_p' is estimated from the entire sample by equations (6.65)- (6.67).

These two values should be close to each other in general. We can make them precisely equal if we use the approximation (6.134) for the latter along with the approximation of A and B of the former corresponding to unbiased quantile (equation (5.70)). Equating the two we have:

$$\Lambda_\infty p + (\Lambda_1 - \Lambda_\infty) = \frac{n + \Lambda_1/\Lambda_\infty - 1}{n - i + 1/\Lambda_\infty} \tag{6.135}$$

and hence:

$$p = 1 + \frac{i - 1 - (\Lambda_1 - 1)(n-i)}{\Lambda_\infty(n-i)+1} \Leftrightarrow i = n - \frac{n-p}{\Lambda_1 + \Lambda_\infty(p-1)} \tag{6.136}$$

Note that for $p = 1, i = n - (n-1)/\Lambda_1$ (e.g. $i \approx n/2$ if $\Lambda_1 = 2$) and for $p = n, i = n$.

This cannot work for $p < 1$ or for $i = n - (n-1)/\Lambda_1$. For small i we can continue with tail moments of order $\overline{p} \geq 1$, assuming return period of nonexceedance of $\overline{K}_p'$ equal to $\overline{T} = \overline{\Lambda}_\infty \overline{p} + \left(\overline{\Lambda}_1 - \overline{\Lambda}_\infty\right)$, with $\overline{T} = 1/(1-1/T)$. Thus:

$$\frac{1}{1-1/\left(\overline{\Lambda}_\infty \overline{p}+\left(\overline{\Lambda}_1-\overline{\Lambda}_\infty\right)\right)}=\frac{n+\Lambda_1/\Lambda_\infty-1}{n-i+1/\Lambda_\infty} \tag{6.137}$$

and taking also (6.119) into account and setting $\overline{\Lambda}_\infty = \Lambda_\infty\,(1+\delta)/(\Lambda_1-1)$ (for some δ which, as we will see, turns out to be very small) we find:

$$\overline{p}=1+\frac{(\Lambda_1-1)(n-i)-i+1}{(1+\delta)(\Lambda_\infty(i-1)+\Lambda_1-1)}, i=1+\frac{(\Lambda_1-1)(n-1-(1+\delta)(\overline{p}-1))}{\Lambda_1+\Lambda_\infty(1+\delta)(\overline{p}-1)} \tag{6.138}$$

For a symmetric distribution, $\Lambda_1 = 2$ and $\overline{\Lambda}_\infty = \Lambda_\infty$ and thus $\delta = 0$. For the skewed distributions contained in Table 6.7 by comparing with Table 6.6 and investigating numerically, it can be verified that δ is small, of the order of 2 % - 5 %. Thus, neglecting δ we find:

$$\overline{p}=1+\frac{(\Lambda_1-1)(n-i)-i+1}{\Lambda_\infty(i-1)+\Lambda_1-1}, \qquad i=1+\frac{(n-\overline{p})(\Lambda_1-1)}{\Lambda_1+\Lambda_\infty(\overline{p}-1)} \tag{6.139}$$

Based on the above results, for a given moment order p a *quick-and-dirty estimate* of the noncentral moment K_p' is the value $x_{(i:n)}$, the *i*th smallest value of the sample, with *i* determined from equation (6.136). Likewise, if *i* is determined from equation (6.139) for a given $\overline{p}$, then the value $x_{(i:n)}$ is a *quick-and-dirty estimate* of the tail K-moment $\overline{K}_{\overline{p}}'$. In essence, the quick-and-dirty K-moments approach is equivalent to the order statistics one.

For numerical illustration we use the Pareto distribution with tail index ξ = 0.15 and other parameters as shown in the caption of Figure 6.13. In this case $T_{(i:n)}$ will be determined either from the unbiased estimator of the quantile of the Pareto distribution (corresponding to unbiased quantile, case VI of Table 5.5) with ξ = 0.15:

$$\frac{T_{(i:n)}}{D}=\frac{n+0.452}{n-i+0.491}$$

For the noncentral K-moment of order $(p, 1)$ and for $p \geq 1$ we use linear approximation

$$G(p)=\Lambda_\infty p+(\Lambda_1-\Lambda_\infty)=2.035p+0.92$$

Thus,

$$p=\frac{T/D-0.92}{2.035}=0.491\,T/D-0.452$$

This is precisely the result we get from the quantile-unbiased estimator of the order statistics if we set $n = i = p$. For return periods $T/D < \Lambda_1 = 2.95$ the resulting p is smaller than 1. This is not a problem as the definition of K-moments has already been extended for non-integer order p except that the linear approximation is no longer accurate.

Therefore, we use the tail K-moment of order $(\overline{p}, 1)$; for $\overline{p} \geq 1$ we apply the linear approximation:

$$\overline{G}(p)=\overline{\Lambda}_\infty \overline{p}+\left(\overline{\Lambda}_1-\overline{\Lambda}_\infty\right)=\overline{p}+0.512$$

Thus,

$$\overline{p}=\overline{T}/D-0.512=1/(1-\ D/T)\ -0.512$$

Note that for $\overline{p}=n$, $1/(1-\ D/T)=\overline{p}+0.512$ or, after the algebraic manipulations, $T/D = (n+0.512)/(n-0.488)$. This is slightly different from $(n+0.452)/(n-0.509)$, as given by the order statistics approach. For $n = 100$, the former formula gives $T/D = 1.0100$ and the latter 1.00966, a difference of 0.04 %. This slight difference is due to the fact that δ is not exactly zero (namely, $\delta = \left(\overline{\Lambda}_\infty/\Lambda_\infty\right)(\Lambda_1-1)-1 = (2.035/1)\times 0.512-1 = 0.042$).

Alternatively, as the Pareto distribution yields an exact solution (see Table 6.2), we could use that instead of the approximations. Namely, this is:

$$G(p) = \left((p + 1 - \xi)\,\mathrm{B}(1 - \xi, p + 1)\right)^{1/\xi}$$

and is valid for all p, integer or not, including for $p < 1$.

Simulation results for our example are shown in Figure 6.13. On the average, the order-statistics and the K-moment approaches give equivalent results and equally good, both in terms of averages and uncertainty limits (prediction limits). However, if we focus on a single realization, such as the one also shown in Figure 6.13, the K-moments approach yields a smooth arrangement of empirical points, while that of the order-statistics approach indicates a greater variability and a rougher arrangement. The reason is that in the K-moments approach each K-moment value is a weighted average of several points, while in the order statistics only one value is used each time. As regards the order-statistics approach, we note that there would be a substantial difference in the largest value if we adopted the Weibull plotting position formula, which, as explained in section 5.6, we have deemed inappropriate.

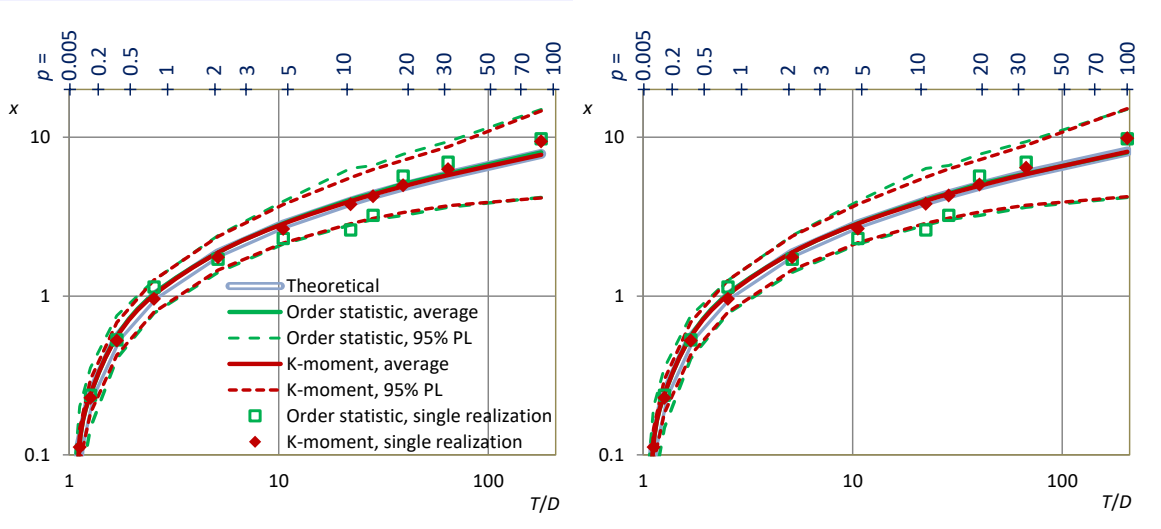

Figure 6.13 Simulation results of empirical return periods assigned to Pareto quantiles (for tail index $\xi = 0.15$, scale parameter $\lambda = 1$ and lower bound zero). Averages and prediction limits (PL) were calculated from 200 simulations each with $n = 100$. The curves of averages for both the order statistics and the K-moment approaches are indistinguishable from the theoretical curves. The return periods $T_{(i:n)}$ were assigned by **(left)** the generic option of unbiased $\ln T$ and **(right)** the unbiased quantile option. The correspondence between the K-moment of order p and the return period T is also shown through the upper horizontal axis. The plots of a single realization are also shown (but for part of the empirical points to avoid an overcrowded graph).

The above illustration helps us to formulate a few focus points of more general validity:

1. The largest value in an observed sample of size n does not have a return period of about n time units as commonly assumed. Rather, it is the order p of the maximum K-moment that is equal to n. Thus, the return period of the maximum observation is about $\Lambda_\infty n$, usually 1.8 n to 2 n.
2. In both approaches, the order-statistics and K-moments, the results are virtually (or even exactly) the same in terms of expected values and uncertainty.
3. While with order statistics we can empirically assign return periods only to the observations, thus designating only n specific values of return period, with K-moments there is no such restriction. Rather, we can empirically assign a return period to any quantile value between the smallest and the largest observation.
4. Because of the more accurate formulae for K-moments discussed in section 6.14, in comparison to those of order statistics discussed in section 5.6, the return periods empirically assigned by the K-moments approach are typically more accurate or at most equivalent to those of order statistics.

5. In addition, while with order statistics only one observation is used for each assignment of return period, in the K-moments approach each K-moment value is a weighted average of several observations, thus contributing to the accuracy of estimation.

For all these reasons the K-moments approach is deemed preferable. We note though that computationally it is more demanding.

6.15 Extreme-oriented estimation via K-moments

As we have discussed in section 2.19, using the entire data set to model extremes is preferable than using block (e.g. annual) maxima. Usually, the values above a specified threshold are chosen, while the remaining values are discarded in modelling. However, if we use the K-moments there is no need to set a threshold. All values can be used, but as we have seen in section 6.9, in a sample of size n the estimation of the K-moment of orders $(p, 1)$ relies only the $n - p + 1$ largest values, thus rendering thresholding unnecessary.

In Chapter 4 we have discussed two different approaches for fitting distribution functions to data. The method of maximum likelihood is well reasoned and is based on an optimization logic. In contrast, the method of moments is based on solving equations and is not quite rigorously argued. Assuming that we fit a two-parameter model (say, a two-parameter gamma distribution) the method of moments uses the first two classical (noncentral) moments and determines the two parameters by equating the sample moments to the theoretical moments of the distributions. One could raise two major questions on the logic of the method:

1. *Why use the first and second moments and not, say, the second and third?* One may easily justify the standard choice of using the lowest possible order of moments by the fact that higher moments are less accurately estimated. On the other hand, one may counter that, when we are interested in extremes, these are better reflected in higher-order moments. It is well known that a model can hardly be a perfect representation of reality. Thus, we cannot expect that a good model fitting on the first and second moment would be equally suitable for the distribution tail, i.e. the behaviour on extremes.
2. *We use two moments and not more?* The standard answer, that two equations suffice to find two unknowns, may be adequate from a theoretical mathematical point of view but it is not from an empirical and engineering one. (The saying is that to draw a straight line a mathematician needs two points but an engineer needs three). Certainly, an optimization framework (as in maximizing likelihood or in minimizing fitting error) is much preferable and superior to an equation solving method.

Having introduced the concept of K-moments we have already seen several advantages, which are particularly strong for an extreme-oriented modelling. The following three properties are highlighted:

1. They are knowable with unbiased estimators (from samples) for high orders, up to the sample size n, while the estimation uncertainty is by orders of magnitude lower than in the classical moments (section 6.9 and Digression 6.A).

2. The estimators can explicitly (albeit approximately) take into account any existing dependence structure (section 6.12).
3. The K-moment values, can directly be assigned return periods, through Λ-coefficients, similar to what happens with order statistics, but with some advantages over the latter (section 6.14).

With the above points in mind, we can now formulate our extreme-oriented distribution fitting based on these postulates.

1. We use all n data.
2. From the data we estimate all K-moments, from orders 1 to n (alternatively, we could choose a subset of them, e.g. 100 values of p arranged in a geometric progression from 1 to n).
3. We make a climacogram of data and assess if there is long range dependence. In the case there is, we adapt the moment orders using equation (6.95).
4. We assume a model, i.e., a marginal distribution function with some parameters and for that model we establish the relationship between moment order p and Λ-coefficients Λ_p, which define return periods $T/d = 1/p\Lambda_p$.
5. We choose a range of return periods depending on our focus on extremes and over that range we fit the model so as to minimize the mean square errors (possibly weighted) between the empirical and theoretical K-T relationships.

In all points the moment order q is assumed to be 1.

A possible criticism on using high order K-moments (up to order n) is that this gives higher weight to the highest observations, which are more uncertain than the low ones. This criticism would be valid if the true distribution function was known to be the one chosen as a model for the real-world process studied. But this is hardly the case. Let us assume that in the time series of flow observations we have three very high values and that we have chosen a certain model, e.g. a Lognormal distribution. How can we be sure are that the model is correct? If we are not sure (which actually is always the case), and if we are to design a certain engineering construction, would we prefer a fitting of the chosen model that is consistent with theoretical considerations, e.g. based on the maximum likelihood method, even if this yields a departure for the three high values? Or would we feel safer if our fitting represents well the three high values?

The framework is illustrated in Digression 6.E for rainfall extremes in Bologna.

Digression 6.E: Extreme-oriented estimation of rainfall in Bologna

The record of daily rainfall in Bologna has already been discussed in section 1.3. The period of observation is T_{obs} = 206 years and it includes n = 19 426 nonzero rainfall depths (all other daily rainfall values are zero. Therefore, the time reference of defining the distribution function (conditional on $\underline{x} > 0$) and return period is:

$$D = T_{\text{obs}}/n = 206 \text{ years}/19426 = 0.01060 \text{ years} = (1/94.3) \text{ years}$$

(which means 94.3 rain days per year). As already discussed in Digression 2.H, the average daily rainfall during rain days is 7.2 mm and the maximum 155.7 mm.

Here we will see several options for fitting a marginal distribution to nonzero daily rainfall, we will study the differences among them, and we will trace a nearly optimal option. Firstly, for the

sake of illustration we intentionally choose the simplest and blatantly unsuitable model, the one-parameter exponential distribution:

$$F(x) = 1 - \mathrm{e}^{-x/\lambda}$$

In this case, one moment suffices to estimate the single (scale) parameter λ —but which moment to choose? The standard option is to choose the first moment, the mean, so that $\lambda = \mu = 7.2$ mm. This is the same if we choose classical moments, L-moments, etc. The maximum likelihood method also results in the same estimate of λ.

What if we chose a moment higher than 1 for the estimation? The exponential distribution is quite convenient and yields simple analytical relationships for all types of moments. Thus, the theoretical K- and classical moments are:

$$K_p = (H_p - 1)\lambda, \qquad K_{p2} = \left(\left(H_{p-1} - 1\right)^2 + H^{(2)}_{p-1}\right)\lambda^2, \qquad K_{pp} = \mu_p = (!\,p)\lambda^p$$

where H_p is the pth harmonic number, $H^{(2)}_p$ is the pth harmonic number of order 2 and $!\,p$ is the subfactorial of p. If we estimate the sample moment $\hat{K}_p$, $\hat{K}_{p2}$ or $\hat{\mu}_p$ and equate it to the respective theoretical quantity as above, we obtain another estimate of λ. The resulting estimates are plotted in Figure 6.14 (left) vs. moment order p, whilst some of the resulting fitted distributions are plotted in Figure 6.14 (right) in comparison to the empirical distribution.

It is evident in Figure 6.14 that the moment order p affects the fitting dramatically. Specifically, the scale parameter λ increases with increasing p and q. If we wish to model maxima, it is better to fit based on the thousandth K-moment than on the first! This clearly shows that, as far as extremes are concerned, high-order K-moments are preferable to low-order K-moments.

In a next step, we fit and compare both the exponential and the Pareto distribution in two cases: for the entire data set (size: 19426 for 206 years) and for values over threshold (VOT), where the threshold (47 mm) was chosen so that the sample contain 206 values (size equal to the number of years). Specifically, the two distribution functions are, respectively,

$$F(x) = 1 - \mathrm{e}^{-(x/\lambda - \varepsilon)}, \qquad F(x) = 1 - (1 + \xi\,(x/\lambda - \varepsilon))^{-1/\xi}$$

Comparisons of empirical and theoretical distributions are depicted in Figure 6.15. The exponential distribution was fitted with one parameter (setting $\varepsilon = 0$) for all data and with two parameters for the VOT case. The Pareto distribution was fitted with two parameters (setting $\varepsilon = 0$) for all data and three for the VOT case. Obviously, physical consistency demands that $\varepsilon = 0$ but violation of this condition can improve the fitting.

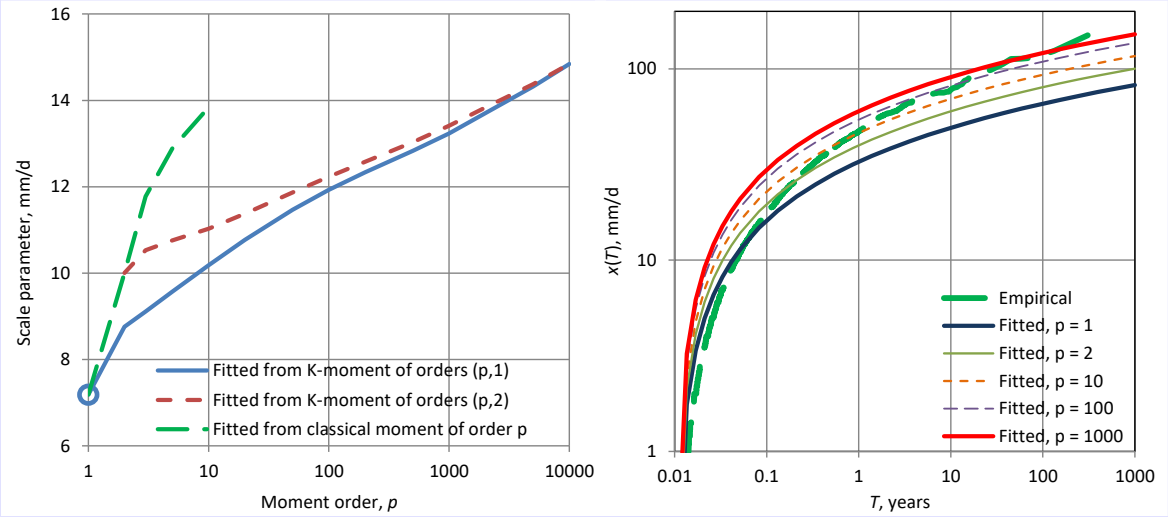

Figure 6.14 (left) Estimate of the scale parameter λ of the exponential distribution from the pth moment, fitted on the Bologna rainfall record; the circle corresponds to the standard estimate by any of the methods of classical moments, L-moments, K-moments and maximum likelihood. **(right)** Resulting fitted distribution, as a graph of x vs. T, for the indicated values of p; the empirical distribution is calculated by formula III of Table 5.5.

For the Pareto case, the methods of moments and L-moments were used, with the lowest orders (1, 2 or 3, depending on the number of parameters of the theoretical distribution). Figure 6.15 shows that there is not a clear winner between the moments and L-moments methods. When the entire data set is used, the fitting is quite unsatisfactory for the distribution tail (extremes). Yet the classical moments fitting shows better performance than the L-moments. If we use the sample over threshold and the three-parameter Pareto, classical moments and L-moments give fittings very close to each other (with slight advantage of the latter on both small and high values). Among the two options, all data and VOT option, the latter gives a better fitting on the maxima—but at the expense of an additional parameter and a physically inconsistent nonzero minimum.

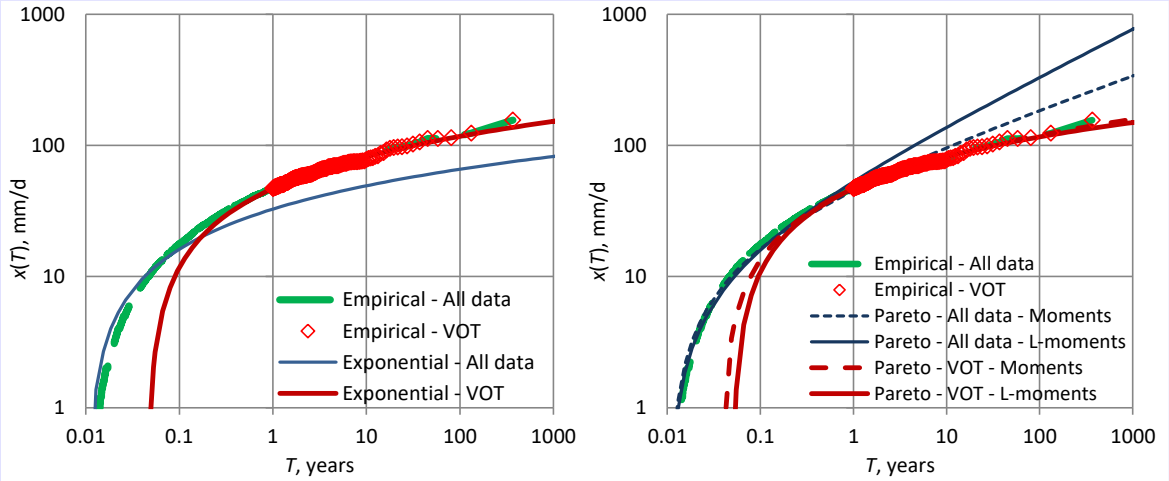

Figure 6.15 Fitting (in terms plot of quantile x vs. return period T) of the exponential **(left)** and Pareto **(right)** distribution on the Bologna daily rainfall record by the indicated methods; the empirical distribution is calculated by formula III of Table 5.5.

Let us now examine two questions: Can we improve the first option, so that the lower bound be zero for physical consistency? Can we use the entire data set and fit on the distribution tail? The replies to both questions are positive and in fact the first one has already been discussed in Digression 2.H. Here we will study them exclusively using the K-moments, both for assigning empirical return periods and for distribution fitting. We assume Pareto distribution with zero lower bound:

$$F(x) = 1 - (1 + \xi\, x/\lambda)^{-\frac{1}{\xi}}, \qquad \frac{T(x)}{D} = (1 + \xi\, x/\lambda)^{\frac{1}{\xi}}, \qquad x = \lambda \frac{(T/D)^{\xi} - 1}{\xi}$$

The estimated K-moments have return period:

$$\frac{\hat{T}\left(\hat{K}'_p\right)}{D} = p\Lambda_p = \left((p + 1 - \xi)\, \mathrm{B}(1 - \xi, p + 1)\right)^{\frac{1}{\xi}}$$

(With negligible error, we could also use the approximation $\hat{T}\left(\hat{K}'_p\right)/D = \Lambda_\infty(p - 1) + \Lambda_1$.) We estimate the parameters ξ and λ by minimizing the mean square error of the logarithms of the empirical $\hat{T}\left(\hat{K}'_p\right)$ from the theoretical $T(\hat{K}'_p)$. (Minimizing the error of $\hat{K}'_p$ with respect to K'_p, without reference to T, is another possibility.) We calculate the error for a range of T from 2 years to the maximum value that the sample size allows. The fitted parameters are shown in Table 6.8. The fitted distribution function is depicted in Figure 6.16, which shows a perfect agreement of theoretical and empirical curves for $T > 1$ year (the two curves are indistinguishable). For comparison, empirical curves for the order statistics are also plotted but these have not been used at any step of the fitting procedure.

The model shown in Figure 6.16 (right) is quite satisfactory, almost perfect, as far as the distribution tail is concerned. The proximity of empirical and theoretical curves is remarkable, as is the physical consistency and parsimony of the model, which contains only a scale parameter and a tail index.

By changing the parameters, we can obtain a better fit on the entire set of values but at the cost of worsening that on large return periods (this has been already done in Figure 2.4). Alternatively, by adding one parameter to the theoretical distribution function, we can obtain a model applicable for the entire range of rainfall depth, without compromising the performance on large return periods. Namely, we use the Pareto-Burr-Fuller (PBF) distribution, again with zero lower bound:

$$F(x) = 1 - \left(1 + \zeta\xi (x/\lambda)^{\zeta}\right)^{-\frac{1}{\zeta\xi}}, \qquad \frac{T}{D} = \left(1 + \zeta\xi \left(\frac{x}{\lambda}\right)^{\zeta}\right)^{\frac{1}{\zeta\xi}}, \qquad x = \lambda\left(\frac{(T/D)^{\zeta\xi} - 1}{\zeta\xi}\right)^{\frac{1}{\zeta}}$$

Now for the fitting we use the same estimation procedure as above but calculate the error on the entire range of values. However, we give less importance to the low quantiles by weighting the square error at each point with the quantile itself. In this case we have two different fitting variants. In the first we do not apply any constraint in parameters and in the second we keep the tail index ξ as estimated for the Pareto distribution (ξ =0.098, in order not to distort the good fitting on the tail). The parameters are shown in Table 6.8. A perfect fit of the model and the empirical curve for the entire range of return periods is seen in Figure 6.17 (referring to the first variant).

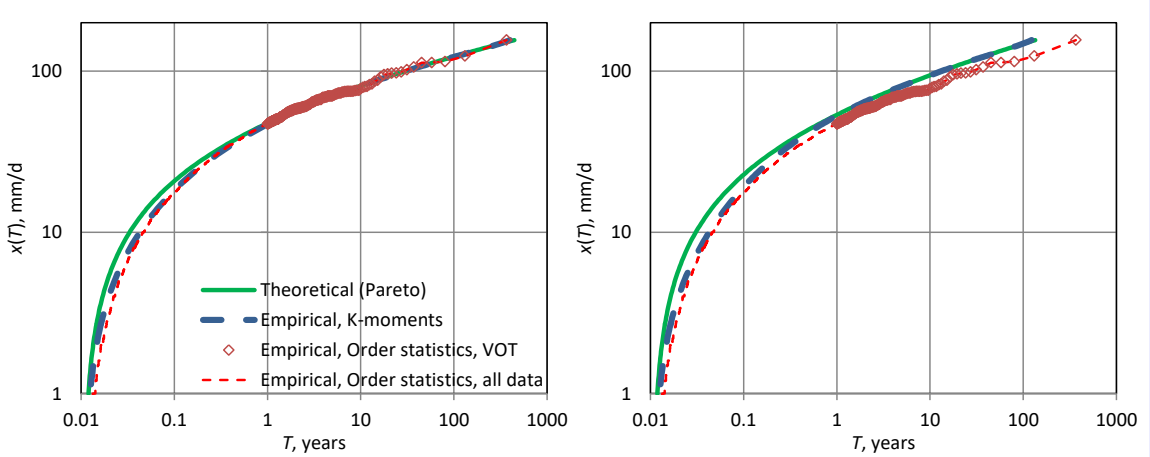

Figure 6.16 Fitting (in terms plot of quantile x vs. return period T) of the Pareto distribution on the Bologna daily rainfall record by the indicated methods **(left)** assuming independence and **(right)** accounting for long-range dependence; the curves of theoretical and empirical K-moments are indistinguishable for $T > 1$ year. The empirical distribution from order statistics (calculated by formula III of Table 5.5 and not considering dependence so that it is the same in both panels) is also plotted for comparison.

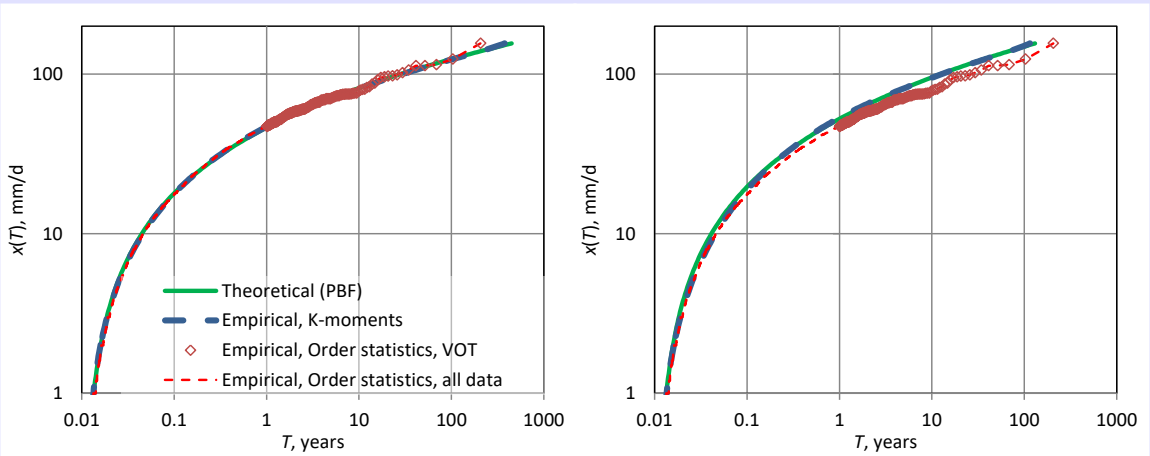

Figure 6.17 Same as Figure 6.16, but for the Pareto-Burr-Feller distribution fitted for the entire range of return period **(left)** assuming independence and **(right)** accounting for long-range dependence; note that empirical return periods based on order statistics do not consider dependence and thus they are the same in both panels.

Referring to the numerical results in Table 6.8, we can provide a final comparison focusing on the question, which is the design value (distribution quantile) for return period T = 1000 years.

- If we followed the dominant approach of using Gumbel (EV1) distribution on annual maxima, which is equivalent to using exponential tail for the parent distribution, then the design value would be 152.4 mm. Note that this is lower than the record observation, which is 155.7 mm.
- If we changed the distribution from exponential to Pareto, thus being more consistent to recent findings (see also discussion and references in Digression 2.H and in Digression 8.F), the design value would be 173.7 mm, a 15% increase.
- By assuming Pareto tail and also accounting for dependence, the design value becomes 218.3 mm, a 43% increase in comparison to the initial estimate of 152.4 mm.
- Additional changes can appear if we use a model with more parameters, such as the PBF distribution; however, it is not clear if these changes would increase or decrease the design value, as the direction depends on additional assumptions. As the additional parameters result in higher uncertainty, it may be preferable to use the more parsimonious Pareto model.

Note that the increases due to methodological improvements of a consistent stationary stochastic framework, are much larger than those usually published in modern literature identifying increases attributed to global warming (see Koutsoyiannis, 2020b). The finding here may lead to the following suggestions for fine scale rainfall extremes:

- Assume stationarity.
- Use Pareto tail.
- Take dependence into account.
- Fit based on K-moments of high order.

Table 6.8 Comparison of model parameters and resulting quantile values for characteristic return periods for the different models fitted.

Distribution	Fitting assumptions[1]	Tail index, ξ	Scale parameter, λ (mm)	Location parameter, ε	Lower-tail index, ζ	T (years)		
						100	1000	10000
Exponential	I-VOT	0	15.27	3.07	1	117.2	152.4	187.6
Pareto	I-HT	0.098	8.30	0	1	122.9	175.5	241.3
PBF	I-A	0.042	6.12	0	0.786	124.1	173.7	230.7
PBF	I-A	0.098	7.07	0	0.928	124.0	179.9	250.5
Pareto	D-HT	0.120	8.85	0	1	147.7	218.3	311.4
PBF	D-A	0.058	6.49	0	0.775	148.7	213.5	290.8
PBF	D-A	0.120	6.98	0	0.891	151.6	229.7	333.9

[1] I: independence; D: dependence; A: all data; VOT: values above threshold; HT: high return period, $T \geq 1$ year.

Appendix 6-I: The binomial identity and the binomial and Bernoulli transforms

The binomial identity is:

$$(x+y)^p = \sum_{i=0}^{p} \binom{p}{i} x^i y^{p-i} = y^p \sum_{i=0}^{p} \binom{p}{i} \left(\frac{x}{y}\right)^i \tag{6.140}$$

where p is a nonnegative integer and x and y are any numbers. The identity can be expanded for any real (or even complex) p. Assuming $|x| < |y|$ (to guarantee convergence), the identity takes the form:

$$(x+y)^p = \sum_{i=0}^{\infty} \binom{p}{i} x^i y^{p-i} = y^p \sum_{i=0}^{\infty} \binom{p}{i} \left(\frac{x}{y}\right)^i \tag{6.141}$$

and since for integer p and for $i > p$ the binomial coefficient $\binom{p}{i}$ is zero, (6.140) is readily recovered from (6.141).

In our stochastic context, it is interesting to study the case where x represents a stochastic variable and y a number. We thus get the following characteristic cases:

- $x \to \underline{x}, y = 1$

$$\left(\underline{x}+1\right)^p = \sum_{i=0}^{\infty} \binom{p}{i} \underline{x}^i \tag{6.142}$$

- $x \to -\underline{x}, y = 1$

$$\left(1-\underline{x}\right)^p = \sum_{i=0}^{p} \binom{p}{i} (-1)^i \underline{x}^i \tag{6.143}$$

- $x \to P\underline{x}, y = 1-P$, typically for $0 < P < 1$:

$$\left(P\underline{x}+1-P\right)^p = \sum_{i=0}^{p} \binom{p}{i} P^i (1-P)^{p-i} \underline{x}^i \tag{6.144}$$

If we take expected values on (6.143), and denote $a_i \coloneqq \mathrm{E}\left[\underline{x}^i\right], b_i \coloneqq \mathrm{E}\left[\left(1-\underline{x}\right)^i\right]$ we can write:

$$b_p = \sum_{i=0}^{p} \binom{p}{i} (-1)^i a_i \tag{6.145}$$

This latter equation defines the *binomial transform*, which is self-inverted (involutory), i.e.:

$$a_p = \sum_{i=0}^{p} \binom{p}{i} (-1)^i b_i \tag{6.146}$$

Using the symbol $\mathcal{B}$ for the binomial transform we can write:

$$b_p = (\mathcal{B}a)_p \Leftrightarrow a_p = (\mathcal{B}b)_p \tag{6.147}$$

Likewise, if we take expected values on (6.144), and denote $c_i \coloneqq \mathrm{E}\left[\left(P\underline{x}+1-P\right)^i\right]$ we can write:

$$c_p = \sum_{i=0}^{p} \binom{p}{i} P^i (1-P)^{p-i} a_i \tag{6.148}$$

This latter equation defines the *Bernoulli transform with parameter P*. If we denote it with the symbol $\mathcal{B}^P$ we can write:

$$c_p = (\mathcal{B}^P a)_p \tag{6.149}$$

The relationship between the Bernoulli and binomial transforms is found as follows:

$$c_p = (1-P)^p \sum_{i=0}^{p} \binom{p}{i} \left(\frac{P}{1-P}\right)^i a_i = (1-P)^p (\mathcal{B}a')_p, \qquad a'_p \coloneqq \left(\frac{-P}{1-P}\right)^i a_p \tag{6.150}$$

Consequently:

$$c_p = (\mathcal{B}^P a)_p = (1-P)^p (\mathcal{B}a')_p \tag{6.151}$$

which can be written as

$$c'_p = (\mathcal{B}a')_p, \qquad c'_p \coloneqq \frac{1}{(1-P)^p} c_p \tag{6.152}$$

and by inverting the binomial transform

$$a'_p = (\mathcal{B}c')_p \Leftrightarrow a_p = \left(\frac{1}{p} - 1\right)^p (\mathcal{B}c')_p = ((\mathcal{B}^P)^{-1} c)_p \tag{6.153}$$

where $(\mathcal{B}^P)^{-1}$ is the inverse Bernoulli transform.

By setting $P = 1/2$, so that $a'_p \coloneqq (-1)^p a_p, c'_p \coloneqq 2^{-p} c_p$we find:`

$$(\mathcal{B}^P a)_p = 2^{-p} (\mathcal{B}a')_p \tag{6.154}$$

which shows that the binomial transform can be viewed as a special case of the Bernoulli transform.

Extending the result in equation (6.143), we multiply both sides by any function $g(\underline{x})$:

$$g(\underline{x})(1-\underline{x})^p = g(\underline{x}) \sum_{i=0}^{p} \binom{p}{i} (-1)^i \underline{x}^i \tag{6.155}$$

and take expected values to find:

$$b_p = \sum_{i=0}^{p} \binom{p}{i} (-1)^i a_i = (\mathcal{B}a)_p \tag{6.156}$$

where now:

$$a_p \coloneqq \mathrm{E}\left[g(\underline{x})\underline{x}^p\right], \qquad b_p \coloneqq \mathrm{E}\left[g(\underline{x})(1-\underline{x})^p\right], \tag{6.157}$$

Appendix 6-II: Relationships between different moment categories

For mean $\mu \neq 0$, the classical central and noncentral moments can, respectively, be written as:

$$\mu_p = \mathrm{E}\left[\left(\underline{x} - \mu\right)^p\right] = (-\mu)^p \mathrm{E}\left[\left(1 - \frac{\underline{x}}{\mu}\right)^p\right], \qquad \mu'_p = \mathrm{E}\left[\underline{x}^p\right] = \mu^p \mathrm{E}\left[\left(\frac{\underline{x}}{\mu}\right)^p\right] \tag{6.158}$$

Thus, the sequences $\mu_p/(-\mu)^p$ and $\mu'_p/\mu^p$are related to each other through the binomial transform (equation (6.143)), i.e.:

$$\frac{\mu_p}{(-\mu)^p} = \sum_{i=0}^{p} \binom{p}{i} (-1)^i \frac{\mu'_i}{\mu^i}, \qquad \frac{\mu'_p}{\mu^p} = \sum_{i=0}^{p} \binom{p}{i} (-1)^i \frac{\mu_i}{(-\mu)^i} \tag{6.159}$$

The latter simplify to:

$$\mu_p = \sum_{i=0}^{p} \binom{p}{i} (-\mu)^{p-i} \mu'_i, \qquad \mu'_p = \sum_{i=0}^{p} \binom{p}{i} \mu^{p-i} \mu_i \tag{6.160}$$

Similar relationships can be obtained between central and noncentral K-moments. In particular, from (6.7) we find

$$K_{pq} = (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q} (\underline{x}-\mu)^q\right] = (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q} \sum_{i=0}^{q} \binom{q}{i} (-\mu)^{q-i} \underline{x}^i\right]$$

$$= \sum_{i=0}^{q} \binom{q}{i} (-\mu)^{q-i} (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q} \underline{x}^i\right] \tag{6.161}$$

and finally

$$K_{pq} = \sum_{i=0}^{q} \binom{q}{i} (-\mu)^{q-i} K'_{p-q+i,i} \tag{6.162}$$

which can also be written in a binomial transform form as

$$\frac{K_{pq}}{(-\mu)^q} = \sum_{i=0}^{q} \binom{q}{i} (-1)^i \frac{K'_{p-q+i,i}}{\mu^i} \tag{6.163}$$

The inverse transform, after algebraic manipulations becomes

$$K'_{pq} = \sum_{i=0}^{q} \binom{q}{i} \mu^{q-i} K_{p-q+i,i} \tag{6.164}$$

In particular, for q = 1:

$$K_p = K'_p - \mu, \qquad K'_p = K_p + \mu \tag{6.165}$$

and for q = 2:

$$K_{p2} = K'_{p2} - 2\mu K'_{p-1,1} + \mu^2, \qquad K'_{p2} = K_{p2} + 2\mu K_{p-1,1} + \mu^2 \tag{6.166}$$

To find the relationship of noncentral to tail-based K-moments, we start from their definitions writing them in the form:

$$\frac{K'_{pq}}{p-q+1} = \mathrm{E}\left[\underline{x}^q \left(F(\underline{x})\right)^{p-q}\right], \qquad \frac{\overline{K}'_{pq}}{p-q+1} = \mathrm{E}\left[\underline{x}^q \left(1-F(\underline{x})\right)^{p-q}\right] \tag{6.167}$$

Setting $j = p - q$ we have:

$$\frac{K'_{q+j,q}}{j+1} = \mathrm{E}\left[\underline{x}^q \left(F(\underline{x})\right)^{j}\right], \qquad \frac{\overline{K}'_{q+j,q}}{j+1} = \mathrm{E}\left[\underline{x}^q \left(1-F(\underline{x})\right)^{j}\right] \tag{6.168}$$

which, combined with (6.156), indicate that the left-hand side parts are related through the binomial transform:

$$\frac{\overline{K}'_{q+j,q}}{j+1} = \sum_{i=0}^{j} \binom{j}{i} (-1)^i \frac{K'_{q+i,q}}{i+1} \tag{6.169}$$

Hence:

$$\overline{K}'_{q+j,q} = \sum_{i=0}^{j} \frac{j+1}{i+1} \binom{j}{i} (-1)^i K'_{q+i,q} = \sum_{i=0}^{j} \binom{j+1}{i+1} (-1)^i K'_{q+i,q} \tag{6.170}$$

which, after the necessary manipulations, yields (6.23) .

To find the relationship of central to hypercentral moments, we proceed as follows:

$$\begin{aligned} K^+_{pq} &= 2^{p-q}(p-q+1)\mathrm{E}\left[\left(F(\underline{x})-1/2\right)^{p-q}\left(\underline{x}-\mu\right)^q\right] \\ &= 2^{p-q}(p-q+1)\mathrm{E}\left[\sum_{i=0}^{p-q}\binom{p-q}{i}\left(-\frac{1}{2}\right)^i F(\underline{x})^{p-i-q}\left(\underline{x}-\mu\right)^q\right] = \\ &= 2^{p-q}(p-q+1)\sum_{i=0}^{p-q}\binom{p-q}{i}\left(-\frac{1}{2}\right)^i \mathrm{E}\left[F(\underline{x})^{p-i-q}\left(\underline{x}-\mu\right)^q\right] = \\ &= 2^{p-q}\sum_{i=0}^{p-q}\binom{p-q}{i}\left(-\frac{1}{2}\right)^i \frac{p-q+1}{p-q-i+1} K_{p-i,q} \end{aligned} \tag{6.171}$$

This results in

$$K^+_{pq} = 2^{p-q}\sum_{i=0}^{p-q}\left(-\frac{1}{2}\right)^i \binom{p-q+1}{i} K_{p-i,q}, \qquad \frac{K^+_{pq}}{2^{p-q}} = \sum_{i=0}^{p-q}(-1)^i \binom{p-q+1}{i}\frac{K_{p-i,q}}{2^i} \tag{6.172}$$

where the inverse is

$$K_{pq} = \left(\frac{1}{2}\right)^{p-q}\sum_{i=0}^{p-q}\binom{p-q+1}{i} K^+_{p-i,q}, \qquad 2^{p-q}K_{pq} = \sum_{i=0}^{p-q}\binom{p-q+1}{i} K^+_{p-i,q} \tag{6.173}$$

Appendix 6-III: Relationships between K-moments of continuous and mixed distributions

The tail-based K-moments $\overline{K}'^*_p$ of a distribution function with a discontinuity $1-P_1$ at the origin are related to those of the distribution without a discontinuity by:

$$\overline{K}'^*_p = P_1^p \overline{K}'_p = -\sum_{l=0}^{p}(-1)^l \binom{p}{l}(1-P_1)^l \overline{K}'_l \tag{6.174}$$

On the other hand, the noncentral K-moment of order p is:

$$\begin{aligned} K'^*_p &= -\sum_{i=1}^{p}(-1)^i \binom{p}{i}\overline{K}'^*_i = -\sum_{i=1}^{p}(-1)^i \binom{p}{i} P_1^i \overline{K}'_i = \sum_{i=1}^{p}(-1)^i \binom{p}{i} P_1^i \sum_{l=1}^{i}(-1)^l \binom{i}{l} K'_l \\ &= \sum_{l=1}^{p}(-1)^l K'_l \sum_{i=l}^{p}(-1)^i \binom{p}{i}\binom{i}{l} P_1^i \end{aligned} \tag{6.175}$$

Using the identity:

$$\binom{p}{n}\binom{n}{l} = \binom{p}{l}\binom{p-l}{n-l} \tag{6.176}$$

we find:

$$K_p'^* = \sum_{l=1}^{p}(-1)^l K_l' \sum_{i=l}^{p}(-1)^i \binom{p}{l}\binom{p-l}{i-l} P_1^i = \sum_{l=1}^{p}(-1)^l \binom{p}{l} K_l' \sum_{i=l}^{p}(-1)^i \binom{p-l}{i-l} P_1^i =$$
$$= \sum_{l=1}^{p}(-1)^l \binom{p}{l} K_l' \sum_{i=0}^{p-l}(-1)^{i+l} \binom{p-l}{i} P_1^{i+l} \tag{6.177}$$
$$= \sum_{l=1}^{p}(-1)^l (-1)^l \binom{p}{l} K_l' P_1^l \sum_{i=0}^{p-l}(-1)^i \binom{p-l}{i} P_1^i$$

and finally:

$$K_p'^* = \sum_{l=1}^{p}\binom{p}{l} K_l' P_1^l (1-P_1)^{p-l} \tag{6.178}$$

Regarding the approximation of equation (6.38), we define p_c so that it correspond to $p' = 2$. Thus, $K_{p_c}'^* = K_2'$. Approximating $K_p'^*$ for $p \le p_c$ with a power function as in the upper case of (6.38), we determine b as the logarithmic slope:

$$b = \frac{\ln(K_2'/P_1K_1')}{\ln p_c} = \frac{\ln(K_2'/K_1') - \ln P_1}{\ln p_c} = \frac{c\ln 2 - \ln P_1}{\ln p_c} = \frac{\ln(2^c/P_1)}{\ln p_c} \tag{6.179}$$

where $c \coloneqq \ln(K_2'/K_1')/\ln 2$. On the other hand, if we assume that $p' = P_1 p$, then at $p' = 2, 2 = P_1 p_c$ or $p_c = 2/P_1$. Making the algebraic operations we find (6.38).

Appendix 6-IV: Proof of estimator unbiasedness

The density function $f_{(i:n)}(x)$ of the ith order statistic, $\underline{x}_{(i:n)}$ is (Papoulis 1990):

$$f_{(i:n)}(x) = (n-i+1)\binom{n}{i-1}\big(F(x)\big)^{i-1}\big(1-F(x)\big)^{n-i} f(x) \tag{6.180}$$

and therefore:

$$\mathrm{E}\left[\underline{x}_{(i:n)}^q\right] = \int_{-\infty}^{\infty} x^q f_{(i:n)}(x)(x)\mathrm{d}x$$
$$= (n-i+1)\binom{n}{i-1}\int_{-\infty}^{\infty} x^q \big(F(x)\big)^{i-1}\big(1-F(x)\big)^{n-i} f(x)\mathrm{d}x \tag{6.181}$$
$$= (n-i+1)\binom{n}{i-1}\int_0^1 \big(x(F)\big)^q F^{i-1}(1-F)^{n-i}\mathrm{d}F$$

Further, the above expected value is related to the K-moments as follows:

$$\begin{aligned}
\mathrm{E}\left[\underline{x}_{(i:n)}^{q}\right] &= (n-i+1)\binom{n}{i-1}\int_0^1 \left(x(F)\right)^q F^{i-1}\sum_{j=0}^{n-i}\binom{n-i}{j}(-1)^j F^j \mathrm{d}F = \\
&= (n-i+1)\binom{n}{i-1}\sum_{j=0}^{n-i}\binom{n-i}{j}(-1)^j \int_0^1 \left(x(F)\right)^q F^{i+j-1+q-q}\mathrm{d}F = \\
&= i\binom{n}{i}\sum_{j=0}^{n-i}\binom{n-i}{j}(-1)^j \frac{K'_{i+j+q-1,q}}{i+j}
\end{aligned} \tag{6.182}$$

Now, we define the noncentral K-moment estimator:

$$\widehat{\underline{K}}'_{pq} = \sum_{i=1}^{n} b_{inpq}\, \underline{x}_{(i:n)}^{q} \tag{6.183}$$

and we seek coefficients b_{inpq} which make it unbiased. It is recalled that the ordering of the sample is meant in terms of $\underline{x}$ and not $\underline{x}^q$ ($\underline{x}_{(i:n)}^{q} \coloneqq \left(\underline{x}_{(i:n)}\right)^q$, which can be different from $\left(\underline{x}^q\right)_{(i:n)}$). This makes equations (6.181) and (6.183) consistent to each other.

The expectation of the estimator (6.183) is:

$$\begin{aligned}
\mathrm{E}\left[\widehat{\underline{K}}'_{pq}\right] &= \sum_{i=1}^{n} b_{inpq}\mathrm{E}\left[\underline{x}_{(i:n)}^{q}\right] \\
&= \sum_{i=1}^{n} b_{inpq}(n-i+1)\binom{n}{i-1}\int_0^1 \left(\left(x(F)\right)^q F^{p-q}\right) F^{i-1-p+q}(1-F)^{n-i}\mathrm{d}F \\
&= \int_0^1 \left(\left(x(F)\right)^q F^{p-q}\right)\sum_{i=1}^{n}(n-i+1)\binom{n}{i-1} b_{inpq}\, F^{i-1-p+q}(1-F)^{n-i}\mathrm{d}F
\end{aligned} \tag{6.184}$$

If we choose:

$$(n-i+1)\binom{n}{i-1} b_{inpq} = \begin{cases} 0 & i \le p-q \\ (p-q+1)\binom{n-p+q-1}{i-p+q-1} & i > p-q \end{cases} \tag{6.185}$$

then the sum in (6.184) is drastically simplified, i.e.,

$$\begin{aligned}
&\sum_{i=1}^{n} b_{inpq}\,(n-i+1)\binom{n}{i-1} F^{i-1-p+q}(1-F)^{n-i} \\
&\qquad = (p-q+1)\sum_{\substack{i=p-q+1 \\ n-p+q-1}}^{n}\binom{n-p+q-1}{i-p+q-1} F^{i-1-p+q}(1-F)^{n-i} \\
&\qquad = (p-q+1)\sum_{j=0}^{}\binom{n-p+q-1}{j} F^{j}(1-F)^{n-p+q-1-j} \\
&\qquad = (p-q+1)\big(F+(1-F)\big)^{n-p+q-1} = (p-q+1)(1)
\end{aligned} \tag{6.186}$$

and, consequently:

$$\mathrm{E}\left[\underline{\widehat{K}}'_{pq}\right] = (p-q+1)\int_0^1 \left(\left(x(F)\right)^q F^{p-q}\right)\mathrm{d}F = K'_{pq} \tag{6.187}$$

In other words, to have an unbiased estimator of a noncentral K-moment it suffices to choose:

$$b_{inpq} = \frac{p-q+1}{n-i+1}\frac{\binom{n-p+q-1}{i-p+q-1}}{\binom{n}{i-1}} = \frac{p-q+1}{n}\frac{(n-p+q-1)!\,(i-1)!}{(i-p+q-1)!\,(n-1)!} \tag{6.188}$$

and this can be generalized for non-integer $p > 0$, i.e.,

$$b_{inpq} = \begin{cases} 0, & i \le p-q \\ \dfrac{p-q+1}{n}\,\dfrac{\Gamma(n-p+q)}{\Gamma(n)}\,\dfrac{\Gamma(i)}{\Gamma(i-p+q)}, & i > p-q \end{cases} \tag{6.189}$$

where we can notice that only the last of the three multiplicative terms depends on *i*.

If we denote b_{inp} (without the last index q) the coefficient for $q = 1$, i.e.:

$$b_{inp} = \begin{cases} 0, & i < p \\ \dfrac{p}{n}\,\dfrac{\Gamma(n-p+1)}{\Gamma(n)}\,\dfrac{\Gamma(i)}{\Gamma(i-p+1)}, & i \ge p \end{cases} \tag{6.190}$$

then it can be readily verified that

$$b_{inpq} = b_{i,n,p-q+1} \tag{6.191}$$

On the other hand, combining (6.183) and (6.188) and taking expected values we find:

$$K'_{pq} = \sum_{i=p-q+1}^{n} \frac{p-q+1}{n}\,\frac{(n-p+q-1)!}{(n-1)!}\,\frac{(i-1)!}{(i-p+q-1)!}\,\mathrm{E}\left[\underline{x}^q_{(i:n)}\right] \tag{6.192}$$

Multiplying and dividing the right-hand side by $(p-q)!$ we get:

$$K'_{pq} = \sum_{i=p-q+1}^{n} \frac{(p-q+1)!\,(n-p+q-1)!}{n!}\,\frac{(i-1)!}{(p-q)!\,(i-p+q-1)!}\,\mathrm{E}\left[\underline{x}^q_{(i:n)}\right] \tag{6.193}$$

which can be written as

$$\binom{n}{p-q+1} K'_{pq} = \sum_{i=p-q+1}^{n} \binom{i-1}{p-q}\,\mathrm{E}\left[\underline{x}^q_{(i:n)}\right] \tag{6.194}$$

Appendix 6-V: Properties of the simplified K-moment estimators

For the case examined, the necessary unbiasedness condition of equation (6.68) takes the form:

$$\frac{p}{n}\sum_{i=1}^{n}\left(\widehat{F}\left(x_{(i:n)}\right)\right)^{p-1} = 1 \tag{6.195}$$

As already stated in section 6.10, the estimator (6.80) precisely fulfils this condition for $p \le 4$, i.e. the difference:

$$\Theta(n,p) \coloneqq \frac{p}{n}\sum_{i=1}^{n}\left(\frac{2i-1-n+\sqrt{n^2-1}}{2\sqrt{n^2-1}}\right)^{p-1} - 1 \tag{6.196}$$

is precisely zero. For higher p it is close, but not precisely equal, to zero. Specifically, it is smaller than 0.05%, for moment order as high as $p = n/10$. Beyond $n/10$ and up to $n/2$, it reaches 1%. It is not difficult to evaluate $\Theta(n,p)$ from equation (6.196) and then divide the K-moment estimate by $\Theta(n,p)+1$ to counter the deviation. Furthermore, a very accurate numerical approximation of $\Theta(n,p)$ is:

$$\Theta(n,p) = \begin{cases} 0, & p \le 4 \\ \approx -\frac{1}{24}\left(\frac{p-3.5}{n}\right)^2, & 4 < p \le \frac{n}{2} \end{cases} \tag{6.197}$$

Here we stress that (6.196) represents a necessary but not sufficient condition for unbiasedness. For that reason, the estimator (6.80) should not be applied for $p > n/2$ as, even after correcting by dividing by $\Theta(n,p)+1$, the bias is present.

In contrast, the approximate estimator (6.82), which does not take into account the smallest sample values (i.e., the $x_{(i:n)}$ values for $i < p$) can be used for any p. In this case the bias $\Theta(n,p)$ takes the form:

$$\Theta(n,p) \coloneqq \frac{p}{n}\sum_{i=p}^{n}\left(\frac{4i(n-4)-2n+7}{4n(n-4)-2}\right)^{p-1} - 1 \tag{6.198}$$

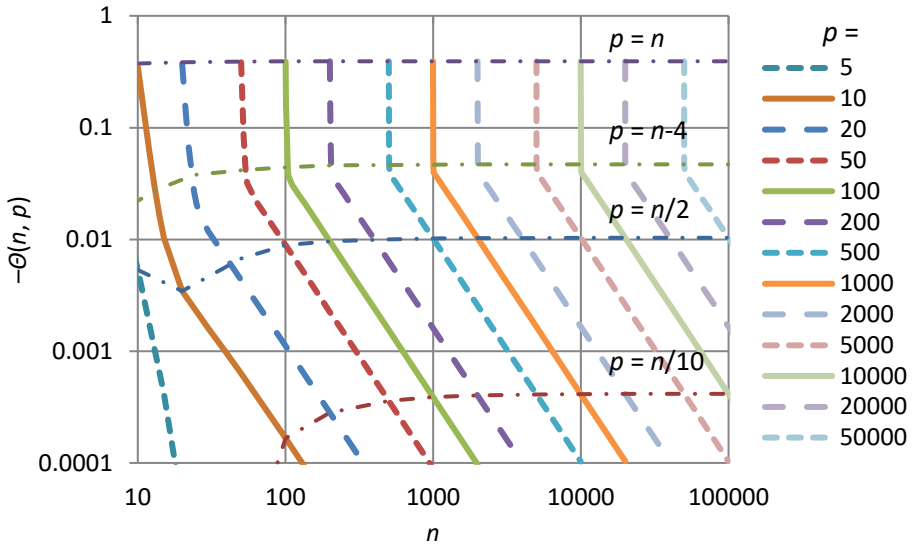

Figure 6.18 Deviation $-\Theta(n,p)$ in fulfilling the necessary unbiasedness condition of the estimator (6.82), as a function of sample size n and K-moment order p. Dash-dot lines correspond to the specified values of p as fractions of n.

This factor is estimated either by direct application of (6.198), or by its theoretical evaluation though the generalized Riemann zeta function ζ, i.e.:

$$\Theta(n,p) = \frac{p\big(\zeta(1-p,p-a(n)) - \zeta(1-p,n+1-a(n))\big)}{n(n-b(n))^{p-1}} - 1 \tag{6.199}$$

where $a(n)$ and $b(n)$ are given by (6.81), or even by the following numerical approximation, which is close to accurate:

$$\Theta(n,p)=\begin{cases}0, & p=1\\ \approx -\dfrac{1}{24}\left(\dfrac{p-4}{n}\right)^2-\left(1+\dfrac{1}{24}-\dfrac{1}{\sqrt{\mathrm{e}}}\right)\mathrm{e}^{p-n}, & p>1\end{cases} \tag{6.200}$$

The bias for the estimator (6.82) is depicted in Figure 6.18. It can be seen there that the deviation is negligible (< 0.05%), for moment order up to $p = n/10$, very small (< 1%) up to $n/2$, small (< 5%) up to $p = n - 4$, and increases rapidly thereafter. It is not difficult to evaluate $\Theta(n,p)$ from equation (6.196) and then divide the K-moment estimate by $\Theta(n,p)+1$ to remove bias.

Appendix 6-VI: Derivation of equations for the effect of autocorrelation

It is more convenient to determine the central K-moment K_2^{d} and then the noncentral one, which will be $K_2'^{\mathrm{d}} = K_2^{\mathrm{d}} + \mu$. For K_2^{d} we may assume that the stochastic variables have been transformed to normal distribution with zero mean and unit variance. As our derivations are approximate, we neglect the effect of that transformation to the autocorrelation. Assuming that the correlation coefficient of the variables $\underline{x}_i, \underline{x}_j$ is r_{ij} and using known results for normal variables (Nadarajah and Kotz, 2008; see also Appendix 5-II), the probability density of $\underline{y}_{ij} \coloneqq \max(\underline{x}_i, \underline{x}_j)$ will be:

$$f_{y_{ij}}(y) = 2f(y)F\left(\frac{1-r_{ij}}{\sqrt{1-r_{ij}^2}}y\right) \tag{6.201}$$

The expectation of $\underline{y}$ is then easily evaluated to:

$$K_2^{\mathrm{d},ij} = \mathrm{E}\left[\max(\underline{x}_i, \underline{x}_j)\right] = \int_{-\infty}^{\infty} y f_{y_{ij}}(y)\mathrm{d}y = \sqrt{\frac{1-r_{ij}}{\pi}} \tag{6.202}$$

We assume that $\underline{x}_i, \underline{x}_j$ are two terms among those in the sequence $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n$, and that the process has autocorrelation function $r_{ij} = r(i-j)$. For $i > j$, there are $(n-1)$ ways to allocate i and j so that $\tau \coloneqq i - j = 1$, $(n-2)$ ways to allocate them so that $\tau \coloneqq i - j = 2$, etc. Thus, the average of all $K_2^{\mathrm{d},ij}$ will be:

$$K_2^{\mathrm{d}} = \sum_{\tau=1}^{n-1}\sqrt{\frac{1-r(\tau)}{\pi}}(n-\tau) \Big/ \sum_{\tau=1}^{n-1}(n-\tau) = \frac{2}{\sqrt{\pi}\,n(n-1)}\sum_{\tau=1}^{n-1}\sqrt{1-r(\tau)}\,(n-\tau) \tag{6.203}$$

As $r(\tau) < 1$, this can be approximated as:

$$K_2^{\mathrm{d}} \approx \frac{2}{\sqrt{\pi}\,n(n-1)}\sum_{\tau=1}^{n-1}(1-r(\tau)/2)\,(n-\tau) = \frac{1}{\sqrt{\pi}} - \frac{1}{\sqrt{\pi}\,n(n-1)}\sum_{\tau=1}^{n-1} r(\tau)\,(n-\tau) \tag{6.204}$$

For an independent process $r(\tau) = 0$ and $K_2 = 1/\sqrt{\pi}$. Thus, if we define the adjustment coefficient as:

$$\Theta^{\mathrm{d}}(n; r(\tau)) \coloneqq \frac{K_2^{\mathrm{d}} - K_2}{K_2} = \frac{K_2^{\mathrm{d}}}{K_2} - 1 \tag{6.205}$$

this will take the form of equation (6.91)

For a Markov process, in which $r(\tau) = r^{\tau}$ equation (6.91) yields:

$$\Theta^{\mathrm{M}}(n,2) \approx -\frac{2r}{(1-r)^2}\frac{1-r-r^n/n}{n-1} \tag{6.206}$$

which results in (6.92). For an HK process, where $r(\tau) \approx H(2H-1)\tau^{2H-2}$, (6.91) yields:

$$\Theta^{\mathrm{HK}}(n,H) \approx -\frac{H(1-2H)}{n(n-1)}\left(H_{n-1}^{(1-2H)} - nH_{n-1}^{(2-2H)}\right) \tag{6.207}$$

A rough approximation of the generalized harmonic number $H_n^{(a)}$ is (Lampret, 2015):

$$H_n^{(a)} = \begin{cases} 1 + \dfrac{n^{1-a}-1}{1-a}, & a \neq 0 \\ 1 + \ln n\,, & a = 0 \end{cases} \tag{6.208}$$

With this approximation, after the algebraic operations and approximations we find (6.93).

Since these results were obtained for the normal distribution, the following relationship holds approximately (see section 6.14, equation (6.112)):

$$\frac{1}{\overline{F}_N(K_p)} \approx \Lambda_\infty p + \Lambda_1 - \Lambda_\infty \tag{6.209}$$

where $\Lambda_1 = 2$ and $\Lambda_\infty = \mathrm{e}^{\gamma} = 1.781$, whilst γ is the Euler constant and the meaning of the Λ-coefficients are explained in section 6.14. For our approximation we will initially neglect the difference $\Lambda_1 - \Lambda_\infty$, thus introducing some error for small p, which we will revoke later. In this case:

$$\overline{F}_{\mathrm{N}}(K_p) \approx \frac{1}{\Lambda_\infty p}, \qquad \overline{F}_{\mathrm{N}}(K_{p'}) \approx \frac{1}{\Lambda_\infty p'} \tag{6.210}$$

where p' is such that:

$$K_{p'} = K_p^{\mathrm{d}} = (1+\Theta)K_p \tag{6.211}$$

where for convenience we have simplified the notation $\Theta^{\mathrm{HK}}(n,H)$ to Θ. From (6.211) and (6.210), solving for p', we find

$$\frac{1}{p'} \approx \Lambda_\infty \overline{F}_{\mathrm{N}}\left((1+\Theta)\overline{F}_{\mathrm{N}}^{-1}\left(\frac{1}{\Lambda_\infty p}\right)\right) \tag{6.212}$$

Now we use the approximation of the normal distribution function derived in Appendix 5-II (equations (5.44) and (5.45) for the distribution function and the quantile function, respectively) and find:

$$\frac{1}{p'} \approx \frac{\Lambda_\infty}{2}\exp\left(\frac{1+\Theta}{2}\left(\Theta\left(\sqrt{4\ln\left(\frac{\Lambda_\infty p}{2}\right)+1}-1\right)\right) - 2(1+\Theta)\ln\left(\frac{\Lambda_\infty p}{2}\right)\right)\right) =: C(p) \tag{6.213}$$

Calculating the log-log derivative of $C(p)$ we find

$$C^{\#}(p) = (1+\Theta)\left(\frac{\Theta}{\sqrt{4\ln(\Lambda_\infty p/2)+1}} - 1 - \Theta\right) \tag{6.214}$$

For $p \to \infty$, $C^{\#}(p) \to -(1+\Theta)^2$, which does not depend on Λ_∞. This allows simplifying the approximation (6.213) as:

$$p' \approx Ap^{((1+\Theta)^2)} + 1 - A \tag{6.215}$$

for some constant A, where the term $1 - A$ in the end was added so as to give $p' = 1$ for $p = 1$, thus recovering from the error introduced by neglecting the difference $\Lambda_1 - \Lambda_\infty$. By numerical investigation it was found that the constant $A = 1 - 2\Theta$ makes the approximation satisfactory, thus resulting in equation (6.95).

Appendix 6-VII: Derivation of limiting Λ factors

With reference to section 2.19 on the relationship of parent and extreme value distribution, combining equations (2.117) and (2.121), for sufficiently large threshold u we find that for a distribution function that belongs to the domain of attraction of the Extreme Value Type I distribution, the following approximation holds for $x \geq u$:

$$F(x) = F(u) + F\big(x|\underline{x} > u\big)\big(1 - F(u)\big) = F(u) + \left(1 - \exp\left(\frac{x-u}{\lambda}\right) - F(u)\right)\big(1 - F(u)\big) \tag{6.216}$$

Inverting F we find that

$$x(F) = u - \lambda \ln\left(\frac{1-F}{1-F_u}\right), \qquad F \geq F_u \coloneqq F(u) \tag{6.217}$$

while for $F < F_u$ the quantile function is unknown, say $x_{\mathrm{U}}(F)$. From equation (6.19) we find

$$K'_p = p\int_0^1 x(F)\,F^{p-1}\mathrm{d}F = p\int_0^{F_u} x(F)\,F^{p-1}\mathrm{d}F + p\int_{F_u}^1 x(F)\,F^{p-1}\mathrm{d}F = A + B + C \tag{6.218}$$

where

$$A \coloneqq p\int_0^{F_u} x(F)\,F^{p-1}\mathrm{d}F, \qquad B \coloneqq -p\int_0^{F_u} x_{\mathrm{U}}(F)\,F^{p-1}\mathrm{d}F, \qquad C \coloneqq p\int_0^1 x(F)\,F^{p-1}\mathrm{d}F \tag{6.219}$$

As $x(F)$ is a non-decreasing function with lower limit $x(0) = u + \lambda\ln(1 - F_u)$ and upper limit $x(F_u) = u$, we have for the term A:

$$p\int_0^{F_u} (u + \lambda\ln(1 - F_u))\,F^{p-1}\mathrm{d}F \leq A \leq p\int_0^{F_u} u\,F^{p-1}\mathrm{d}F \tag{6.220}$$

or

$$(u + \lambda\ln(1 - F_u))F_u^p \leq A \leq uF_u^p \tag{6.221}$$

The unknown $x_{\mathrm{U}}(F)$ should also be a non-decreasing function with lower limit, say, c (assumed finite as happens in hydrometeorological variables, e.g. c = 0) and upper limit $x_{\mathrm{U}}(F_u) = x(F_u) = u$. Thus, we have for the term B:

$$-p\int_0^{F_u} u\,F^{p-1}\mathrm{d}F \leq B \leq -p\int_0^{F_u} c\,F^{p-1}\mathrm{d}F \tag{6.222}$$

or

$$-uF_u^p \leq B \leq -cF_u^p \tag{6.223}$$

The term C evaluates to

$$C = p\int_0^1 \left(u - \lambda \ln\left(\frac{1-F}{1-F_u}\right)\right) F^{p-1}\mathrm{d}F = u + \lambda\left(H_p + \ln(1-F_u\right) \tag{6.224}$$

Now, combining all above we find:

$$C + (\lambda \ln(1-F_u))F_u^p \le K_p' \le C + (u-c)F_u^p \tag{6.225}$$

or

$$1 + \frac{(\lambda \ln(1-F_u))F_u^p}{C} \le \frac{K_p'}{C} \le 1 + \frac{(u-c)F_u^p}{C} \tag{6.226}$$

As $p \to \infty$, clearly $C \to \infty$, while both the lower and upper limit in the above inequality tend to 1 (notice that $F_u < 1$ and thus $F_u^p \to 0$). Thus, as $p \to \infty$ the following approximation holds:

$$K_p' = u + \lambda\left(H_p + \ln(1-F_u)\right) \tag{6.227}$$

From (6.216) we find:

$$F\left(K_p'\right) = 1 - \exp\left(-H_p\right) \tag{6.228}$$

and from (6.108) we obtain:

$$\Lambda_p = \frac{\exp\left(H_p\right)}{p} \tag{6.229}$$

The last relationship holds true precisely for the exponential distribution for any p and at the limit as $p \to \infty$ for any distribution belonging to the domain of attraction of the Extreme Value Type I distribution. This limit is evaluated to:

$$\Lambda_\infty = \mathrm{e}^{\gamma} \tag{6.230}$$

Appendix 6-VIII: Explanations for the approximation of Λ-coefficients

Equation (6.111), i.e.:

$$\Lambda_p = \Lambda_\infty + (\Lambda_1 - \Lambda_\infty)\frac{1}{p} \tag{6.231}$$

provides a first approximation of Λ_p for any distribution but can be improved. While it captures the initial and final values, Λ_1 and Λ_∞, it does not reflect the rate at which Λ_p tends to Λ_∞, which differs in different distributions. We quantify this rate through the difference

$$\Delta\Lambda_p \coloneqq \Lambda_p - \Lambda_\infty \tag{6.232}$$

If we approximate Λ_p with equation (6.113), the same difference will be given by:

$$\Delta\Lambda_p^{\mathrm{A}} \coloneqq \frac{A}{p} - B\ln\left(\beta'\left(1 + \frac{\beta}{(p+1)^\beta - 1}\right)\right), \qquad \beta' \coloneqq \begin{cases} 1, & \beta \ge 0 \\ \dfrac{1}{1-\beta}, & \beta \le 0 \end{cases} \tag{6.233}$$

where the superscript 'A' stands for "approximation". It can be easily demonstrated that for any $\beta \in \mathbb{R}$:

$$\Delta\Lambda_p^{\mathrm{A}} \underset{p\to\infty}{\longrightarrow} 0$$

The rate at which $\Delta\Lambda_p^{\mathrm{A}}$ tends to zero is described by the following asymptotic properties (given here without proof):

$$\Delta\Lambda_p^{\mathrm{A}} \underset{p\to\infty}{\longrightarrow} \begin{cases} -\dfrac{B}{\ln p}, & \beta = 0 \\ -\dfrac{\beta B}{p^{|\beta|}}, & 0 < |\beta| < 1 \\ \dfrac{A-\beta' B}{p}, & |\beta| = 1 \\ \dfrac{A}{p}, & |\beta| > 1 \end{cases} \tag{6.234}$$

Furthermore, we easily find for $p = 1$ that:

$$\Delta\Lambda_1^{\mathrm{A}} = \begin{cases} A - B\ln\left(1+\dfrac{1}{\ln 2}\right), & \beta = 0 \\ A - B\ln(1+\beta'), & |\beta| = 1 \\ A - B\ln\left(\beta'\left(1+\dfrac{\beta}{2^{\beta}-1}\right)\right), & \text{otherwise} \end{cases} \tag{6.235}$$

Now, for each particular distribution function we should find first the parameter β from the asymptotic properties of the function and then match $\Delta\Lambda_p^{\mathrm{A}}$ and $\Delta\Lambda_p$ for $p = 1$ and $p \to \infty$ utilizing the above two equations. A systematic study of several distributions, using both theoretical and numerical analyses, gave the results listed in Table 6.6.

Appendix 6-IX: The invariance of Λ-coefficients under linear transformations

Considering the linear transformation $\underline{w} = b\underline{x} + c$, the distribution of $\underline{w}$ is $F_w(w) = F((w-c)/b)$ and the density is $f_w(w) = f((w-c)/b)/w$. Thus,

$$\begin{aligned} K'_{w_{pq}} &= (p-q+1)\mathrm{E}\left[\left(F_w(\underline{w})\right)^{p-q}\underline{w}^q\right] = (p-q+1)\mathrm{E}\left[\left(F\left((\underline{w}-c)/b\right)\right)^{p-q}\underline{w}^q\right] \\ &= (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}(b\underline{x}+c)^q\right] \\ &= (p-q+1)b^q\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}(\underline{x}+c/b)^q\right] \end{aligned} \tag{6.236}$$

This means that $K'_{w_{pq}}$ depends on b and on the ratio c/b, and its determination requires the expansion of $(\underline{x}+c/b)^q$ through the binomial coefficients $\binom{q}{i}, i = 0, \ldots, q$. Thus, the result will be an involved expression that will contain a weighted sum of K-moments of $\underline{x}$, K'_{pi} for $i = 0, \ldots, q$. Note that the case $q = 1$ is an exception as

$$K'_{w_p} = pb\mathrm{E}\left[\left(F(\underline{x})\right)^{p-1}(\underline{x}+c/b)\right] = pb\mathrm{E}\left[\left(F(\underline{x})\right)^{p-1}\underline{x}\right] + pc\mathrm{E}\left[\left(F(\underline{x})\right)^{p-1}\right] = bK'_{pq} + c \tag{6.237}$$

The situation is much simpler for the central moments of

$$\begin{aligned} K_{w_{pq}} &= (p-q+1)\mathrm{E}\left[\left(F_w(\underline{w})\right)^{p-q}(\underline{w}-\mu_w)^q\right] \\ &= (p-q+1)\mathrm{E}\left[\left(F((\underline{w}-c)/b)\right)^{p-q}(\underline{w}-\mu_w)^q\right] \\ &= (p-q+1)\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}\left((b\underline{x}+c)-(b\mu+c)\right)^q\right] \\ &= (p-q+1)b^q\mathrm{E}\left[\left(F(\underline{x})\right)^{p-q}(\underline{x}-\mu)^q\right] \end{aligned} \tag{6.238}$$

and finally,

$$K_{w_{pq}} = b^q K_{pq} \tag{6.239}$$

Furthermore,

$$F_w\left(\left(K_{w_{pq}}\right)^{1/q}+\mu_w\right) = F\left(\left(\left(K_{w_{pq}}\right)^{1/q}+\mu_w-c\right)/b\right) = F\left(K_{pq}^{1/q}+\mu\right) \tag{6.240}$$

Hence,

$$\Lambda_{w_{pq}} = \Lambda_{pq} \tag{6.241}$$

which proves equation (6.130).

Appendix 6-X: Approximation K-moments of a normal variable

For a probability density that is an even function we have $f(x) = f(-x)$, $\bar{F}(x) \coloneqq 1 - F(x) = F(-x)$, $x(F) = -x(1-F)$. Hence, using (6.19), we find:

$$\begin{aligned} K'_{pq} = K_{pq} &= (p-q+1)\left(\int_0^{1/2} x(F)^q F^{p-q}\mathrm{d}F + \int_{1/2}^{1}\left(x(F)\right)^q F^{p-q}\mathrm{d}F\right) \\ &= (p-q+1)\left(\int_0^{1/2} x(F)^q F^{p-q}\mathrm{d}F + \int_0^{1/2}\left(x(1-A)\right)^q (1-A)^{p-q}\mathrm{d}A\right) \\ &= (p-q+1)\left(\int_0^{1/2} x(F)^q F^{p-q}\mathrm{d}F + \int_0^{1/2}(-1)^q\left(x(F)\right)^q (1-F)^{p-q}\mathrm{d}F\right) \\ &= (p-q+1)\left(\int_0^{1/2} x(F)^q\left(F^{p-q}+(-1)^q(1-F)^{p-q}\right)\mathrm{d}F\right) \end{aligned} \tag{6.242}$$

For $q = 1$ this becomes

$$K'_p = K_p = p\left(\int_0^{1/2} x(F)\left(F^{p-1}-(1-F)^{p-1}\right)\mathrm{d}F\right) \tag{6.243}$$

Let

$$B_p \coloneqq p\left(\int_0^{1/2} x(F)\,F^{p-1}\,\mathrm{d}F\right) \tag{6.244}$$

Then

$$p\left(\int_0^{1/2} x(F)\,(1-F)^{p-1}\,\mathrm{d}F\right) = p\sum_{k=0}^{p-1}\binom{p-1}{k}(-1)^k\left(\int_0^{1/2} x(F)\,F^k\,\mathrm{d}F\right)$$
$$= \sum_{k=0}^{p-1}\frac{p}{k+1}\frac{(p-1)!}{k!\,(p-k-1)!}(-1)^k B_{k+1} = -\sum_{k=1}^{p}\frac{p!}{(k)!\,(p-k)!}(-1)^k B_k \tag{6.245}$$

and hence

$$K'_p = B_p + \sum_{k=1}^{p}\binom{p}{k}(-1)^k B_k \tag{6.246}$$

For the normal distribution B_p can be approximated using (5.45), i.e.,

$$B_p \approx p\left(\int_0^{1/2} -\frac{3}{4}\left(\sqrt{1-4\ln(2F)}-1\right)F^{p-1}\,\mathrm{d}F\right) = -\frac{3\sqrt{\pi}\mathrm{e}^{p/4}\,\mathrm{erfc}\left(\sqrt{p}/2\right)}{2^{p+2}} \tag{6.247}$$

To find the Λ-coefficients we use the approximation (5.44) noting that $K'_p > 0$ for any $p > 0$:

$$\Lambda_p \approx \frac{2}{p\exp\left(-\frac{2}{3}K'_p\left(1+\frac{2}{3}K'_{p1}\right)\right)} \tag{6.248}$$

Chapter 7. Stochastic simulation of hydroclimatic processes

7.1 Desiderata of a simulation scheme

In several instances in the previous chapters we had to deal with problems that do not admit an analytical solution. A most promising alternative for such problems is the stochastic (or Monte Carlo) simulation, which has been introduced in section 2.6. If the processes we had to deal with could be effectively modelled as white noise, then the simple random number generators presented in section 2.6 would be enough for our simulations. However, hydroclimatic processes are characterized by several behaviours which we need to respect and reproduce in our simulations, both qualitatively and quantitatively. While here we avoid to provide all details about these behaviours and to review the variety of methods devised to deal with them, we provide a rather simple generic scheme that can used in most problems related to extremes of hydroclimatic processes. In Digression 7.A we also discuss non-conventional types of stochastic simulation, by conversion of deterministic models into stochastic.

Before we discuss simulation schemes per se it is useful to introduce the characteristic behaviours.

Periodicity. When the time scale of interest is finer than annual, hydroclimatic processes exhibit seasonality, related to the annual motion of Earth around the Sun. In addition, when the time scale of interest is finer than daily, some of those processes may exhibit regular diurnal variation, related to the daily rotation of Earth. The most appropriate technique to deal with these regular variations is to build a so-called *cyclostationary model*, with single or double periodicity, depending on intensity of the periodic variation and its effect for the very problem of interest. In a cyclostationary model the parameters of the n*th order distribution function* vary according to periodic (apparently, deterministic) functions of time.

Here we will not discuss the rather sophisticated methods of this category, but we will resort to simpler methods in which only the first-order (marginal) distribution function of the process is dealt with. We list the following techniques of this category of approximate methods, from the most to the least complex.

- *Nonlinear transformation* of the process by "season" and/or "hour", where "season" is a part of the year (e.g. one or more months) in which the seasonal variation is no longer substantial and likewise for "hour" (which may mean one or more hours); the standard transformation of this type is a transformation making the distribution standard normal (*normalization*).
- *Linear transformation* or else *standardization* of the process, usually expressed as $\underline{y}_\tau = (\underline{x}_\tau - \mu_\tau)/\sigma_\tau$, where μ_τ and σ_τ are periodic function of the time τ; this is followed by modelling the process and recovering of $\underline{x}_\tau$ by applying the inverse transformation.
- *Proportional adjustment* (or *linear mapping*), which is similar to the linear transformation expect that there is no subtraction, i.e., $\underline{y}_\tau = \underline{x}_\tau/\alpha_\tau$, where α_τ is a

periodic function of the time τ; the advantages of this technique are its parsimony and avoidance of negative values in the case that the process is nonnegative (e.g. rainfall).

- *Null case* (or do-nothing) is an option when the periodicity entails a negligible effect on the problem we study; an example where the null case is applicable is discussed in Digression 6.B.

Dependence (and particularly **long-range dependence**). The omnipresence of dependence in natural processes is a sufficient reason to replace the classical IID statistics with stochastic processes. Short-range dependence has been the basis of using ARMA-type models, but these prove inadequate for many natural processes. Therefore, our simulation scheme should be able to reproduce long-range dependence. Dependence is typically handled through the second-order characteristics of a stochastic process, while long-range dependence is identified through the asymptotic LLDs of the second order characteristics (section 3.8). Among them, the climacogram and climacospectrum are most useful for the model identification and fitting phases, while for the simulation phase the autocovariance becomes also very useful. Preservation of any one of the second-order characteristics results in preservation of all other.

Intermittence. At fine time scales, hydroclimatic processes exhibit intermittent behaviour. This is most clear in the rainfall process, where it is quantified by the probability dry (probability of dry state). Similar is the situation with the streamflow in ephemeral streams. However, intermittence may appear in a less visible manner in the streamflow of large rivers with permanent flow, where the state switches between baseflow and flood. The baseflow is characterized by its own variability, and therefore a characterization by a single parameter, such as probability of the baseflow state, would be inefficient. A more general characterization of intermittence can be made in terms of high-order moments, starting from the skewness.

Skewness and high-order moments. At fine and intermediate time scales, most hydroclimatic processes have positively skewed distribution functions. The skewness is mainly caused by the fact that hydroclimatic variables are non-negative and sometimes intermittent. This is not so common in other scientific fields whose processes can safely be regarded as Gaussian. Thus, the preservation of skewness becomes important for hydroclimatic processes, while in combination with intermittence, proper modelling should include preservation of moments of order higher than 3. Unlike the second-order characteristics, where simulation schemes are able to preserve joint and marginal moments, for orders of ≥ 3 only marginal moments can be dealt with in an explicit manner.

Time irreversibility. In the streamflow process, time irreversibility (the asymmetry in time, manifested e.g. with rapid increases followed by gradual decreases) is evident up to time scales of several days, while in atmospheric processes irreversibility appears only at very fine scales (Koutsoyiannis, 2019b). Irreversibility can be quantified by the skewness of the time differenced process $\underline{\tilde{x}}_\tau := \underline{x}_\tau - \underline{x}_{\tau-1}$, which in turn has to be

preserved in simulations. This preservation is feasible only if $\underline{x}_\tau$ per se has a skewed distribution, while processes with symmetric distribution are also time symmetric.

Spatial variation and dependence. Hydroclimatic processes evolve both in time and space. Typically, simulation deals only with the temporal evolution. The most precise mathematical representation of hydroclimatic processes can be achieved extending the index set of the process from one dimension (representing time) to three dimensions (one for time and two for space). However, multidimensional modelling is not easy and has been implemented only in few cases. A midway solution, which is more common in applications, is to use multivariate models, which describe the temporal evolution of the process simultaneously at a number of points. Thus, instead of have a vector index set $\boldsymbol{\tau}$ in $\underline{x}_{\boldsymbol{\tau}}$, we vectorize the process state $\underline{\boldsymbol{x}}_\tau$ keeping τ scalar. This vectorization type can be directly used also to model more than one cross-correlated process (e.g. rainfall and runoff) at the same location simultaneously. In the remaining of the chapter we will deal only with scalar processes with scalar index set; the reader interested about multivariate or multidimensional processes is referred to Koutsoyiannis (2000) or Koutsoyiannis et al. (2011), respectively (see also Dimitriadis et al. 2019; Sargentis et al., 2020).

Digression 7.A: Non-conventional stochastic simulation incorporating deterministic models

Deterministic models have been widely used in hydroclimatic processes. In many cases their use has been very effective in providing reasonable predictions yet they suffer from the fact that they neglect uncertainty, which in inherent in such processes.

It is possible to convert a deterministic models into stochastic and perform stochastic simulation to assess the uncertainty. A relevant technique, sometimes called *an ensemble method*, is to shift from one to many applications of the deterministic model. Each simulation is performed after stochastically perturbing either input data, model parameters, model output, or all of them. In particular, perturbing the model error is done by adding random outcomes from the population of model errors, whose probability distribution is conditioned on input data and model parameters. Montanari and Koutsoyiannis (2012) have provided a blueprint of this approach which was further applied in a data-driven mode by Sikorska et al. (2017) and further advanced by Papacharalampous et al. (2020a,b).

In a different approach, deterministic model outputs can be converted to stochastic by connecting a (single-run) deterministic output to a stochastic model of the process using a Bayesian framework. Such an approach, accompanied with hydroclimatic applications, has been studied by Tyralis and Koutsoyiannis (2017).

7.2 Simple discrete-time processes of the Time Series School

The simplest of the processes of the Time Series School have been already described in section 3.11 and Digression 3.E, where it has been also explained why more complex models of that kind are not recommended. Instead of using complex time-series models, it is preferable to follow the general methodology summarized in the next sections. Nonetheless, when there is no persistence (or antipersistence) in the process of interest, the simple models, which are contained in Table 7.1 along with all equations needed for their application, are convenient and readily applicable to simulate a process $\underline{x}_\tau$ by filtering white noise $\underline{v}_\tau$ as indicated in the second column of Table 7.1.

Table 7.1 Equations of the simplest models of the Time Series School and their characteristics.

Name	Process equation	Equations for second-order characteristics	Equation for marginal moments of any order	Eqn. no.
AR(1)	$\underline{x}_\tau = a\underline{x}_{\tau-1} + \underline{v}_\tau$	$c_0(1-a^2) = \sigma_v^2$ $c_\eta = a^{\lvert\eta\rvert}c_0$	$(1-a^p)\mu_p = \mu_p^{(v)}$	(3.77)-(3.78)
AR(2)	$\underline{x}_\tau = a_1\underline{x}_{\tau-1} + a_2\underline{x}_{\tau-2} + \underline{v}_\tau$	$c_0 = a_1c_1 + a_2c_2 + \sigma_v^2$ $c_1 = a_1c_0 + a_2c_1$ $c_\eta = a_1c_{\eta-1} + a_2c_{\eta-2}, \eta \geq 1$		(3.79)-(3.80)
ARMA(1,1)	$\underline{x}_\tau = a\underline{x}_{\tau-1} + \underline{v}_\tau + b\underline{v}_{\tau-1}$	$c_0 = ac_1 + (1+ab+b^2)\sigma_v^2$ $c_1 = ac_0 + b\sigma_v^2$ $c_\eta = a^{\eta-1}c_1, \quad \eta \geq 1$	$(1-a^p)\mu_p = (1+b^p)\mu_p^{(v)}$	(3.71)-(3.72)

All models of Table 7.1 can reproduce the marginal mean and variance of the process, while the AR(1) and ARMA(1,1) can also reproduce marginal moments of higher order. In terms of characteristics of the joint distribution, the AR(1) model can reproduce the lag-one autocovariance, while the other two models can, additionally, reproduce the lag two autocovariance. The model parameters a and b can be determined by solving the equations of the third column of Table 7.1, while the high-order moments of the white noise process can be preserved by specifying the high order moments of $\underline{v}_\tau$, $\mu_p^{(v)}$, so as to satisfy the equations of the fourth column of Table 7.1.

7.3 Generic simulation method for any stochastic structure

To simulate the discrete-time stochastic process $\underline{x}_\tau$ with any autocovariance function c_η we can use the generalized moving average scheme (Koutsoyiannis 2000):

$$\underline{x}_\tau = \sum_{j=-J}^{J} a_j \underline{v}_{\tau-j} \tag{7.1}$$

where a_j are weights to be calculated from the autocovariance function, $\underline{v}_j$ is white noise averaged in discrete-time (and not necessarily Gaussian), also known as *innovation process*, and J is theoretically infinite, so that in all theoretical calculations we will assume $J = \infty$, while in the generation J is a large integer chosen so that the resulting truncation error be negligible. Here we stress that the above scheme is just the contrary to the schemes of the Time Series School. Specifically, (a) we use a purely moving average scheme without any autoregressive term and (b) we do not relate our scheme with observations, as the observations have been already used in the model fitting phase, which is totally isolated from the generation scheme.

Writing equation (7.1) for $\underline{x}_{\tau+\eta}$, multiplying it by (7.1) and taking expected values we find the convolution expression for $J = \infty$:

$$c_\eta = \sum_{l=-\infty}^{\infty} a_l a_{\eta+l} \tag{7.2}$$

We need to find the sequence of $a_\eta, \eta = \cdots, -1, 0, 1, \ldots$, so that (7.2) holds true. The following generic solution of the generating scheme, giving the coefficients a_η, has been proposed by Koutsoyiannis (2020a):

$$a_\eta = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi \mathrm{i}(\theta(\omega)-\eta\omega)} A^{\mathrm{R}}(\omega)\,\mathrm{d}\omega \tag{7.3}$$

where $\theta(\omega)$ is any (arbitrary) odd real function (meaning $\theta(-\omega) = -\theta(\omega)$) and

$$A^{\mathrm{R}}(\omega) \coloneqq \sqrt{2 s_{\mathrm{d}}(\omega)} \tag{7.4}$$

As proved in Koutsoyiannis (2020a) the sequence of a_η:

(1) consists of real numbers, despite the expression in (7.3) involving complex numbers;
(2) satisfies precisely equation (7.2); and
(3) is easy and fast to calculate using the fast Fourier transform (FFT).

This theoretical result is readily converted into a numerical algorithm, which consists of the following steps:

(a) From the continuous-time stochastic model, expressed through its climacogram $\gamma(k)$, we calculate its autocovariance function in discrete time (assuming time step D):

$$c_\eta = \frac{(\eta+1)^2\gamma(|\eta+1|D) + (\eta-1)^2\gamma\big((|\eta-1|D\big)}{2} - \eta^2\gamma(|\eta|D) \tag{7.5}$$

(This step is obviously omitted if the model is already expressed in discrete time through its autocovariance function.)

(b) We choose an appropriate number of coefficients J that is a power of 2 and perform inverse FFT (using common software) to calculate the discrete-time power spectrum and the frequency function $A^{\mathrm{R}}(\omega)$ for an array of $\omega_j = j\,w_1, j = 0,1,\dots,J, w_1 \coloneqq 1/JD$:

$$s_{\mathrm{d}}(\omega_j) = 2c_0 + 4\sum_{\eta=1}^{J} c_\eta \cos(2\pi\eta\omega_j), \qquad A^{\mathrm{R}}(\omega_j) = \sqrt{2 s_{\mathrm{d}}(\omega_j)} \tag{7.6}$$

(c) We choose $\theta(\omega)$ (see below) and we form the arrays (vectors) $\boldsymbol{A}^{\mathrm{R}}$ and $\boldsymbol{A}^{\mathrm{I}}$, both of size $2J$ indexed as $0,\dots,2J-1$, with the superscripts R and I standing for the real and imaginary part of a vector of complex numbers, respectively:

$$[\boldsymbol{A}^{\mathrm{R}}]_j = \begin{cases} A^{\mathrm{R}}(\omega_j)\cos\left(2\pi\theta(\omega_j)\right)/2, & j = 0,\dots,J \\ [\boldsymbol{A}^{\mathrm{R}}]_{2J-j}, & j = J+1,\dots,2J-1 \end{cases} \tag{7.7}$$

$$[\boldsymbol{A}^{\mathrm{I}}]_j = \begin{cases} -A^{\mathrm{R}}(\omega_j)\sin\left(2\pi\theta(\omega_j)\right)/2, & j = 0,\dots,J-1 \\ 0 & j = J \\ -[\boldsymbol{A}^{\mathrm{I}}]_{2J-j}. & j = J+1,\dots,2J-1 \end{cases} \tag{7.8}$$

(d) We perform FFT on the vector $\boldsymbol{A}^{\mathrm{R}} + \mathrm{i}\,\boldsymbol{A}^{\mathrm{I}}$ (using common software), and get the real part of the result for $j = 0,\dots,J$, which is precisely the sequence of a_η.

We note that by choosing J as a power of 2, the vectors $\boldsymbol{A}^{\mathrm{R}}$ and $\boldsymbol{A}^{\mathrm{I}}$ will have size $2J$ which is also a power of 2, thus achieving maximum speed in the FFT calculations. (More details are contained in a supplementary file in Koutsoyiannis, 2020a, which includes numerical examples along with the simple code needed to do these calculations on a spreadsheet). It may be useful to note the following additional points about the method:

- Equation (7.3) gives not a single solution, but a variety of infinitely many ones, all of which preserve exactly the second-order characteristics of the process.
- A particular solution is characterized by the chosen function $\theta(\omega)$.
- Even assuming $\theta(\omega) = \theta_0 \operatorname{sign} \omega$ with constant θ_0, again there are infinitely many solutions.
- The availability of infinitely many solutions enables preservation of additional statistics (e.g. those related to time asymmetry; see section 7.5).
- In addition, we always have several options related to the distribution of the white noise $\underline{v}_\tau$, which in general is not Gaussian, thus enabling preservation of moments of any order (see section 7.4).

The special case $\theta(\omega) = 0$ gives a symmetric solution with respect to positive and negative η:

$$A^{\mathrm{S}}(\omega) \equiv A^{\mathrm{R}}(\omega) = \sqrt{2 s_{\mathrm{d}}(\omega)}, \qquad a_\eta^{\mathrm{S}} = \int_0^{1/2} \sqrt{2 s_{\mathrm{d}}(\omega)} \cos(2\pi\eta\omega)\, \mathrm{d}\omega = a_{-\eta}^{\mathrm{S}} \tag{7.9}$$

where the superscript S stands for symmetric. This has been known as the *symmetric moving average* (SMA) scheme (Koutsoyiannis 2000). All other solutions denote *asymmetric moving average* (AMA) schemes. An interesting special AMA case is obtained for $\theta(\omega) = 1/4 \operatorname{sign} \omega$ (or $2\pi\theta(\omega) = \pi/2 \operatorname{sign} \omega$). This corresponds to an *antisymmetric AMA* scheme (ANTAMA) with:

$$A^{\mathrm{A}}(\omega) = A^{\mathrm{R}}(\omega)\delta(\omega) + \mathrm{i}A^{\mathrm{R}}(\omega), \qquad a_\eta^{\mathrm{A}} = \delta_0 + \frac{1}{2}\int_0^{1/2} \sqrt{2 s_{\mathrm{d}}(\omega)} \sin(2\pi\eta\omega)\, \mathrm{d}\omega \tag{7.10}$$

where the superscript A stands for antisymmetric, $\delta(\omega)$ is the Dirac delta function, and

$$\delta_0 \coloneqq \frac{\sqrt{2 s_{\mathrm{d}}(0)}}{2(2J+1)} \approx \sqrt{\frac{\gamma_J}{2J+1}} \tag{7.11}$$

with δ_0 approaching zero as J becomes large. Any other case of constant θ_0 (where $\theta(\omega) = \theta_0 \operatorname{sign} \omega$) can be expressed in terms of the above two limiting cases through:

$$a_\eta = \delta_0 + \left(a_\eta^{\mathrm{S}} - \delta_0\right)\cos(2\pi\theta_0) + \left(a_\eta^{\mathrm{A}} - \delta_0\right)\sin(2\pi\theta_0) \tag{7.12}$$

For example, the case $\theta_0 = 1/8$ (or $2\pi\theta_0 = \pi/4$) yields the interesting result:

$$a_\eta = \frac{\sqrt{2}}{2}\left(a_\eta^{\mathrm{S}} + a_\eta^{\mathrm{A}}\right) - \left(\sqrt{2} - 1\right)\delta_0 \tag{7.13}$$

A most common solution is the *ordinary backward AMA* (OBAMA) scheme in which $a_\eta = 0$ for any $\eta < 0$; this latter is typically formulated in a different manner and denoted as simply moving average—MA, but as here we study a richer family of schemes, we use the distinct acronym OBAMA. A constant θ_0 does not give a precise OBAMA and therefore a non-constant function $\theta(\omega)$ is needed in this case. A generic analytical solution of $\theta(\omega)$ that would give a precise OBAMA is not simple (this problem is known as factoring of the power spectrum; see Papoulis 1991, p. 402). However, solutions for simple special cases are not too difficult to find (e.g. for rational spectra; Papoulis 1991, p. 402-404; Koutsoyiannis (2020a) for the Markov process in continuous time, as well as for the ARMA(1,1) process, including its special cases AR(1) and MA(1)).

However approximate OBAMA solutions can be found rather easily. First, if for some θ_0 and for $\eta < 0$ it happens that $a_\eta \approx 0$, then it can be verified that:

$$a_\eta \approx \begin{cases} 0, & \eta < 0 \\ a_\eta^{\mathrm{S}} \cos(2\pi\theta_0) + (1 - \cos(2\pi\theta_0) - \sin(2\pi\theta_0))\delta_0, & \eta = 0 \\ \sqrt{2} a_\eta^{\mathrm{S}} + \left(2 - \sqrt{2}\right)\delta_0, & \eta > 0 \end{cases} \tag{7.14}$$

Such a sequence with almost zero coefficients for negative η, will be close to the OBAMA scheme. It is interesting to notice that in this approximate solution only a_0 depends on the constant θ_0, while for $\eta > 0$ the coefficients are approximately equal to those in the SMA, multiplied by $\sqrt{2}$.

For stochastic structures with LRD, the OBAMA scheme this approximation may not be satisfactory and a better approximation can be found by adopting a parametric expression for $\theta(\omega)$ and optimizing its parameters (see examples in Koutsoyiannis, 2020a).

The method is illustrated in Figure 7.1 using two example processes. The first is the Markov process, whose basic properties are shown in Table 3.5. The second is the FHK-C model defined in equation (3.87), which gives its climacogram, whilst all its other characteristics are evaluated through the equations listed in Table 3.3. Specifically, Figure 7.1 shows three special cases, SMA (equation (7.9)), ANTAMA (equation (7.10)) and OBAMA for the two processes. For the OBAMA case and the Markov process the solution plotted is exact, while for the FHK-C process the sequence of a_η is an OBAMA approximation. Some slight (rather invisible) deviations from zero are present in the left-bottom panel, which at a later step will be set to zero and the small resulting effect will be further handled as a truncation error (in the manner described by Koutsoyiannis, 2016) to obtain an exact OBAMA scheme.

All in all, this simple method of the AMA scheme renders ARMA models unnecessary, particularly because of the generic, analytical and fast solution it offers. Here it is important to stress that, while optimization of coefficients involved in the function $\theta(\omega)$ could sometimes be required, it is not necessary in general. Any odd real function $\theta(\omega)$, chosen arbitrarily, will give a_η that will satisfy equation (7.2) (apart from a truncation error) and thus can directly be used in generation. Even if the sequence of $\theta(\omega_j)$ is constructed at random (e.g. as a sequence of random numbers in the interval [0,1/4]),

again equation (7.2) will be satisfied and the resulting a_η can directly be used in generation.

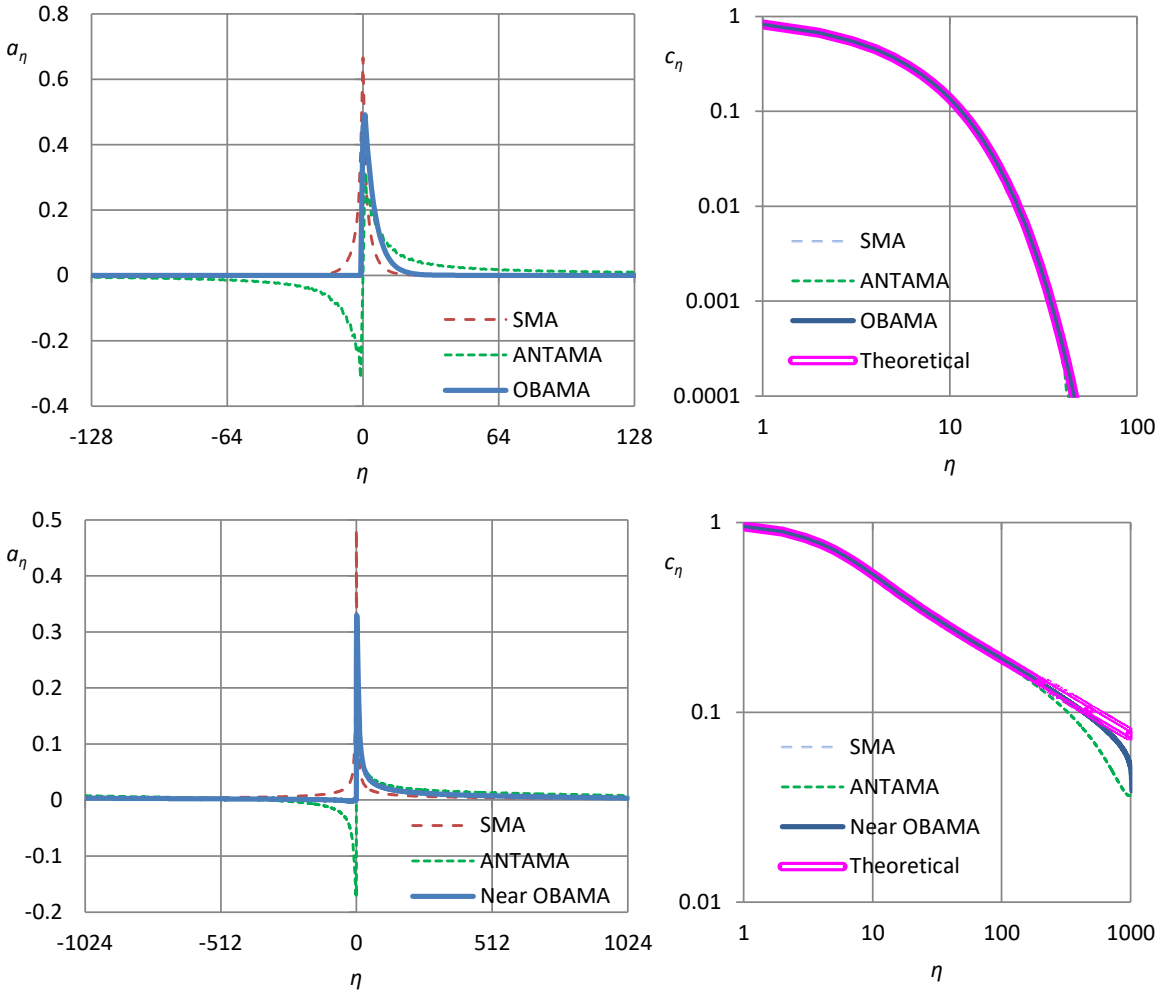

Figure 7.1 Illustration of the symmetric (SMA), antisymmetric (ANTAMA) and ordinary-backword (OBAMA) cases of the generic AMA model for **(upper row)** a Markov process and **(lower row)** an FHK-C process. The parameter values are α = 10, λ = 1 (in both processes), H = 0.8, M = 0.7, and the number of weights is 2049 (J = 1024 = 2^{10}); **(left column)** coefficients a; **(right column)** autocovariance function. (Source: Koutsoyiannis, 2020a.)

Digression 7.B: A simple analytical solution for the HK model

According to the algorithm presented in section 7.3, to calculate the series of coefficients a_η we need to perform the discrete Fourier transform (preferably in its FFT variant) twice, the first time to find the power spectrum of the process $s_{\mathrm{d}}(\omega)$ and the second time to determine a_η from the vector $\boldsymbol{A}^{\mathrm{R}} + \mathrm{i}\boldsymbol{A}^{\mathrm{I}}$. Generally, these transformations are performed numerically. However, the HK process allows analytical calculations. Specifically, there is an explicit analytical SMA solution (Koutsoyiannis, 2016):

$$a_l = \sqrt{b(H)\gamma_1}\left(\frac{|l+1|^{H+0.5} + |l-1|^{H+0.5}}{2} - |l|^{H+0.5}\right) \tag{7.15}$$

where $b(H)$ is a function of the Hurst coefficient H. For $H > 0.5$, the proximity of the power spectrum of the averaged process with that of the continuous-time process (equation (3.86)) allows the theoretical derivation of a consistent expression of $b(H)$, i.e. (Koutsoyiannis, 2016):

$$b(H) = \frac{2\Gamma(2H+1)\sin(\pi H)\,\gamma_1}{\Gamma^2(H+3/2)(1+\sin(\pi H))} \tag{7.16}$$

For $H < 0.5$ the proximity is not good and thus equation (7.16) does not perform well. However, a very good approximation, valid for any H, is Koutsoyiannis (2002, 2016):

$$b(H) \approx \frac{2(1-H)}{(3/2-H)^2 + 0.2(1/2-H)^2} \approx \frac{2(1-H)}{(3-2H)^2} \tag{7.17}$$

7.4 Preservation of high-order moments

The AMA and the SMA schemes allow preserving moments of any order by the method outlined below. In most applications, preservation of moments up to the fourth order gives adequate representation of hydroclimatic processes, as illustrated in Dimitriadis and Koutsoyiannis (2018) and Koutsoyiannis et al. (2018). It should be stressed that in typical sample sizes, high order moments should be evaluated theoretically through the distributional parameters (see Table 2.3) rather than estimated from the data, as their sample estimates are unreliable (Lombardo et al. 2014).

To more conveniently deal with moments of order > 2, we utilize the properties of cumulants of independent variables, and particularly homogeneity and additivity. The cumulants are directly determined from moments and vice versa (equation (2.36). For the pth cumulant, κ_p of $\underline{x}_\tau$, by virtue of (2.38) these properties result in:

$$\kappa_p = \sum_{l=-J}^{J} a_j^p \, \kappa_p^{(v)} \tag{7.18}$$

where $\kappa_p^{(v)}$ is pth cumulant of $\underline{v}_\tau$. Solving for $\kappa_p^{(v)}$ we find:

$$\kappa_p^{(v)} = \frac{\kappa_p}{\sum_{l=-J}^{J} a_j^p} \tag{7.19}$$

Based on the above discourse, we can formulate the following steps of a general simulation strategy, starting from the observed data (noting that alternative modelling strategies can be seen in a series of references provided by Dimitriadis and Koutsoyiannis, 2018, and Tsoukalas et al., 2018):

1. We estimate K-moments for q = 1 (and possibly 2), and we choose a marginal distribution for the process based on K-moments and possibly relevant theoretical considerations (e.g. entropy maximizing distributions).
2. We construct the climacogram and climacospectrum, and we choose a suitable model of second-order dependence.
3. We estimate the marginal and joint distribution parameters of the model (with appropriate provision for fitting issues, such as bias).
4. Based on the model parameters we calculate theoretically (and not estimate from data) the classical moments of the process of interest.
5. From equation (2.36) we calculate the cumulants of the process of interest.

6. From equation (7.19) we calculate the cumulants of the white noise process and from (2.36) we calculate its moments.
7. We choose an appropriate distribution for the white noise, calculate its parameters theoretically from its moments and generate a random sample with the required length.
8. Filtering with equation (7.1) we synthesize the simulated series for the process of interest.

In the case of a tail index $\xi > 0$, the moments and cumulants of $\underline{x}_\tau$ of order $\geq 1/\xi$ will be infinite, and hence those of $\underline{v}_\tau$ will be infinite, too. The moments involved in the modelling framework and the manner they are treated are summarized in Figure 7.2.

Moment order p	K-moment for $q = 1$	Classical moment	Cumulant	K-moment for $q = 2$
1	$K_1' \equiv \mu \equiv \kappa_1$			
2	$K_2', \quad K_2 = K_2' - \mu$	$\mu_2 \equiv \sigma^2 \equiv \kappa_2 \equiv K_{22}$		
3	$K_3', \quad K_3 = K_3' - \mu$	$\mu_3 \equiv \kappa_3$		K_{32}
4	$K_4', \quad K_4 = K_4' - \mu$	μ_4	$\kappa_4 \equiv \mu_4 - \mu_2^2$	K_{42}
⋮	⋮	⋮	⋮	⋮
i	$K_i', \quad K_i = K_i' - \mu$	μ_i	κ_i	K_{i2}
⋮	⋮			⋮
n (sample size)	$K_n', \quad K_n = K_n' - \mu$			K_{n2}

Legend

	Does not exist
	Can be estimated from data
	Estimation from data is possible but usually not necessary

	May exist but it is not necessary to derive
	Should be derived by theoretical calculations
	Estimation from data is possible for large sample but not recommended

Figure 7.2 Schematic of the moments involved in stochastic modelling and the manner they are treated. It is assumed that we wish to preserve classical moments up to order $i \ll n$.

7.5 Preservation of time irreversibility

Time irreversibility can very easily be handled within the AMA framework. Assuming that the white noise $\underline{v}_\tau$ (in discrete time τ) has variance 1 and coefficient of skewness $C_S^{(v)}$, we will have:

$$\mathrm{var}[\underline{x}_\tau] = \sum_{j=-J}^{J} a_j^2, \qquad \mu_3[\underline{x}_\tau] = \sum_{j=-J}^{J} a_j^3 \, C_S^{(v)} \tag{7.20}$$

where $\mu_3[\underline{x}_\tau]$ is the third moment of the process $\underline{x}_\tau$. Its coefficient of skewness will be:

$$C_S := \frac{\mu_3[\underline{x}_\tau]}{\left(\mathrm{var}[\underline{x}_\tau]\right)^{3/2}} = \frac{\sum_{j=-J}^{J} a_j^3}{\left(\sum_{j=-J}^{J} a_j^2\right)^{3/2}} C_S^{(v)} \tag{7.21}$$

Time asymmetry is quantified through the skewness of the differenced process $\underline{\tilde{x}}_\tau$, which by virtue of (7.1) is written as:

$$\underline{\tilde{x}}_\tau \coloneqq \underline{x}_\tau - \underline{x}_{\tau-1} = \sum_{j=-J}^{J} (a_j - a_{j-1})\underline{v}_{\tau-j} \tag{7.22}$$

with $a_{-J-1} = 0$. Thus, its skewness will be:

$$\tilde{C}_\mathrm{S} \coloneqq \frac{\mu_3[\underline{\tilde{x}}_\tau]}{\left(\mathrm{var}[\underline{\tilde{x}}_\tau]\right)^{3/2}} = \frac{\sum_{j=-J}^{J}(a_j - a_{j-1})^3}{\left(\sum_{j=-J}^{J}(a_j - a_{j-1})^2\right)^{3/2}} C_\mathrm{S}^{(v)} \tag{7.23}$$

The ratio:

$$\frac{\tilde{C}_\mathrm{S}}{C_\mathrm{S}} = \frac{\sum_{j=-J}^{J}(a_j - a_{j-1})^3}{\left(\sum_{j=-J}^{J}(a_j - a_{j-1})^2\right)^{3/2}} \frac{\left(\sum_{j=-J}^{J} a_j^2\right)^{3/2}}{\sum_{j=-J}^{J} a_j^3} \tag{7.24}$$

is independent of $C_\mathrm{S}^{(v)}$ and primarily depends on $\theta(\omega)$, which determines the sequence of a_η. The case $\theta(\omega) = 0$, i.e. the SMA, results in complete time symmetry. However, a constant $\theta_0 \neq 0$ (appropriately chosen) can make the ratio $\tilde{C}_\mathrm{S}/C_\mathrm{S}$ as high as we wish, thus enabling preservation of time asymmetry.

The above results make it clear that without skewness in the original process $\underline{x}_\tau$ (e.g. in the case of Gaussian processes), there cannot be time asymmetry.

Chapter 8. Rainfall extremes and ombrian modelling

8.1 From ombrian curves to ombrian models

One of the major tools in hydrological design is the *ombrian relationships*, more widely known by the misnomer rainfall *intensity-duration-frequency* (IDF) curves. An ombrian relationship (from the Greek 'όμβρος', rainfall) is a mathematical relationship connecting the time-averaged rainfall intensity x over a given time scale k (sometimes incorrectly referred to as duration) for a given return period T (also commonly referred to as frequency, although frequency is generally understood as reciprocal to period). Several forms of ombrian relationships are found in the literature, most of which have been empirically derived and validated by the long use in hydrological practice. Attempts to give them a theoretical basis have often used inappropriate assumptions and resulted in oversimplified relationships that are not good for engineering application.

Usually the ombrian curves are constructed for time scales of some minutes to several hours. This range of time scales has been dictated from engineering needs. However, with just a few adaptations of ombrian curves we can have a complete and decent stochastic model of rainfall, an *ombrian model*. The adaptations needed are basically two: an extension of the temporal coverage for large time scales and a more consistent theoretical formulation, in connection to the stochastic concepts we have already developed. From a practical point of view, there is no long way nor is a big effort required to move from the ombrian curves to an ombrian model. And once we have the model, we directly get the ombrian relationships ready for engineering application.

But do we really need a stochastic model? And if we do need, why not choose one of the many stochastic rainfall models of the literature? An easy reply to the second question is that of course we can choose any available model, but a two-in-one solution, a theoretically consistent model and a practical tool, both in one expression, is a better choice. Coming to the first question, our reply is positive for two reasons:

- While in traditional engineering design, the ombrian relationships are directly used in calculations, current hydrosystem configurations, which are increasingly complex, may require stochastic simulation, which is allowed by modern computational facilities. Stochastic simulation enables determination of risk at the end component of the hydrosystem, which actually is at risk, without relying on common simplifying assumptions, such as the equality of probability of occurrence of rainfall and flood discharge.
- As already discussed, estimation from data always involves bias and uncertainty, whose determination requires a model. Both bias and uncertainty become substantial when there is persistence. As we have already seen in several examples, this is the case with the rainfall process and, as we will see in this chapter, even the very common expressions of the relationship of rainfall and time scale suggest none other than persistence.

We will now discuss the basic postulates of an ombrian model, recalling that a model is always an approximation of reality and needs several assumptions to construct.

1. A basic desideratum is that the end result should be readily used in typical engineering tasks even without resorting to simulation. It should thus be as easy to use as traditional ombrian relationships. To this aim we could sacrifice perfect theoretical consistency, if this results in too involved expressions.
2. On the other hand, a basic requirement of any stochastic model is to handle and preserve first and second order characteristics of a process. As in ombrian relationships the variable of interest is the temporal average intensity $\underline{x}^{(k)}$ over any time scale k, it is natural to base our model on the climacogram, i.e. $\mathrm{var}\left[\underline{x}^{(k)}\right]$, recalling from Chapter 3 that by preserving the climacogram we preserve any other second order characteristic. The need for preserving a constant mean is self-evident, even though as we will see (Digression 8.B), no particular interest has been given in this requirement in common expressions of ombrian curves.
3. The process variance should be finite for $k \to 0$ (otherwise it will not be physically consistent; see section 2.17) and zero for $k \to \infty$ (otherwise the process will not be ergodic; see section 3.4).
4. The model should incorporate the fact that the probability dry, $P_0^{(k)} := P\{\underline{x}^{(k)} = 0\}$ is nonzero for small time scales. This means that the probability wet, $P_1^{(k)} := \overline{F}^{(k)}(0) = 1 - P_0^{(k)}$ is smaller than 1 for small k, even for $k \to 0$.
5. As the emphasis of an ombrian model is on maxima, moments of order higher than two are important to consider.
6. In particular, the tail index of the distribution for all scales should be constant for all time scales. Theoretical justification of this requirement is given in Appendix 8-I.
7. Because of its simplicity and explicit relationship between the time-averaged intensity and return period, the Pareto distribution is an optimal choice for small time scales; its suitability has been already verified in examples of previous sections.

In Digression 8.A and Digression 8.B we see that most of these requirements are violated in common ombrian relationships.

Digression 8.A: Inconsistencies of common ombrian relationships

The most common expression of ombrian curves (in particular in fractal-oriented studies) is:

$$x = \frac{\lambda T^{\xi}}{k^{\eta}} \tag{8.1}$$

where λ, ξ, η are parameters, all positive numbers and $\xi \leq \eta \leq 1$. Apparently, it is not dimensionally consistent but this can easily be remedied by introducing the parameters α and β with units of time and rewrite (8.1) in a dimensionally consistent manner, as:

$$x = \frac{\lambda'(T/\beta)^{\xi}}{(k/\alpha)^{\eta}} \tag{8.2}$$

where $\lambda' \coloneqq \lambda\beta^{\xi}/\alpha^{\eta}$.

According to equation (5.50), the return period $T = T^{(k)}$ is associated to time scale k and related to the latter by $T/k = 1/\overline{F}^{(k)}(x)$. Hence:

$$F^{(k)}(x) = 1 - \frac{\alpha}{\beta}\left(\frac{a}{k}\right)^{\eta/\xi - 1}\left(\frac{\lambda'}{x}\right)^{1/\xi} \tag{8.3}$$

This is not a proper probability distribution function as for $x = 0, F^{(k)}(x) = -\infty$. Also, for $k = 0, F^{(k)}(x) = -\infty$, irrespective of x, which again is an inconsistency.

Another ombrian relationship has been proposed by Koutsoyiannis et al. (1998) and refined in Koutsoyiannis (2007):

$$x = \lambda' \frac{(T/D)^{\xi} - \psi'}{(1 + k/\alpha)^{\eta}} \tag{8.4}$$

where D is a time unit, typically 1 year. At first glance this looks consistent with most of the requirements of section 8.1. In section 8.3 we will derive it in a slightly different form as a simplified ombrian model. As we will show, it is not free of inconsistencies, yet for small time scales is a good approximation of our consistent model, and can be useful in engineering application.

8.2 Building an ombrian model

To build a proper model in agreement with the postulates or, at least, without severe violations of the requirements set in section 8.1, we make the following assumptions:

1. Pareto distribution with discontinuity at the origin for small time scales:

$$F^{(k)}(x) = 1 - P_1^{(k)}\left(1 + \xi\frac{x}{\lambda(k)}\right)^{-1/\xi} \tag{8.5}$$

   As explained in section 8.1, the tail index ξ should be constant, while the probability wet, $P_1^{(k)}$, and the state scale parameter, $\lambda(k)$, are functions of the time scale k.

2. Continuous PBF distribution with discontinuity at zero for large time scales, i.e.:

$$F^{(k)}(x) = 1 - P_1^{(k)}\left(1 + \xi\left(\frac{x}{\lambda(k)}\right)^{\zeta(k)}\right)^{-1/\xi} \tag{8.6}$$

   In this case a new parameter $\zeta(k)$ is introduced, which is again a function of time scale. The Pareto distribution is a special case of (8.6) for $\zeta(k) = 1$. In contrast to the Pareto distribution, whose density is a consistently decreasing function of x, the PBF tends to be bell-shaped for increasing $\zeta(k)$, a property consistent with empirical observation and reason.

3. Constant mean of the time-averaged process:

$$\mathrm{E}\left[\underline{x}^{(k)}\right] = \mu \tag{8.7}$$

4. Climacogram of type FHK-C (equation (3.87)), i.e.:

$$\mathrm{var}\left[\underline{x}^{(k)}\right] = \gamma(k) = \lambda_1 \left(1 + \left(\frac{k}{\alpha}\right)^{2M}\right)^{\frac{H-1}{M}} \tag{8.8}$$

or of type FHK-CD (equation (3.90)). This has six parameters in total. To avoid an overparametrized model we set both time scale parameters equal and, as we expect $H > 0.5$ due to persistence and $M < 0.5$ due to roughness, we set $M = 1 - H$, thus getting:

$$\mathrm{var}\left[\underline{x}^{(k)}\right] = \gamma(k) = \lambda_1 \left(1 + \frac{k}{\alpha}\right)^{2H-2} + \lambda_2 \left(1 - \left(1 + \frac{\alpha}{k}\right)^{2H-2}\right) \tag{8.9}$$

Clearly, in both cases, $\gamma(k) \to 0$, as $k \to \infty$, which makes the process ergodic; for $k = 0$, $\gamma(0) = \gamma_0 = \lambda_1$ in the case of (8.8) and $\gamma(0) = \gamma_0 = \lambda_1 + \lambda_2$ in the case of (8.9). In both cases $\gamma(0)$ is finite and the number of parameters is four.

5. Probability wet and dry, $P_1^{(k)} = 1 - P_0^{(k)}$, varying with time scale according to:

$$\ln P_0^{(k)} = \ln P_0^{(k^*)} \, (k/k^*)^{\theta}, \qquad k \geq k^* \tag{8.10}$$

where k^* is the transition time scale from Pareto to PBF distribution, for which $P_0^{(k^*)} > 0$ and $\zeta(k^*) = 1$ (for continuity of the transition), and θ is a parameter ($0 \leq \theta \leq 1$). This equation has been derived in Koutsoyiannis (2006a) based on maximum entropy considerations.* As we will see, in the Pareto distribution, the probabilities dry and wet are derived directly from the distribution, and thus no equation additional to (8.10) is needed. The transition time scale k^* is chosen at a point where the deviation of probability dry derived from the Pareto model from the empirical one is marginally acceptable.

Both the decreasing (Pareto) and the bell-shaped (PBF) types of probability densities are consistent with natural behaviours for small and large time scales, respectively. It must be noted though that the tail index of the PBF distribution in the form of equation (8.6) is not ξ but $\xi/\zeta(k)$ and tends to zero as $k \to \infty$. Thus, with equation (8.6) we have sacrificed the requirement of a constant tail index, but this violation happens only for large time scales. The alternative to keep (8.6) and replace ξ with $\xi\zeta(k)$, thus recovering a constant tail index ξ, is not an option because it would result in a finite variance as $k \to \infty$ (with a coefficient of variation $\xi/\sqrt{1-2\xi}$), i.e., in a nonergodic process. There is also the alternative to replace (8.6) with the distribution of the sum of correlated Pareto variables. Actually, there is an analytical solution for this (Arendarczyk et al., 2018, albeit not for distribution with discontinuity at zero) but it is quite complicated and would severely violate the basic desideratum (point 1 of section 8.1). Therefore, we deem that

* A slight modification of equation (8.10), i.e., $\ln P_0^{(k)} = \ln P_0^{(0)} + \ln\left(P_0^{(k^*)}/P_0^{(0)}\right)(k/k^*)^{\theta}$, where $P_0^{(0)}$ is an additional parameter representing the probability dry of the instantaneous process, with value close to 1, can cover the entire range of scales. However, here the assumption of the Pareto distribution for small scales renders the additional parameter unnecessary.

the sacrifice of the constant tail index for the very large scales, which usually are not of interest in engineering practice, is unimportant.

What it remains to complete the model is to determine the functions $\lambda(k)$ and $\zeta(k)$ from the mean μ and the climacogram $\gamma(k)$. In Appendix 8-II we derive these functions as well as approximations thereof which are sufficiently good and much more practical in application:

$$\frac{1}{\zeta(k)} \approx \sqrt{(1-2\xi)\left(P_1^{(k)}\frac{\gamma(k)+\mu^2}{\mu^2}-1\right)} \tag{8.11}$$

$$\frac{1}{\lambda(k)} \approx \frac{P_1^{(k)}}{\mu}\left(1+\frac{1}{(1-\xi)\left(\zeta(k)\right)^2}-\frac{1}{\left(\zeta(k)\right)^{\sqrt{2}}}\right) \tag{8.12}$$

These correspond to the PBF distribution. In the Pareto case, $\zeta(k)=1$, and hence (8.11) can be used to derive the probability wet as:

$$P_1^{(k)} = \frac{1-\xi}{1/2-\xi}\frac{\mu^2}{\gamma(k)+\mu^2} \tag{8.13}$$

while (8.12) simplifies to:

$$\frac{1}{\lambda(k)} = \frac{P_1^{(k)}}{\mu(1-\xi)} = \frac{\mu}{(1/2-\xi)(\gamma(k)+\mu^2)} \tag{8.14}$$

Note that in the Pareto case, the equations are exact. The special case $P_1^{(k)}=1$ signifies the maximum time scale $k^*_{\max}$, at which the Pareto distribution is mathematically feasible, at which:

$$P_1^{(k^*_{\max})}=1, \qquad \frac{\gamma(k^*_{\max})}{\mu^2}=\frac{1}{1-2\xi}, \qquad \lambda(k^*_{\max})=\mu(1-\xi) \tag{8.15}$$

However, if we are interested in preserving the probabilities dry/wet according to equation (8.10), we should choose the time scale k^* (of transition from Pareto to PBF) smaller enough than $k^*_{\max}$.

The PBF distribution is feasible for any time scale, even when $P_1^{(k)}=1$ which actually is the case for large scales. In that case, equation (8.11) simplifies to:

$$\frac{1}{\zeta(k)} = \frac{\sqrt{(1-2\xi)\gamma(k)}}{\mu} \tag{8.16}$$

By setting $T=1/(1-F^{(k)}(x))$, the ombrian expression resulting from equation (8.6) is:

$$x = \lambda(k)\left(\frac{\left(P_1^{(k)}T/k\right)^{\xi}-1}{\xi}\right)^{\frac{1}{\zeta(k)}} \tag{8.17}$$

and for $\zeta(k) = 1$ (Pareto), it simplifies to:

$$x = \lambda(k)\frac{\left(P_1^{(k)}\,T/k\right)^{\xi} - 1}{\xi} \tag{8.18}$$

For $\xi = 0$ the PBF and Pareto expressions switch to Weibul and exponential, respectively, i.e.:

$$x = \lambda(k)\left(\ln\left(P_1^{(k)}\,T/k\right)\right)^{\frac{1}{\zeta(k)}}, \qquad x = \lambda(k)\ln\left(P_1^{(k)}\,T/k\right) \tag{8.19}$$

Recapitulating the above discourse, our ombrian model gives directly the ombrian curves in the form of (8.17) and its special case (8.18) for the Pareto distribution, applying on small scales, which are of greatest interest from an engineering point of view. These relationships rely on the mean μ, the climacogram $\gamma(k)$, the probability wet $P_1^{(k)}$ and the tail index of the distribution function of rainfall intensity. These relationhips are reproduced in Table 8.1.

Table 8.1 Mathematical relationships of the ombrian model. The ombrian curves per se are given in the last two rows.

Quantity	Small scales, $k \le k^*$ (Pareto)[1]	Large scales, $k \ge k^*$ (PBF)[1]
$\mathrm{E}\left[\underline{x}^{(k)}\right]$	μ	
$\gamma(k)$	$\lambda_1(1+(k/\alpha)^{2M})^{\frac{H-1}{M}}$ or $\lambda_1\left(1+\frac{k}{\alpha}\right)^{2H-2} + \lambda_2\left(1-\left(1+\frac{\alpha}{k}\right)^{2H-2}\right)$	
$P_1^{(k)}$	$\frac{1-\xi}{1/2-\xi}\frac{\mu^2}{\gamma(k)+\mu^2}$	$1-\left(1-P_1^{(k^*)}\right)^{(k/k^*)^{\theta}}$
$\frac{1}{\zeta(k)}$	1	$\sqrt{(1-2\xi)\left(P_1^{(k)}(\gamma(k)/\mu^2+1)-1\right)}$
$\frac{1}{\lambda(k)}$	$\frac{P_1^{(k)}}{\mu(1-\xi)}$	$\frac{P_1^{(k)}}{\mu}\left(1+\frac{1}{(1-\xi)\left(\zeta(k)\right)^2}-\frac{1}{\left(\zeta(k)\right)^{\sqrt{2}}}\right)$
x for $\xi > 0$	$\lambda(k)\frac{\left(P_1^{(k)}\,T/k\right)^{\xi}-1}{\xi}$	$\lambda(k)\left(\frac{\left(P_1^{(k)}\,T/k\right)^{\xi}-1}{\xi}\right)^{\frac{1}{\zeta(k)}}$
x for $\xi = 0$	$x = \lambda(k)\ln\left(P_1^{(k)}\,T/k\right)$	$x = \lambda(k)\left(\ln\left(P_1^{(k)}\,T/k\right)\right)^{\frac{1}{\zeta(k)}}$

[1] The transition time scale k^* is the time scale at which the empirical probability wet $P_1^{(k)}$ deviates from the expression given for the Pareto distribution; values are typically of the order of 10 to 100 h. Note that for $k \gg k^*$, the probability wet becomes 1; this simplifies the relationships for the PBF distribution for very large scales.

Our ombrian model offers:

(a) mathematical and physical consistency;
(b) coverage of all time scales, from zero to infinity;
(c) good behaviour on the very fine time scales, through the fractal parameter M;
(d) good behaviour on very large time scales, through the Hurst parameter H and the mean μ whose effect becomes important as time scale increases;
(e) simultaneous treatment and preservation of the climacogram; and
(f) simultaneous treatment and preservation of the probability dry/wet.

The ombrian model uses a total of seven parameters listed in Table 8.2. This number is greater than that in the conventional ombrian curves, which is typically five. If the data cannot support the estimation of seven parameters, this number can be reduced by using default values (e.g. $\theta = 1$, $M = 0.5$).

It is useful to note two inequality relationships among the parameters which would be useful in the model fitting phase. The first is implied by equation (8.13) and the fact that $P_1^{(k)} \leq 1$; as the domain of the Pareto distribution extends up to the transition time scale k^*, the following should hold.

$$\xi \leq \frac{1}{2} - \frac{\mu^2}{2\gamma(k^*)} \tag{8.20}$$

Furthermore, in order for (8.11) to be valid, the following inequality should hold for the entire domain of the PBF distribution, i.e. for any $k \geq k^*$:

$$P_1^{(k)} \left(\frac{\gamma(k)}{\mu^2} + 1 \right) \geq 1 \tag{8.21}$$

Table 8.2 Parameters of the ombrian model.

Parameter	Meaning of parameter	Related tool	Related equation
μ	Mean intensity	Mean, μ	(8.7)
λ_1, λ_2	Intensity scale parameters[1]	Climacogram, $\gamma(k)$	(8.8) or (8.9)
α	Time scale parameter	Climacogram, $\gamma(k)$	(8.8) or (8.9)
M	Fractal (smoothness) parameter[2]	Climacogram, $\gamma(k)$	(8.8)
H	Hurst parameter	Climacogram, $\gamma(k)$	(8.8) or (8.9)
θ	Exponent of the expression of probability dry	Probability wet, $P_1^{(k)}$	(8.10) [3]
ξ	Tail index	Probability distribution, $F(x)$	(8.5)-(8.6)

[1] One or two parameters for the cases that the climacogram is given by (8.8) or (8.9), respectively.
[2] The fractal (roughness/smoothness) parameter M is an independent parameter in the case that the climacogram is given by (8.8), while if it is given by (8.9) it is assumed $M = 1 - H$.
[3] The expression includes also the transition time scale k^* but this is not regarded a parameter but a modelling choice.

8.3 Model simplification for small time scales

The Pareto ombrian expression in equation (8.18), which is applicable only for small scales $k \leq k^*$, can be written as:

$$x = \lambda(k)\frac{(T/\beta(k))^{\xi} - 1}{\xi} \tag{8.22}$$

where $\beta(k)$ is a function of time scale with units of time, i.e.:

$$\beta(k) \coloneqq \frac{k}{P_1^{(k)}} \tag{8.23}$$

By virtue of (8.14) we will have:

$$x = \frac{(1/2 - \xi)(\gamma(k) + \mu^2)}{\xi\mu}\left(\left(\frac{T}{\beta(k)}\right)^{\xi} - 1\right) \tag{8.24}$$

Now, we make the simplifying assumption $P_1^{(k)} \propto k$, which can stand as an approximation for small k; hence:

$$\beta(k) = \beta = \text{constant} \tag{8.25}$$

Then (8.24) can then be written as:

$$x = \frac{(1/2 - \xi)(\gamma(k) + \mu^2)}{\xi\mu}\left(\left(\frac{T}{\beta}\right)^{\xi} - 1\right) \tag{8.26}$$

Further, by noting that for small time scales $\gamma(k) \gg \mu^2$, we can neglect the latter term in their sum. Assuming a climacogram in the form (8.8) and taking the neutral value $M = 1/2$ as default, we find:

$$x = \lambda_1\frac{(1/2 - \xi)}{\xi\mu}\left(1 + \frac{k}{\alpha}\right)^{2H-2}\left(\left(\frac{T}{\beta}\right)^{\xi} - 1\right) \tag{8.27}$$

We can now see that, thanks to the simplifying assumption (8.25), the rainfall intensity x is determined as the product of a function of time scale k and return period T. This facilitates calculations and particularly the parameter estimation. We can write this property in a more concise form as:

$$x = \lambda\frac{b(T)}{a(k)} \tag{8.28}$$

where we have made the product as quotient in order for both $a(k)$ and $b(T)$ to be increasing functions of their arguments. The function $a(k)$ is:

$$a(k) = \left(1 + \frac{k}{\alpha}\right)^{\eta}, \qquad \eta \coloneqq 2 - 2H \tag{8.29}$$

The parameter λ and the function $b(T)$ are for $\xi > 0$:

$$\lambda = \frac{(1/2 - \xi)\lambda_1}{\xi\mu}, \qquad b(T) = \left(\frac{T}{\beta}\right)^{\xi} - 1 \tag{8.30}$$

and for $\xi = 0$:

$$\lambda = \frac{\lambda_1}{2\mu}, \qquad b(T) = \ln\left(\frac{T}{\beta}\right) \tag{8.31}$$

This simplified ombrian relationship, constituted of equations (8.28)-(8.31), has five parameters in total, falling in three categories, namely: (a) λ with units same as x (typically mm/h); (b) α and β with units of time (typically in h, even though it may be convenient to express β in years); and (c) the dimensionless ξ and η ($0 < \xi < 0.5, 0 < \eta < 1$). Comparison of the parameters of the ombrian model and the simplified ombrian relationship is provided in Table 8.3. Interestingly, the parameter η is related, through equation (8.29), to the Hurst parameter, which is $H = 1 - \eta/2$. Clearly, any value of $\eta < 1$ results in $H > 0.5$, i.e., a process with persistence. Only the case $\eta = 1$ results in $H = 0.5$, but empirical evidence does not support the value $\eta = 1$. For typical values of η = 0.5-0.7, the resulting Hurst parameter H is 0.75-0.65. However, this is not a proper way to estimate the Hurst parameter because equation (8.29) is an approximation good for small scales, while the Hurst behaviour should be assessed on large time scales.

Table 8.3 Comparison of the parameters of the ombrian model and the simplified ombrian relationship.

Ombrian model		Simplified ombrian relationship	
Parameter	Meaning of parameter	Parameter	Meaning of parameter
μ	Mean intensity		
λ_1, λ_2	Intensity scale parameters	λ	Intensity scale parameter
α	Time scale parameter for k	α	Time scale parameter for k
		β	Time scale parameter for T
M	Fractal (smoothness) parameter[2]	η	Exponent of the expression of the time scale function $a(k)$
H	Hurst parameter		
θ	Exponent of the expression of probability dry		
ξ	Tail index	ξ	Tail index

Digression 8.B: Limits of the simplified ombrian relationship

For $\xi > 0$, combining equations (8.28)-(8.30) we can write:

$$x = \lambda \frac{\left((T/\beta)^{\xi} - 1\right)}{(1 + k/\alpha)^{\eta}} \tag{8.32}$$

By comparing it with equation (8.4), we see that the two equations are mathematically equivalent, with the parameters α, ξ and η being identical in the two cases, and the remaining two related by:

$$\left\{\beta = \psi'^{1/\xi} D, \qquad \lambda = \lambda' \psi'\right\} \Leftrightarrow \left\{\psi' = (\beta/D)^{\xi}, \qquad \lambda' = \lambda (D/\beta)^{\xi}\right\} \tag{8.33}$$

As the particular form of equation (8.4) has been widespread (for example, in Greece the ombrian curves of all country have been expressed in this form), equation (8.33) is useful for conversion between the two forms in engineering application.

Solving equation (8.32) for T we find the expression of the distribution function of mean intensity x at time scale k as:

$$F^{(k)}(x) = 1 - \frac{k}{T} = 1 - \frac{k}{\beta}\left(1 + \frac{x}{\lambda}\left(1 + \frac{k}{\alpha}\right)^{\eta}\right)^{-1/\xi} \tag{8.34}$$

This latter indeed reflects a Pareto distribution with a discontinuity at zero, which is:

$$P_0^{(k)} = F^{(k)}(0) = 1 - P_1^{(k)} = 1 - \frac{k}{\beta} \tag{8.35}$$

In this respect, at first glance it is consistent with respect to point 4 of section 8.1. However, for large k, this probability may become negative, which is a mathematical inconsistency. In addition, if $k = 0$, $P_0^{(k)} = F^{(k)}(0) = 1$, which means that only the value $x = 0$ is allowed. This is also an inconsistency.

Furthermore, it is easy to find that its mean and squared coefficient of variation, are:

$$E\left[\underline{x}^{(k)}\right] = \frac{\xi\lambda}{(1-\xi)(1+k/\alpha)^{\eta}}\frac{k}{\beta}, \qquad C_v^2\left[\underline{x}^{(k)}\right] = \frac{1-\xi}{1/2-\xi}\frac{\beta}{k} - 1 \tag{8.36}$$

Both these expressions signify inconsistencies with respect to points 2 and 3 of section 8.1. The mean is clearly an increasing function of time scale, tending to infinity as $k \to \infty$, while it should be constant, and becoming zero if $k = 0$, which is absurd. The squared coefficient of variation may become negative for large k, tending to -1 as $k \to \infty$, which is absurd as a square of a real number cannot be negative, and to $+\infty$ as $k \to 0$.

However, if we restrict k so that $k < \beta$ and hence the probability be reasonable ($P_1^{(k)} < 1$), then we can easily infer from (8.36) that $C_v\left[\underline{x}^{(k)}\right] > 1/(1-2\xi) > 0$. In other words, the simplified ombrian relationship has reasonable behaviour for time scales sufficiently smaller than β, even though the constant mean condition will always be violated. Furthermore, to avoid an absurd behaviour close to $k = 0$, we should also restrict k from below. A safe lower bound is the smallest value at which data were available and were used in the construction of ombrian curves.

To get a more specific quantified view of the above, we use as an example the ombrian relationship of Greater Athens (Kephisos River basin). This was derived by Koutsoyiannis et al. (2010) using data of time scales from 5 min to 48 h and assuming a validity for time scales 5 min to 100 h. For altitudes up to 200 m, the ombrian expression is:

$$x = 207\frac{T^{0.15} - 0.61}{(1 + k/0.17)^{0.77}} \tag{8.37}$$

with x in mm/h, T in years and k in h. Using equation (8.33), we can express this in the form of (8.32) with parameters λ = 126 mm/h, α = 0.17 h, β = 325 h, ξ = 0.15 and η = 0.77. Using equations (8.35) and (8.36), we can calculate the probability wet, the mean and the coefficient of variation. These are plotted in Figure 8.1 for time scales covering three orders of magnitude, from 0.1 h (6 min) to 100 h. The probability wet and the coefficient of variation have a reasonable behaviour in this range, as 100 h is much smaller than β =325 h. The constancy of the mean is violated, but this is not a severe limitation for this range of scales, which is dominated by the variance rather than of the mean.

Summarizing the above discourse, the simplified ombrian relationship (equations (8.28)-(8.31)), can give an acceptable approximation of the ombrian relationship for a range of time scales of three orders of magnitude, provided that we have data in that range to fit this model. If we want to go to a wider range of time scales, or if we want to perform tasks other than direct application of the ombrian relationship—e.g. stochastic simulation—then we should use the full model of section 8.2.

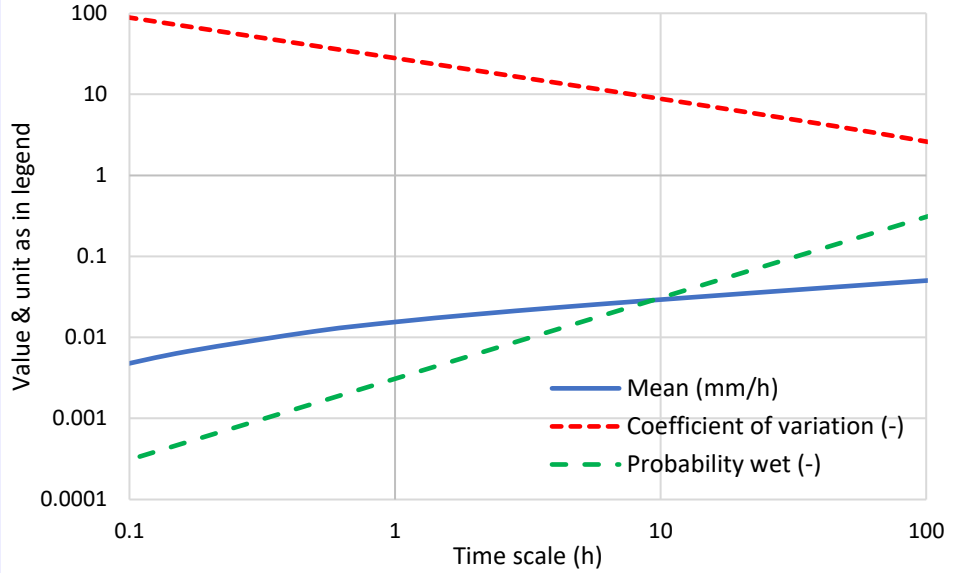

Figure 8.1 Mean, coefficient of variation and probability wet derived from the ombrian relationship of Athens.

8.4 Data availability and processing

For a reliable estimation of ombrian curves, it is important to utilize all available rainfall observations on all time scales. Modern rainfall measuring devices are sensors which readily provide digital information at small time steps (e.g. 10 min) but the older mechanical autographic devices should never be neglected, even though digitization of the archive of their recording charts is tedious.

It has been a common practice to base the construction of the ombrian curves of a certain area on the data of subdaily time scale only. However, this is a problematic practice that leaves out important information. As first noted in Koutsoyiannis et al. (1998), the (usually much longer) daily rainfall observation records can be fully utilized for a more reliable model fitting.

It is a strong suggestion of this text to combine and use the entire data sets of all types of devices and, as explained in Digression 2.K, work on the parent distribution rather than extracting values over threshold or, even worse, time-block (e.g. annual) extremes. It is noted though that in some cases the availability of data is such that does not allow access to the full information. For example, in several cases only a few the extreme rainfall events have been digitized while the majority of rainfall recordings remain in charts. In such cases adaptation of the ombrian model may be required, followed by conversion of the final results so that they correspond to the parent distribution, which is the natural basis for estimating design quantities. In particular, if the data used are block maxima, then the Pareto parent distribution corresponds to EV2 distribution of extremes. Therefore, the model fitting should be made on the EV2, rather than the Pareto, distribution but the final model should be formulated for the Pareto distribution, using the parameter values that were estimated for EV2 distribution (see more details in section 2.19 and Digression 2.K).

Some of the model parameters are more sensitive to the availability of data of a particular temporal resolution, as summarized in Table 8.4. Thus, the reliable estimation

of those parameters depends crucially on the availability of the particular resolution on which is more sensitive.

Table 8.4 Crucial sensitivity on particular temporal resolution of the parameters of the ombrian model and the simplified ombrian relationship.

Temporal resolution	Parameters that are most sensitive to the data type
Sub-hourly	α (Time scale parameter for k), M (Fractal/smoothness parameter)
Sub-daily	θ (Exponent of the expression of probability dry), η (Exponent of the expression of the time scale function $a(k)$)
Daily and higher	μ (Mean intensity), ξ (Tail index), H (Hurst parameter), β (Time scale parameter for T), λ_1, λ_2 (Intensity scale parameters)

Digression 8.C Do we need a sliding window and a Hershfield coefficient?

When studying a process on multiple scales (e.g. to infer the climacogram of the process), we aggregate the available data from several time series to different time scales. No particular provision is made for the starting point for aggregation of each time series. To make this clearer, let us assume that we have a daily time series x_τ and from this we construct the 2-day time series $x_\tau^{(2)}$. Actually, we can construct two different time series $x_\tau^{(2)}$, depending on the selection we have made for the first term. Namely, the $x_1^{(2)}$ that contains the daily term x_2 could be either $(x_1 + x_2)/2$ or $(x_2 + x_3)/2$. Likewise, if we construct a time series at time scale 10, there are 10 variants (the first term $x_1^{(10)}$ that contains the daily term x_{10} could be anyone among $(x_1 + \cdots + x_{10})/10$ through $(x_{10} + \cdots + x_{19})/10$). All these are numerically different time series. But their statistical characteristics are precisely the same. Since we are doing stochastics and we are interested on the statistical characteristics (and not on time series values) all options are equivalent. Thus, the notion of a sliding window is unnecessary for our study.

However, when studying extremes, it has been a tradition in hydrological practice (e.g. Linsley et al., 1975, p. 357) to use a sliding window and take the maximum value among the variants. In the above example, instead of constructing a time series $x_\tau^{(2)}$ whose first term would be, e.g., $x_1^{(2)}=(x_1 + x_2)/2$, we use the notion of a sliding window to construct the time series $y_\tau^{(2)}$ whose first term is:

$$y_1^{(2)} \coloneqq \max\{(x_1 + x_2)/2, (x_2 + x_3)/2\} = (x_2 + \max\{x_1, x_3\})/2 \tag{8.38}$$

Let us consider the ratio of the expectations of the two cases, $\mathrm{E}\left[\underline{y}_1^{(2)}\right]/\mathrm{E}\left[\underline{x}_1^{(2)}\right]$. If we temporarily assume a fully random process, then this ratio can be expressed in terms of noncentral K-moments as:

$$\frac{\mathrm{E}\left[y_1^{(2)}\right]}{\mathrm{E}\left[x_1^{(2)}\right]} = \frac{K_1' + K_2'}{2K_1'} = \frac{1}{2} + \frac{K_2'}{2K_1'} \tag{8.39}$$

Thus, in an exponential distribution of the intensity, in which $K_2'/K_1' = 1.5$ (see Table 6.3), we will have a ratio equal to 1.25. As we increase the time scale k beyond 2, the corresponding ratio $\mathrm{E}\left[\underline{y}_1^{(k)}\right]/\mathrm{E}\left[\underline{x}_1^{(k)}\right]$ will converge fast to a value slightly higher than 1.25. However, because of time dependence this coefficient turns out to be much smaller. This decreased value, empirically estimated from rainfall data, is usually termed the *Hershfield coefficient* after Hershfield and Wilson (1957) who first studied it and noted "*It has been determined that, on the average, the maximum rainfall in any consecutive 60-minute period is 13 percent greater than the clock-hour*

rainfall for the same frequency for the corresponding period of record at most stations. Similarly, and by coincidence, the same factor applies to daily rainfall; to convert observation-day rainfall for a particular frequency to the maximum 1440-minute rainfall for the same frequency, multiply by 1.13."

The value of 1.13 has been extensively used worldwide and several later studies confirmed, rather than invalidated it. Specifically, studies of maxima on multiple time scales have been based on the series $y_\tau^{(k)}$, rather than $x_\tau^{(k)}$, determined as above. For the lowest available time scale, $k = 1$ (the time step of the original series) as this method can no longer be applied, the values $y_\tau^{(1)}$ are calculated by multiplying x_τ by 1.13.

However, this tactic distorts the stochastic behaviour of the process $\underline{x}_\tau^{(k)}$ which is to be studied. As clarified above, the ombrian model is a stochastic model of the average intensity $x_\tau^{(k)}$ at any time scale k. The quantity $y_\tau^{(k)}$ is something different from $x_\tau^{(k)}$ and there is no need to study it at all. Therefore, the notion of the sliding window is not recommended to use. A fixed time window, with any arbitrary starting time is what is actually needed, without any conversion of the original time series, except taking temporal averages at several time scales. For consistency, a fixed, rather than sliding, window should be used even in extracting block maxima.

8.5 Ombrian model fitting

Assuming that the ombrian model parameters are known, we can determine theoretically, based on the equations grouped together in Table 8.1, the following quantities:

- the climacogram as a function of time scale $\gamma(k)$;
- the probability wet as a function of time scale, $P_1^{(k)}$; and
- the rainfall intensity as a function of the time scale and the return period, $x(k, T)$.

On the other hand, from the available data series, each one referring to a specific time scale k, we can determine empirical estimates of:

- the standard climacogram estimate $\hat{\gamma}(k)$ using equation (4.23); and
- the probability wet, $\hat{P}_1^{(k)} = \hat{n}_1/n$, where $\hat{n}_1$ is the number of nonzero observations and n is the total number of observations in the time series of observations (where both $\hat{n}_1$ and n depend on k).

If we have both the model and the data series, then for each series referring to a specific time scale k, we can make tables of empirical values of intensities, x and corresponding return periods, T, in two ways:

- based on K-moments, or
- based or order statistics.

For the approach based on the K-moments (Chapter 6), we can implement the following algorithm for each specified time scale k.

1. We calculate the theoretical probability, $P_1^{(k)}$, from the number of observations n in the sample we specify $n_1 = P_1^{(k)} n$ and we choose the n_1 largest values from the time series for further processing. (In a perfect model fit, $n_1 = \hat{n}_1$, i.e., the observed number of nonzero values).

2. By adapting equation (6.93), we calculate the bias correction factor Θ. Since the model is not a pure HK model, to estimate Θ we modify equation (6.93) neglecting the first term of its right-hand side, which is small, and adapting the second term as:

$$\Theta(k, L, H) \approx -\frac{\gamma(L)}{2\gamma(k)} \tag{8.40}$$

   This is obtained by inspection of equations (6.93) (4.24) and (3.81)
3. From the equations of Table 6.6 (entries on the Pareto and PBF distributions, given the tail index ξ of the model, we calculate the Λ-coefficients Λ_1 and Λ_∞.
4. We choose a number m of moment orders p ranging (in geometric progression) from 1 to n_1 and for each one we estimate the noncentral K-moment $\underline{\widehat{K}}'_p$ using the equations (6.65)-(6.67).
5. For each order p we estimate the bias corrected order p' from equation (6.95).
6. For each p' we estimate the Λ-coefficients $\Lambda_{p'}$ from equation (6.111) and the return period from (6.112), which we adapt to the following formula*:

$$T\left(\underline{\widehat{K}}'_p\right) = \frac{k}{P_1^{(k)}} p' \Lambda_{p'} \approx \frac{k}{P_1^{(k)}} \left(\Lambda_\infty p' + (\Lambda_1 - \Lambda_\infty)\right) \tag{8.41}$$

7. To the return period $T\left(\underline{\widehat{K}}'_p\right)$ so calculated, there corresponds a value $x = \underline{\widehat{K}}'_p$; repeating this procedure for all p we make a table of empirical values of x and corresponding T.

Alternatively, we can construct the required table using order statistics. In this case, the standard procedure of assigning return period to sample values is simpler (section 5.6) but it does not take into account the persistence, as in the case of K-moments. Here we adapt the standard procedure to take it into account by the following quick-and-dirty manner for each specified time scale k.

1. Using the tail index ξ of the model we calculate the coefficients A and B of equation (5.57) and Table 5.5 (formula VI for approximation of the Pareto distribution).
2. We make a first estimate T of the return period of each nonzero value x from equation (5.57) based on the rank i of each sample observation $x_{(i:n)}$, sorted in ascending order; this estimate is based on the assumption of independence.
3. From equation (8.40) we calculate the bias correction factor $\Theta(k, L, H)$.
4. Based on the simplified approximation (rule of thumb) of the relationship of return period and order of K-moment, expressed in equations (6.106)-(6.107), and

* To check the formula, let us consider k = 1 h, $P_1^{(k)} = 0.03$, and period of observations of 100 years = 876 600 h, and assume $\Lambda_1 = \Lambda_\infty = 2$. Then n = 876 600, $n_1 = P_1^{(k)} n = 26\,296$; for the maximum value of $p = n_1 = 26\,298$, assuming independence so that $p' = p$, the return period will be $T\left(K'_p\right) = \left(k/P_1^{(k)}\right)\left(\Lambda_\infty p' + (\Lambda_1 - \Lambda_\infty)\right) = (1/0.03)(2 \times 26\,298) = 1\,753\,200$ h = 200 years (as expected, because $\Lambda_\infty = 2$).

combining with equation (6.95), we estimate the adapted return period to take persistence into account as follows:

$$T' \approx \min\left(\left(2\Theta + (1-2\Theta)\left(\frac{T\hat{P}_1^{(k)}}{2k}\right)^{(1+\Theta)^2}\right)\frac{2k}{\hat{P}_1^{(k)}}, T\right) \tag{8.42}$$

5. Repeating this procedure for all nonzero $x_{(i:n)}$ we make a table of empirical values of x and corresponding T'.

Prior to most of the above calculations, we need to have assumed an ombrian model and specify its parameter values. We may start from some guesses and find the final values by minimizing the error between theoretical and estimated statistics. Such errors are nonlinear functions of the parameters and we need a nonlinear solver to perform the minimization. Such solvers are now common in most numerical software platforms (including spreadsheets).

If we wish to optimize the climacogram, then we can formulate the fitting error as:

$$e_\gamma \coloneqq \sum_k w_\gamma(k)\big(\ln\big(\gamma(k)-\gamma(L)\big) - \ln\big(\hat{\gamma}(k)\big)\big)^2 \tag{8.43}$$

where L is the observation period and $w_\gamma(k)$ denotes some weight, which can be chosen as a function of k. As the climacogram spans several orders of magnitude, it is advisable to compare logarithms rather than actual values. Furthermore, as articulated in section 4.6, due to the presence of bias, the estimate $\hat{\gamma}(k)$ is not comparable to the theoretical climacogram $\gamma(k)$ but rather to the theoretical expectation of its estimator, based on equation (4.24), i.e., $\mathrm{E}\left[\underline{\hat{\gamma}}(k)\right] = \gamma(k) - \gamma(L)$. These notes explain the specific mathematical form of equation (8.43). By minimizing e_γ we can determine the parameters related to the climacogram. However, the exponents θ and ξ cannot be determined from the minimization of e_γ.

In a similar manner, we can define the fitting error in the probability wet (or dry) as:

$$e_P \coloneqq \sum_k w_P(k)\left(P_1^{(k)} - \hat{P}_1^{(k)}\right)^2 \tag{8.44}$$

where $w_P(k)$ denotes some weight, which can again be chosen as a function of k. As all parameters of the ombrian model are involved in the mathematical expression of $P_1^{(k)}$, the minimization of e_P can determine all parameters. However, as no model is a perfect description of reality, this type of specification of parameters, which focuses on the dry part of the rainfall process, is not good for extreme rainfall.

For this reason, it is more advisable to fit by minimizing an error metric focusing on distribution quantiles $x(k,T)$ for all available time scales k and a series of return periods, as described above. The total fitting error in this case is:

$$e_x \coloneqq \sum_k \frac{1}{\gamma(k)} \frac{1}{n_k} \sum_T w_x(T)\left(x(k,T) - \hat{x}(k,T)\right)^2 \tag{8.45}$$

where $w_x(T)$ is a weighting factor, determined as a function of the return period T, and n_k is the number of x values at time scale k. The total square error over the entire set of return periods (the second sum in the right-hand side of (8.45)) is further normalized by the climacogram $\gamma(k)$.

A combined optimization would take into account all above three error metrics in a linear combination with weights a_γ, a_P, a_x, i.e.:

$$e \coloneqq a_\gamma e_\gamma + a_P e_P + a_x e_x \tag{8.46}$$

This can give a best compromise in a simple manner, even though multivariate (Pareto) optimization would provide a more sophisticated approach.

Digression 8.D: Ombrian model for Uccle, Belgium

We will illustrate the ombrian model against data using very long observational records. We start with the meteorological station at Uccle (a suburb of Brussels), Belgium (50.80°N, 4.37°E, 100.0 m), which is perhaps the one with longest sub-hourly rainfall record worldwide. It belongs to the Royal Meteorological Institute of Belgium (RMIB) and its recording started in 1898. Here the data from 1898 to 2017 (not publicly available) with a gap in 2003 (119 years in total) have been used at the minimum available time step, which is 10 min, and at aggregate time scales up to 96 h. In addition, the daily precipitation record, publicly available through KNMI's (Koninklijk Nederlands Meteorologisch Instituut) Climexp system, by accessing the European Climate Assessment & Dataset, has been used[1]. This covers the period 1880-2018 (139 years) with only very few missing daily values, which were left unfilled. The time scales of investigation start from the minimum available, i.e., daily, and advance up to 13 years (so that the aggregate time series have at least 10 values). The sub-hourly and daily records are generally in good agreement with each other up to 1999 but later there are notable deviations.

We fit the model simultaneously at both sources of data. We aggregate the data of the original 10 min time step to time scales of 0.5, 1, 2, 4, 6, 12, 24 h, and 1, 2, 4 d (11 time scales in total). Also, we aggregate the data of the original 1 d time step to time scales of 2, 4, 8, 16, 32, 64, 128 d, and 0.5, 1, 2, 4, 8, 13 years (14 time scales in total). By choosing not to use time scales > 96 h for the sub-hourly data and also to use a larger number of time scales for the daily data, we give more emphasis on the latter, as daily data are generally deemed more reliable than (sub)hourly, particularly on the large time scales.

Here we use the fitting approach based on the order statistics, as described in section 8.5. For each of the time scales of investigation we estimate from the sample the variance (climacogram) and probability wet and, once model parameters are assumed, the return period of each nonzero intensity value. The form of the theoretical climacogram we use is the FHK-C (equation (8.8)).

A first model fitting has been based on merely the climacogram, on the basis of equation (8.43), assuming equal weights, i.e., $w_\gamma(k) = 1$. In this case we estimated merely the four parameters that appear in equation (8.8). Another fitting has been based on the probability wet, on the basis of equation (8.44) assuming equal weights, i.e., $w_P(k) = 1$. In this case all seven parameters are estimated. However, from an engineering point of view a more useful fitting is that based on the total fitting error of distribution quantiles on the basis of equation (8.45). As in the approach based on order statistics the low return periods appear much more frequently than the high ones, if the weights $w_x(T)$ are set equal, then the fitting emphasis will be given on the small return periods. To avoid this, we have chosen an increasing function $w_x(T)$, namely $w_x(T) \propto \sqrt{T}$. Finally, a combined fitting on the basis of equation (8.46) has been performed with weights $a_\gamma = 0.1, a_P =$

$100, a_x = 1$. (Note that the chosen high value of a_P counterbalances the fact that e_P is much smaller than the other error components.)

The fitted parameters in all cases are shown in Table 8.5. The optimization cases for quantiles and combined resulted in virtually the same parameter values and thus one entry appears in Table 8.5 for both. By inspecting the table, we see that the parameters of different optimization objectives are fairly stable (do not change much with change of the objective function) are μ, H, θ and ξ. Notable is the high value of ξ (0.20, against typical values of 0.1-0.15) and the moderate H (0.6).

Table 8.5 Parameters of the ombrian model of Uccle.

Case of optimization[1]	μ (mm/h) [2]	λ(mm^2/h^2)	α (h)	M (-)	H (-)	θ (-)	ξ (-)
Climacogram (e_γ)	-	0.281	0.645	0.28	0.58	-	-
Probability wet (e_P)	0.0905	0.794	0.722	0.20	0.60	0.630	0.200
Quantiles (e_x) & Combined (e)	0.0916	1.387	0.140	0.50	0.62	0.573	0.194
Quantiles (e_x) for subdomain[3]	0.2454	2.848	0.250	1	0.56	1	0.123

[1] The transition time scale k^* was chosen 12 h.

[2] The mean estimate is 0.0916 mm/h for the hourly series and 0.0905 for the daily series. In the quantiles/combined optimization cases the value was derived by optimization.

[3] The subdomain is defined as $k \leq 2$ d & $T > 2$ years.

The empirical and theoretical climacograms are shown in Figure 8.2. The model can obtain a perfect climacogram fitting, if the optimization objective is the climacogram per se, but even in the combined optimization the fitting remains good. Likewise, as seen in Figure 8.3, the model can obtain a perfect fitting on the probability wet, if the optimization objective is this latter, but even in the combined optimization the fitting remains relatively good.

The fitting on distribution quantiles for the combined optimization is shown in Figure 8.4, which is close to a typical depiction of ombrian relationships, except for the fact that here the time scales span 6 orders of magnitude (10 min = 0.17 h to 13 years = 113 958 h) and the return periods span almost 5 orders of magnitude. The fitting is generally good for those impressively wide spans of time scales and return periods and thus supports the suitability of the ombrian model.

If we delimit the ranges of time scales and return periods to those used in typical ombrian curves, we can obtain an even better fitting. This is illustrated in Figure 8.5 for $k \leq 2$ d and $T \geq 1$ year. The ombrian relationships appear in this subdomain as straight lines in the double logarithmic plot of Figure 8.5, a fact that is characteristic for the tails of the Pareto distribution and has enticed the fans of multifractals to perceive this behaviour as the magic of power laws.

The fitting on quantiles described above has also been used in this case and the resulting parameter values also appear in Table 8.5. Interestingly, the parameter values differ substantially in this case and, in particular, the mean does not approach its standard empirical estimate, as also happens in common approaches of construction of ombrian curves.

It is noted, though, that a more simplified model (section 8.3) and a simplified model fitting method (section 8.6) can be used if we are interested in that subdomain only. The real power of the full ombrian model is its coverage of all time scales and return periods, as well as its direct applicability in stochastic simulation.

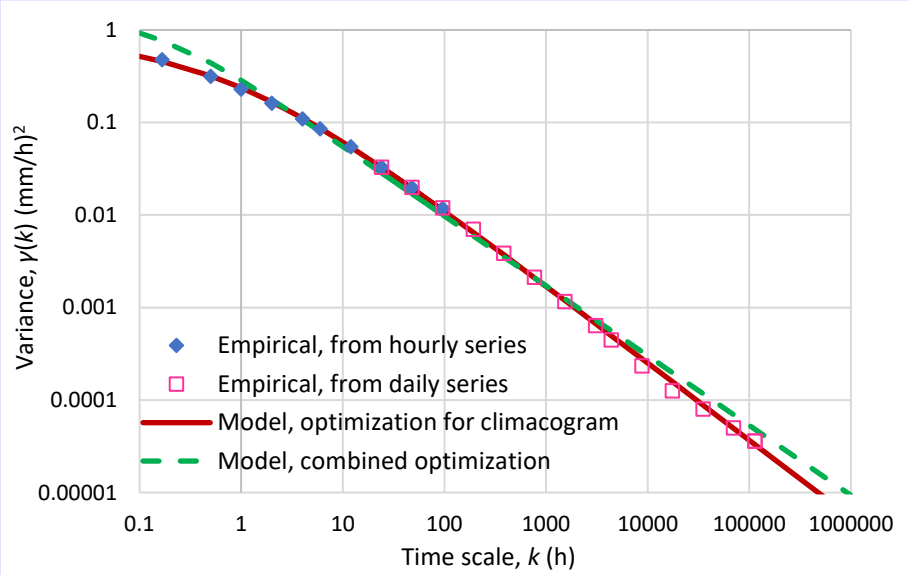

Figure 8.2 Fitting of the ombrian model (equation (8.8)) to the empirical estimates of the climacogram for Uccle. Note that the bias in the climacogram is negligible due to low Hurst parameter.

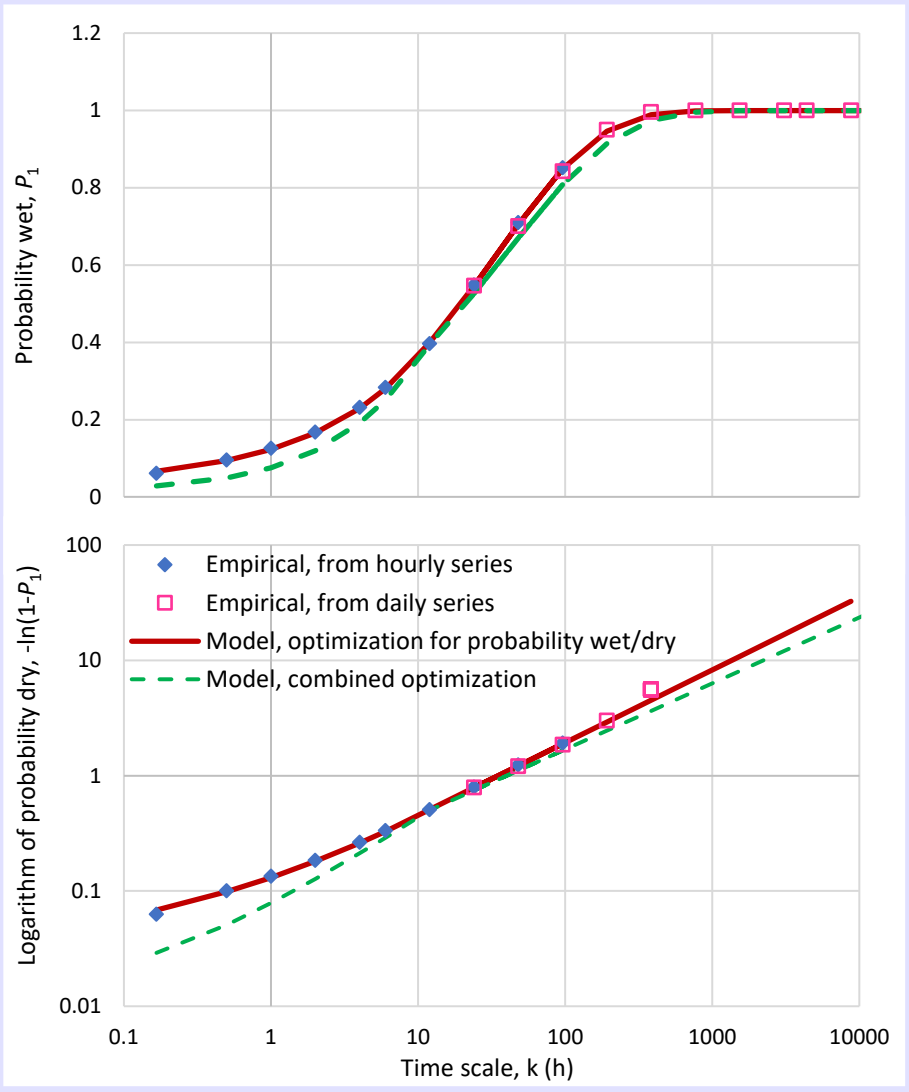

Figure 8.3 Fitting of the ombrian model (equations (8.10) and (8.13)) to the empirical estimates of probability wet (P_1) or dry ($1 - P_1$) for Uccle.

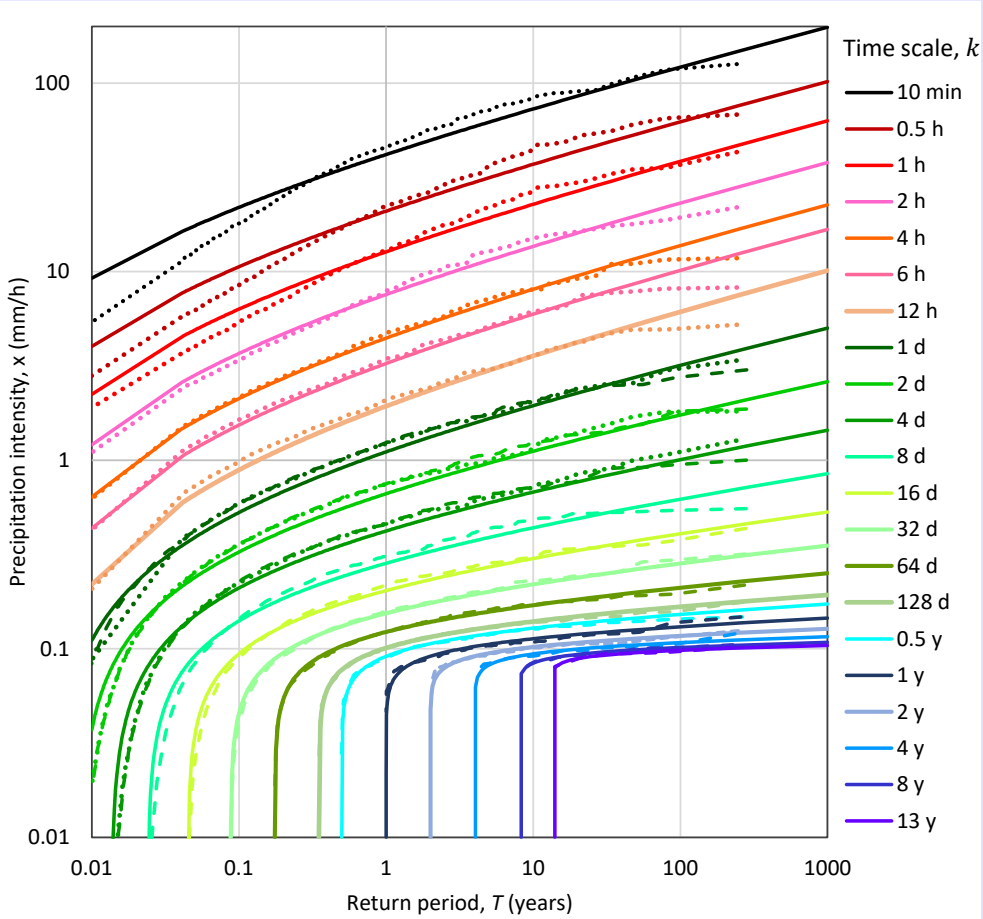

Figure 8.4 Ombrian relationships as resulted from the ombrian model for Uccle for time scales spanning 6 orders of magnitude (10 min = 0.17 h to 13 years = 113 958 h). The empirical points are estimated from order statistics (using formula VI of Table 5.5) taking into account the effect of persistence. Continuous, dashed and dotted lines represent the theoretical values of model, the empirical estimates of daily series and those of the hourly series, respectively. The abbreviation "y" stands for year.

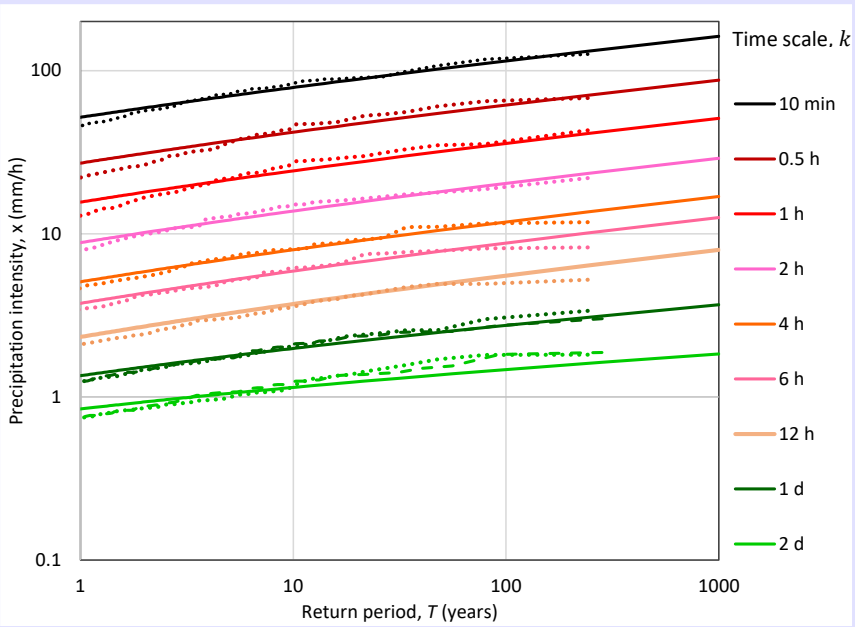

Figure 8.5 As in Figure 8.4 but with fitting focused on time scales ≤ 2 d and return periods ≥ 1 year.

[1] Climexp section "blended ECA&D", data access 2020-08-10; data time stamp 2019-05-23.

Digression 8.E: Ombrian model for Bologna, Italy

As already described in section 1.3, Bologna, Italy, has one of the longest daily rainfall records worldwide, currently 206 years. Hourly rainfall data of the Bologna station are also available but for a much shorter period, 1990-2013, and are provided again by the Dext3r repository (retrieved and processed by Lombardo et al., 2019). The total length is 23 years, as the entire 2008 is missing.

Here we use the same ombrian model as in the Uccle case (Digression 8.D) except that we use the FHK-CD type climacogram (equation (8.9), which is more appropriate for the more complex shape of the empirical climacogram, seen in Figure 8.6. In addition to applying the procedure used in Uccle, which was based on order statistics, here we also apply the K-moments approach.

The parameter values for the optimization cases examined by all approaches and cases are shown in Table 8.6. In comparison to Uccle, the most notable difference is the very high Hurst parameter (≥ 0.92, comparable to the value 0.90 that has been already estimated for the annual rainfall in Bologna in section 1.3). This has visible effects in terms of high bias in the climacogram for scales $k > 1000$ h (about 40 d) in Figure 8.6.

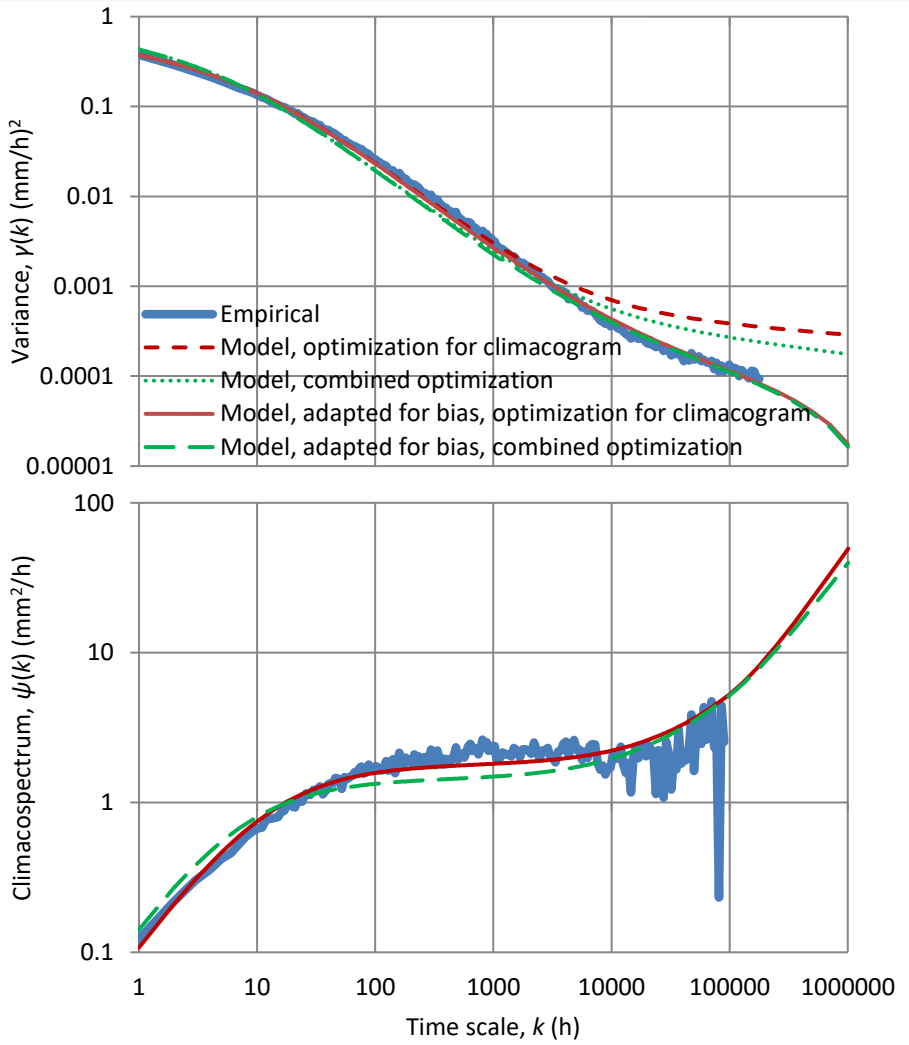

Figure 8.6 Fitting of the ombrian model (equation (8.9)) to the empirical estimates of the climacogram **(upper)** and climacospectrum **(lower)** for Bologna. The empirical estimates for time scales smaller than or greater than 1000 h (~42 d) are taken from the hourly and daily series, respectively. Note that the bias in the climacospectrum graph is negligible.

Table 8.6 Parameters of the ombrian model of Bologna.

Case of optimization[1]	μ (mm/h)	λ_1(mm^2/h^2)	λ_2(mm^2/h^2)	α (h)	H (-)	θ (-)	ξ (-)
Climacogram (e_γ)	-	0.000864	1.51	16.4	0.95	-	-
Probability wet (e_P)	0.0773	0.00775	0.836	14.15	0.95	0.795	0.121
Based on order statistics							
Quantiles (e_x)	0.0788	0.00407	1.60	7.70	0.93	0.693	0.125
Combined (e)	0.0823	0.00110	1.43	8.74	0.92	0.787	0.121
Based on K-moments							
Quantiles (e_x), entire domain	0.0775	0.00088	2.24	8.42	0.95	0.801	0.114
Quantiles (e_x), subdomain[2]	0.1515	0.01238	5.01	5.46	0.95	-	0.083

[1] The transition time scale k^* was chosen 96 h (= 4 d).

[2] The subdomain is defined as $T > 2$ years and $k \leq 4$ d; for these time scales the distribution is Pareto and therefore θ cannot be estimated.

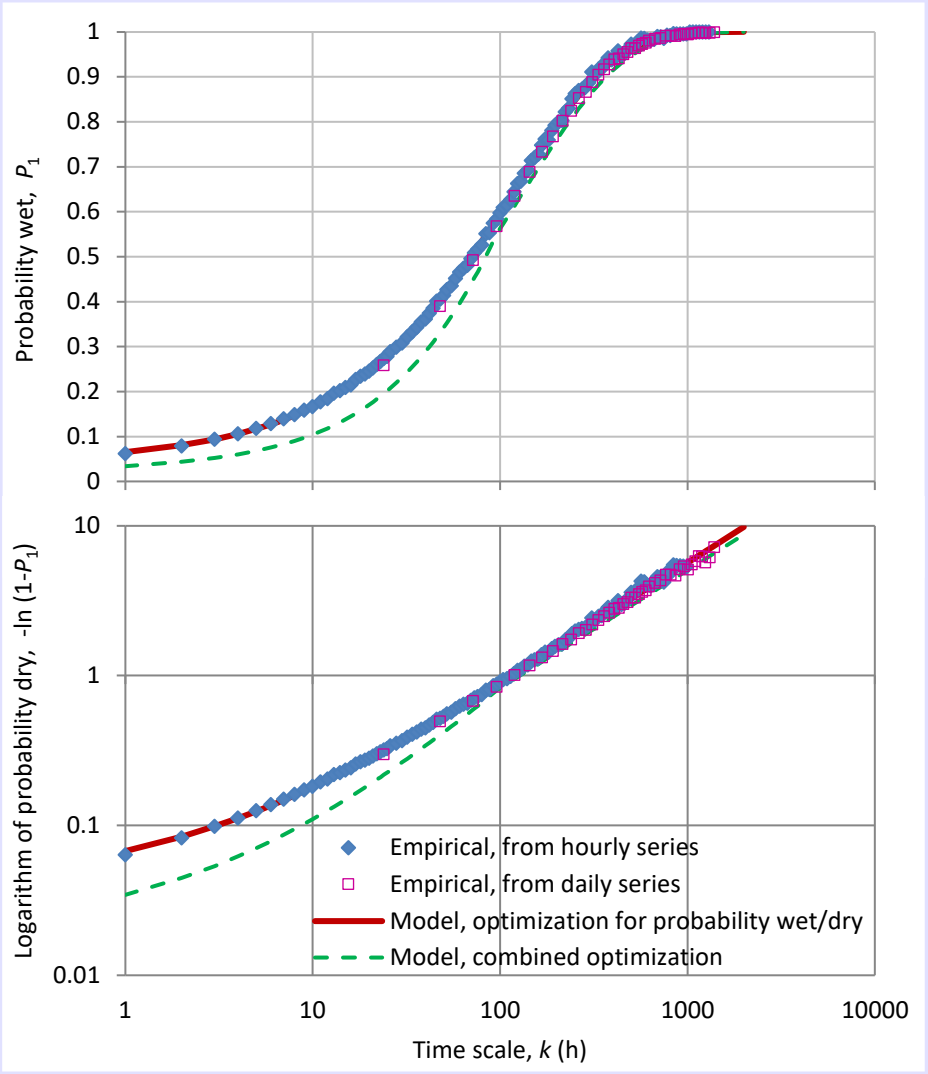

Figure 8.7 Fitting of the ombrian model (equations (8.10) and (8.13)) to the empirical estimates of probability wet (P_1) or dry ($1 - P_1$) of Bologna.

As seen in Figure 8.6, the empirical and theoretical climacograms compare very well or even perfectly for a fitting on the basis of the climacogram. This figure also includes the climacospectrum of the process, where the fitting is good enough. Likewise, as seen in Figure 8.7, the model can obtain a perfect fitting on the probability wet, if the optimization objective is this, but even in the combined optimization the fitting remains relatively good.

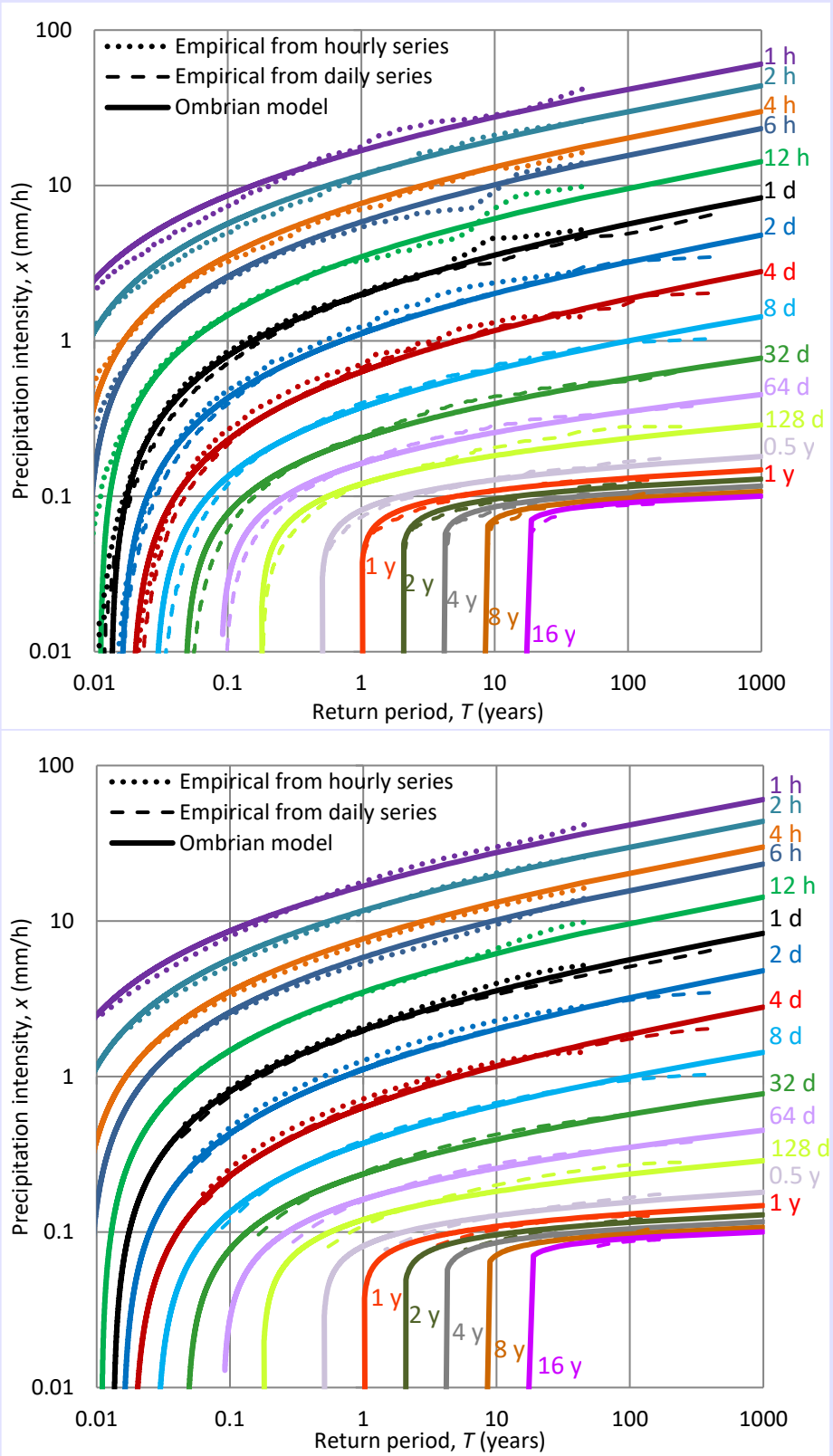

Figure 8.8 Ombrian curves as resulted from the ombrian model for Bologna for time scales spanning 5 orders of magnitude (1 h to 16 years = 140256 h). The empirical points are estimated from order statistics (**upper**; using formula VI of Table 5.5) and K-moments **(lower)**. The optimization was based on quantiles estimated from order statistics (Table 8.6; case of optimization: combined). In both cases the effect of persistence was taken into account; the ombrian model results were plotted for bias-adapted variance in order to be comparable with empirical plots (thus, for $k > 1000$ h or about 40 d, the true intensity resulting from the model is higher than what is shown in the graph). The abbreviation "y" stands for year.

The fitting on distribution quantiles for the combined optimization is shown in Figure 8.8, with return periods assigned by order statistics (upper graph) or by K-moments (lower graph). In both cases adaptations to take into account the bias due to the intense HK behaviour have been performed. There are no noteworthy differences between the two graphs, except the smother empirical curves in K-moments graph. The fitting is good for the entire range of time scales and return periods, each of which spans five orders of magnitude. Again, the good fitting supports the suitability of the ombrian model.

While the lower panel of Figure 8.8 shows empirical quantiles based on K-moments, the model parameters are the same as in the upper panel, i.e. those estimated using order statistics. Now, Figure 8.9 shows similar information but for an optimization based on empirical quantiles estimated by K-moments (see Table 8.6, penultimate raw). We note that the latter are noncentral K-moments estimated at a number $m + 1$ orders, where m = 80 i.e., $p_i = n^{i/m}, i = 0, \ldots, m$, with $p_0 = 1, p_m = n$. The choice of $m \ll n$ substantially reduces the computational burden without compromising accuracy. The orders p_i need not be natural numbers. The estimation of the return periods of the noncentral K-moments K'_{p_i} of orders p_i is made using the linear approximation. (It will be shown in Chapter 9 that this approximation is quite satisfactory for fitting purposes, with negligible differences from the nonlinear approximation). The lowest order $p_0 = 1$ yields the mean $K'_1 \equiv \mu$ and therefore this procedure does not allow to derive estimates below the mean. This is absolutely consistent with the purpose of fitting an ombrian model for rainfall extremes, where obviously rainfall intensities lower than the mean are not of interest. However, in other applications, such as in estimation of low flows, small values are important. In this case we can estimate low quantiles using the tail K-moments $\overline{K}'_{p_i}$, as will be detailed in Chapter 9 . For the same reason, the plots of theoretical distributions below the overall mean (μ = 0.0775 mm/h) are been shown in the figure.

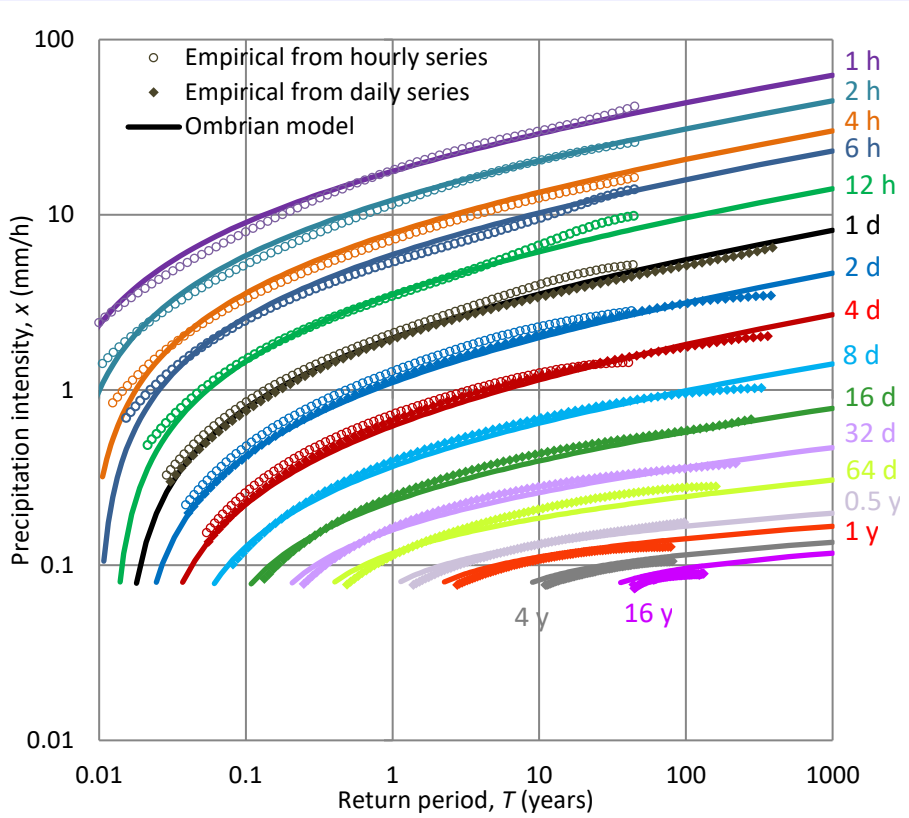

Figure 8.9 Ombrian curves for Bologna as in Figure 8.8 but with the optimization based on quantiles estimated from noncentral K-moments, which are also plotted in the figure along with the theoretical model. The effect of persistence was taken into account. The abbreviation "y" stands for year.

If we delimit the ranges of time scales and return periods to those used in typical ombrian curves, we can reduce the fitting error. The resulting fitting is shown in the right panel of Figure 8.10, where the optimization of the model error (e_x) was made on the subdomain $T \geq 2$ years and $k \leq 4$ d. In this case, the Pareto distribution applies in all subdomain. The ombrian relationships appear in this subdomain as straight lines in the double logarithmic plot, a fact that is characteristic for the tails of the Pareto distribution. The left panel of Figure 8.10 is similar to the right, but the parameters are those fitted on the entire domain, as in Figure 8.9. While, by construction, the model error (e_x) is smaller in the right panel than in the first, a visual inspection shows that the model fitting of the left panel may be preferable from a practical point of view, and has the additional advantage of being applicable on the entire range of return periods and time scales.

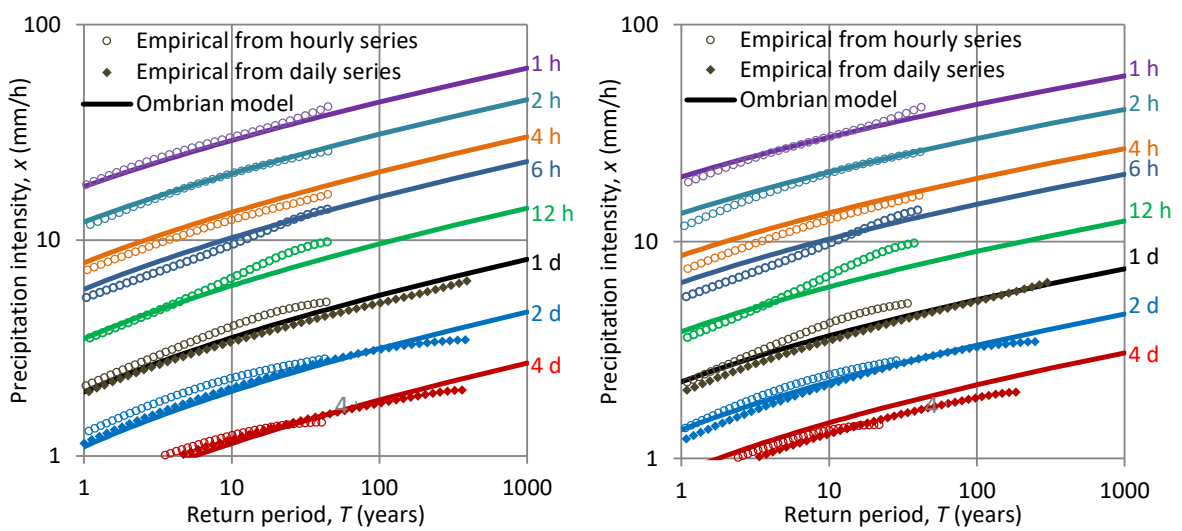

Figure 8.10 Ombrian curves for Bologna as Figure 8.9 but focusing on large return periods and small time scales, with the optimization based on quantiles estimated from noncentral K-moments on **(left)**, the entire domain (same is Figure 8.9) and **(right)** the subdomain $T > 2$, $k \leq 4$ d (see Table 8.6).

8.6 Simplified ombrian relationship fitting

The procedures of section 8.5 can also be applied in the simplified version of ombrian relationships of section 8.3, in the form of equation (8.28). In addition, the separability of functions $a(k)$ and $b(T)$ allows a two-step approach, with each step determining the parameters of each of the two functions separately. The procedure was introduced by Koutsoyiannis et al. (1998) and is based on expressing (8.28) in the form:

$$a(k)x = \lambda b(T) \tag{8.47}$$

We observe that the time scale k is not a stochastic variable (rather it takes on a set of values depending on the data availability) and $a(k)$ is a deterministic function thereof, while the right-hand side of the equation in essence is an expression of the (Pareto) distribution function, which does not depend on k. By substitution of equations (8.29) and (8.30), we can write (8.47) as:

$$\left(1+\frac{k}{\alpha}\right)^{\eta} x = \lambda\left(\left(\frac{T}{\beta}\right)^{\xi} - 1\right) \tag{8.48}$$

Hence, for the different time scales k_j the stochastic variables $\underline{y_j} = a(k_j)\underline{x} = (1 + k/\alpha)^{\eta}\underline{x}$ have a common distribution function. Thus, the $\underline{y_j}$ for different k_j can be regarded as samples from the same distribution. Let $\underline{y_{ji}} \coloneqq a(k_j)\underline{x_{ji}}$ where $\underline{x_{ji}}$ is the ith item of the sub-sample of size n_j corresponding to the time scale k_j and let $\underline{r_{ji}}$ be its rank in the merged sample of all the $\underline{y_{ji}}$ of size $n = \sum_j n_j$. Let the mean rank of each sub-sample be $\underline{r_j} = \sum_i \underline{r_{ji}}\,/n_j$. We replace all $\underline{r_{ji}}$ of each sub-sample with its mean $\underline{r_j}$ and we get a sample of size n with n_1 values equal to $\underline{r_1}$, n_2 values equal to $\underline{r_2}$, etc. The estimators of its mean and variance* will be:

$$\overline{\underline{r}} \coloneqq \frac{1}{n}\sum_j n_j \underline{r_j}\,, \qquad \underline{\gamma_r} \coloneqq \frac{1}{n}\sum_j n_j \left(\underline{r_j} - \overline{\underline{r}}\right)^2 \tag{8.49}$$

where if there are no ties among different mean ranks, it is easy to see that $\overline{\underline{r}} = (n + 1)/2$ (a constant value as $\underline{r_{ji}}$ vary from 1 to n).

Now, it is easy to understand that if the samples are from the same distribution, that in the right-hand side of the equation (8.47), then each $\underline{r_j}$ should be close to the mean and the variance $\underline{\gamma_r}$ should be minimal. Furthermore, given the observations x_{ji}, the variance estimate γ_r depends on the parameters α and η. Thus, we form a minimization problem, seeking to find the values α and η that minimize γ_r. With current computational tools (even common spreadsheet software) numerical minimization of a function of two variables is an easy task (it can even be solved even without using a solver, by a trial-and-error method).

We further note that the method could even work with the variables $\underline{y_{ji}}$ instead of their ranks $\underline{r_{ji}}$. Nonetheless, using ranks makes the method more robust, i.e., not affected by the presence of outliers in the samples.

For the sake of improving the fitting of $b(d)$ in the region of higher intensities (and also to simplify the calculations) it may be preferable to use in this first step of calculations a part of the data values of each group instead of the complete series. For example, we can use the highest 1/2 or 1/3 of intensity values for each time scale (Koutsoyiannis et al., 1998).

Once the values of α and η are determined, we proceed to the second step of calculations, which is very easy. Assuming that, with these values, all $\underline{y_{ji}}$ are from the same distribution, we merge all k groups of values $\underline{y_{ji}}$ thus forming a single sample. To finalize the task, it suffices to estimate the parameters of the Pareto distribution using e.g. the method of K-moments. This defines completely the form and the parameters of $b(T)$.

An advantage of the two-step method is that it allows giving different roles to different data sets in the fitting procedure. Thus, in the first step the parameters α (which is typically smaller than 1 h and needs sub-hourly data to be reliably estimated) and η (for

* We notice that the variance resembles the Kruskal-Wallis statistic used to test whether several samples are from the same distribution. However, here we do not apply any test, nor would it be possible, as the test assumes independent samples, while clearly here they are dependent.

which hourly or multi-hour time scales are most appropriate) should be based on subdaily and even sub-hourly data. In contrast, the parameters of $b(T)$ are better deduced from daily raingauge data rather than from autographic rain recorder data, because the latter are more susceptible to measurement errors and also of shorter length. In particular the tail index ξ of $b(T)$ should ideally be based on multi-station data of the area, or be assumed independently of data, according to experience in the area of study.

Digression 8.F: Tail index of rainfall intensity worldwide

Both the full ombrian model and the simplified ombrian relationships share the same parameters ξ and α. The tail index ξ determines the behaviour of the distribution tail and it is the most difficult to estimate. In the examples of Uccle (Digression 8.D) and Bologna (Digression 8.E) the available data sets were quite long and supported a reliable estimate of these parameter, which was found 0.194 for Uccle and 0.121 for Bologna. However, for short data records this is not possible and thus it is useful to refer to the global behaviour as revealed from the analyses of global data sets.

For years, the Gumbel distribution have been the prevailing models for rainfall extremes, which entails an exponential tail of the parent distribution, or $\xi = 0$. This is the smallest possible value for a distribution that is unbounded from above. Unjustified specification of ξ to its smallest possible value results in unsafe (too small) design rainfall values for large return periods. Recently, however, the appropriateness for rainfall of the exponential tail and the Gumbel distribution has been questioned. Koutsoyiannis (2004a, 2005a, 2007) discussed several theoretical reasons that favour the Pareto/EV2/Fréchet distribution over the exponential/EV1/Gumbel case. By now, several studies have provided empirical evidence supporting the Pareto case ($\xi > 0$). Some of them, based on empirical evidence from daily rainfall records worldwide, are explicitly mentioned below:

1. The data set compiled by Hershfield (1961) with 95 000 station-years, which he used to formulate his PMP method, in the study by Koutsoyiannis (1999), was found consistent with the EV2 distribution with shape parameter $\xi = 0.13$, or slightly varying with the average annual maximum rainfall $\overline{h}$ (in mm) as

$$\xi = 0.183 - 0.00049\,\overline{h} \tag{8.50}$$

2. Koutsoyiannis (2004b, 2005a) compiled an ensemble of annual maximum daily rainfall series from 169 stations in the Northern Hemisphere (28 from Europe and 141 from the USA) roughly belonging to six major climatic zones, all having lengths from 100 to 154 years, and comprising a total of 18 065 station-years. The analysis provided sufficient support for the general applicability of a positive tail index. Furthermore, the ensemble of all samples supported the estimation of a unique shape parameter ξ for all stations. The estimated value of ξ varied for different methods of estimation and was found $\xi = 0.09$ for the maximum likelihood method, $\xi = 0.10$ for the L-moments method, $\xi = 0.13$ for the method of moments and $\xi = 0.15$ for a weighted least squares method. The latter method, by assuming weights equal to the empirical quantiles, gives higher importance to the high values and, as the resulting value leads to more conservative design, the value $\xi = 0.15$ was suggested as the preferred one.
3. Papalexiou and Koutsoyiannis (2013) analysed the annual maximum daily rainfall of 15 137 records from the GHCN daily database, with lengths varying from 40 to 163 years. Using the L-moments method, they fitted to all stations the GEV distribution, which comprises all three cases of extreme value distributions. The results clearly suggested that the EV3 distribution (a distribution bounded from above, with negative tail index) is completely inappropriate for rainfall, while the EV2/Fréchet law ($\xi > 0$) prevails over the EV1/Gumbel law $\xi = 0$. The mean value of the shape parameter ξ for all stations was found to be 0.114. However, this value was not found to be representative for all parts of the world as there is variability. The statistical

sampling effect explains a big part of the observed variability of the shape parameter around its mean value $\xi = 0.114$, but not the total variability. The authors concluded that the geographical location on the globe may affect the value of the shape parameter and constructed a map of the geographical distribution of the GEV shape parameter, which shows that large areas of the world share approximately the same GEV shape parameter. As a final remark, the authors suggested not to follow blindly the statistical estimate of ξ based on whatever statistical method. In particular, they proposed that in the case where data suggest a negative tail index (distribution bounded from above), this should not be used. Instead it is more reasonable to use in this case a Gumbel or, for additional safety, a GEV distribution with a tail index value equal to 0.114.

4. Cavanaugh et al. (2015) analysed again a subset of the GHCN daily database, selecting over 22 000 high quality stations across the globe, which pass certain quality control and temporal completeness criteria. They utilized an advanced test for differentiating between exponential- and heavy-tailed distributions of precipitation, and their results indicated that the majority of precipitation exceedance probabilities are of Pareto type and, therefore, most precipitation records have Pareto tails, not exponential.

Additionally, Veneziano et al. (2009) used multifractal analysis to show that the annual rainfall maximum for time scale d can be approximated by a GEV distribution and that typical values of ξ lie in the range 0.09 to 0.15 with the larger values being associated with more arid climates. Similar results were provided by Chaouche (2001) and Chaouche et al. (2002). Chaouche (2001) exploited a data base of 200 rainfall series of various time steps (month, day, hour, minute) from the five continents, each including more than 100 years of data. Using multifractal analyses, it was found that (a) a Pareto/EV2 type law describes the rainfall amounts for large return periods; (b) the exponent of this law is scale invariant over scales greater than an hour (in fact, this is dictated by theoretical reasons; see Appendix 8-I); and (c) this exponent is almost space invariant. Other studies have also expressed scepticism for the appropriateness of the Gumbel distribution for the case of rainfall extremes. Coles et al. (2003) and Coles and Pericchi (2003) concluded that inference based on a Gumbel distribution model fitted to the annual maxima may result in unrealistically high return periods for certain observed events and suggested a number of modifications to standard methods, among which is the replacement of the Gumbel model with the GEV model. Mora et al. (2005) and Bacro and Chaouche (2006) confirmed that rainfall in Marseille (a raingauge included in the study by Koutsoyiannis, 2004b) and other raingages in southern France are not in the Gumbel law domain. Sisson et al. (2006) highlighted the fact that standard Gumbel analyses routinely assign near-zero probability to subsequently observed disasters, and that for San Juan, Puerto Rico, standard 100-year predicted rainfall estimates may be routinely underestimated by a factor of two. Schaefer et al. (2006) using the methodology by Hosking and Wallis (1997) for regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour time scales for the Washington State, USA, found that the distribution of rainfall maxima in this State generally follows the EV2 distribution type.

Digression 8.G: Area-reduction of point ombrian curves

The statistical analysis of rainfall extremes and the construction of ombrian relationships typically refer to a point (i.e., the raingauge station). On the other hand, in hydrology the transformation of rainfall to runoff occurs at the catchment scale and thus in engineering applications the rainfall intensity should refer to the catchment area. This should require additional statistical analyses for the areally averaged rainfall intensity. However, this is usually too difficult or impossible, because of the sparse network of raingauges as well as synchronization problems among the recordings of different devices. Therefore, a common method for a transformation of point estimates, to account for the spatiotemporal variability of rainfall across the river basin, suggests applying a reduction coefficient, called the *area-reduction factor* (or *areal reduction factor*, ARF).

The ARF is defined to be the ratio of the areally averaged precipitation depth over a certain area A for a specified return period T and time scale k to the precipitation depth over any point of the area (assumed to be climatically homogeneous) for the same return period and time scale. Accordingly, to find the ARF we need to determine the distribution functions of both areal and point rainfall and divide the two for several return periods and time scales. A prerequisite for this is to form statistical samples of areal rainfall with sufficient length and for various time scales. Another prerequisite for the definition to apply is the climatic homogeneity of the entire area, so that the same ombrian relationship applies to any point at the given area.

Some studies miss the above definition and determine the ARF empirically, e.g. by averaging precipitation per event and considering the ratio of maximum point precipitation (also known as the centre point precipitation) to the areal precipitation; this does not make much sense. In fact, empirical procedures like the latter imply different empirical definitions of ARF. A comprehensive review of empirical procedures and alternative definitions can be found in Svensson and Jones (2010). Despite theoretical inconsistencies, results from empirical studies of ARF have certainly some usefulness. Recent studies which adopt the consistent definition have been made by Lombardo et al. (2006) and Overeem et al. (2010). Both of these studies use radar data to estimate ARF, which certainly provide a great potential for studying the spatial variability of extreme precipitation due to the improved spatial coverage, resulting in good indications of the spatial patterns of rainfall. Major improvements in ARF estimation are anticipated in the near future, as radar and satellite data of rainfall will become more reliable and will accumulate in time providing samples with lengths adequate enough to enable reliable investigation of the probability distribution of areal rainfall. It is noted though that the poorer quality of these data, compared to raingauge data, is also expected to affect ARF estimation. Indeed, Allen and DeGaetano (2005) found that radar-based ARF decays at a faster rate (with increasing area) than gauge-based ARF.

Current literature typically gives ARF as a function of A and k, neglecting the effect of T, which is deemed small. Comprehensive investigations were carried out in the UK by NERC (1975) which provided tabulated values of ARF for a wide range of areas (1 to 30 000 km^2) and time scales (1 min to 25 days). Koutsoyiannis and Xanthopoulos (1999, p. 154) fitted the following empirical expression to those tabulated values:

$$\varphi = \max\left(0.25, 1 - \frac{0.048A^{0.36-0.01\ln A}}{k^{0.35}}\right) \tag{8.51}$$

where A is given in km^2 and k in h. The same relationship has been compared with nomographs by Hershfield and Wilson (1957) for the eastern USA and by the US Weather Bureau (1960) for the western USA; differences are visible but not very substantial and this supports applicability of equation (8.51) in other parts of the world.

Appendix 8-I: -Proof that the tail index of a time-averaged process is constant at any time scale

We assume that the stochastic variables $\underline{x}$ and $\underline{y}$ are nonnegative (if they are not, we truncate their distributions) and we let $\underline{z} \coloneqq \underline{x} + \underline{y}$. With the help of Figure 8.11 we can write:

$$P\{\underline{z} > z\} \geq P\{\underline{x} > z\}, \qquad P\{\underline{z} > z\} \geq P\left\{\underline{y} > z\right\} \tag{8.52}$$

and

$$P\{\underline{z} \leq z\} \geq P\left\{\underline{x} \leq z/2, \underline{y} \leq z/2\right\} = P\left\{\underline{x} \leq z/2\}|\underline{y} \leq z/2\right\} P\left\{\underline{y} \leq z/2\right\} \tag{8.53}$$

For independent and positively dependent $\underline{x}$ and $\underline{y}$ we have $P\left\{\underline{x} \leq z/2\}|\underline{y} \leq z/2\right\} \geq P\{\underline{x} \leq z/2\}$. Thus, $P\{\underline{z} \leq z\} \geq P\left\{\underline{x} \leq z/2, \underline{y} \leq z/2\right\} \geq P\{\underline{x} \leq z/2\}P\left\{\underline{y} \leq z/2\right\}$ and consequently:

$$\begin{aligned} P\{\underline{z} > z\} &\le P\{\underline{x} > z/2\} + P\left\{\underline{y} > z/2\right\} - P\{\underline{x} > z/2\}P\left\{\underline{y} > z/2\right\} \\ &\le P\{\underline{x} > z/2\} + P\left\{\underline{y} > z/2\right\} \end{aligned} \tag{8.54}$$

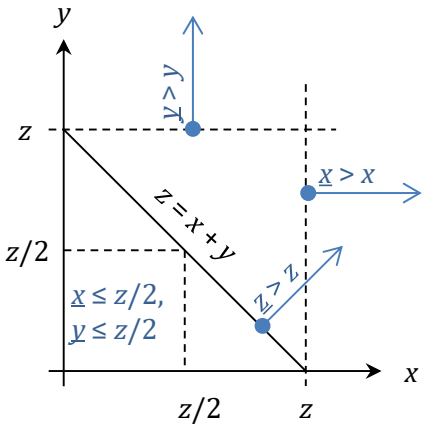

Figure 8.11 Auxiliary sketch of the proof of the constancy of tail index.

As a result:

$$\max\left(\overline{F}_x(z), \overline{F}_y(z)\right) \le \overline{F}_z(z) \le \overline{F}_x(z/2) + \overline{F}_y(z/2) \tag{8.55}$$

Multiplying by $z^{1/\xi}$ for $\xi > 0$ and taking limits we obtain

$$\lim_{z\to\infty} z^{\nu} \max\left(\overline{F}_x(z), \overline{F}_y(z)\right) \le \lim_{z\to\infty} z^{\nu}\overline{F}_z(z) \le \lim_{z\to\infty} z^{\nu}\overline{F}_x(z/2) + \lim_{z\to\infty} z^{\nu}\overline{F}_y(z/2) \tag{8.56}$$

Let ξ_x and ξ_y be the tail indices of $\underline{x}$ and $\underline{y}$, respectively, and assume $\xi_x \ge \xi_y$ (if not, we interchange x and y and have the same results). According to the definition of tail index (equation (2.63)), this means that:

$$\lim_{x\to\infty} x^{1/\xi_x}\bar{F}_x(x) = l_x, \qquad \lim_{y\to\infty} y^{1/\xi_y}\bar{F}_y(y) = l_y \tag{8.57}$$

where l_x and l_y are finite. At the same time it means that $\lim_{x\to\infty} x^{1/\xi}\overline{F}_x(x) = 0$ for any $\xi > \xi_x$ and $\lim_{x\to\infty} x^{1/\xi}\overline{F}_x(x) = \infty$ for any $\xi < \xi_x$, and likewise for y. If we assume $\xi_x > \xi_y$ and take $\xi = \xi_x$, the rightmost part of (8.56) becomes:

$$\begin{aligned} \lim_{z\to\infty} z^{1/\xi_x}\overline{F}_x(z/2) + \lim_{z\to\infty} z^{1/\xi_x}\overline{F}_y(z/2) &= \lim_{z\to\infty}(2z')^{1/\xi_x}\overline{F}_x(z') + \lim_{z\to\infty}(2z')^{1/\xi_x}\overline{F}_y(z') \\ &= 2^{1/\xi_x} l_x + 0 = 2^{1/\xi_x} l_x \end{aligned} \tag{8.58}$$

and the leftmost part is $\lim_{z\to\infty} \max\left(z^{1/\xi_x}\overline{F}_x(z), z^{1/\xi_x}\overline{F}_y(z)\right) = \max(l_x, 0) = l_x$. Thus, (8.56) becomes:

$$l_x \le \lim_{z\to\infty} z^{1/\xi_x}\overline{F}_z(z) \le 2^{1/\xi_x}\, l_x \tag{8.59}$$

Furthermore, for any $\xi < \xi_x$ the leftmost quantity in (8.56) is ∞ and thus $\lim_{z\to\infty} z^{1/\xi}\overline{F}_z(z) = \infty$. Also for any $\xi > \xi_x$ the rightmost quantity in (8.56) is 0 and thus $\lim_{z\to\infty} z^{1/\xi}\overline{F}_z(z) = 0$. If $\xi_x = \xi_y$ the above results are valid except that the rightmost part of (8.59) becomes $2^{1/\xi_x+1} l_x$. Summarizing these results, we have:

$$\lim_{z\to\infty} z^{1/\xi}\overline{F}_z(z) = \begin{cases} 0, & \xi > \xi_x \\ l_z & \xi = \xi_x \\ \infty & \xi < \xi_x \end{cases} \tag{8.60}$$

where l_z is a finite number satisfying $l_x \le l_z \le 2^{1/\xi_x+1} l_x$. This proves that the tail index of the distribution of the sum of two variables equals the maximum of the tail indices of the two variables. This result is readily expanded for many variables. Consequently, for a stationary stochastic process, the tail index is preserved in the cumulative process, which is a sum of many variables with same tail index, and hence of the averaged process.

Appendix 8-II: Relationships of climacogram and parameters of the ombrian model

By standard algebra on equation (8.6) we find that the pth moment of $\underline{x}^{(k)}$ is:

$$\mathrm{E}\left[\left(\underline{x}^{(k)}\right)^p\right] = \mu_p' = \frac{P_1^{(k)}\left(\lambda(k)\right)^p p}{\zeta(k)\xi^{p/\zeta(k)}}\mathrm{B}\left(\frac{p}{\zeta(k)}, \frac{1}{\xi} - \frac{p}{\zeta(k)}\right) \tag{8.61}$$

Hence, the mean is

$$\mathrm{E}\left[\underline{x}^{(k)}\right] = \mu = \frac{P_1^{(k)}\lambda(k)}{\zeta(k)\xi^{1/\zeta(k)}}\mathrm{B}\left(\frac{1}{\zeta(k)}, \frac{1}{\xi} - \frac{1}{\zeta(k)}\right) \tag{8.62}$$

and the squared coefficient of variation is:

$$C_{\mathrm{v}}^2\left[\underline{x}^{(k)}\right] = \frac{\gamma(k)}{\mu^2} = \frac{2\mathrm{B}\left(\frac{2}{\zeta(k)}, \frac{1}{\xi} - \frac{2}{\zeta(k)}\right)}{P_1^{(k)}\zeta(k)\,\mathrm{B}\left(\frac{1}{\zeta(k)}, \frac{1}{\xi} - \frac{1}{\zeta(k)}\right)^2} - 1 \tag{8.63}$$

For the special case that $\zeta(k) = 1$ (equation (8.5); Pareto), the mean and squared coefficient of variation simplify to:

$$\mathrm{E}\left[\underline{x}^{(k)}\right] = \mu = \frac{P_1^{(k)}\lambda(k)}{1-\xi}, \qquad C_{\mathrm{v}}^2\left[\underline{x}^{(k)}\right] = \frac{\gamma(k)}{\mu^2} = \frac{2(1-\xi)}{(1-2\xi)P_1^{(k)}} - 1 \tag{8.64}$$

Combining these with equations (8.9) and (8.15) we find:

$$\frac{\gamma_0(1 + (k_{\max}^*/\alpha)^{2M})^{\frac{H-1}{M}}}{\mu^2} = \frac{1}{1-2\xi} \Leftrightarrow k_{\max}^* = \alpha\left(\left(\frac{(1-2\xi)\gamma_0}{\mu^2}\right)^{\frac{M}{1-H}} - 1\right)^{\frac{1}{2M}} \tag{8.65}$$

However, for the general case, equations (8.62) and (8.63) are implicit for ξ and $\zeta(k)$, and too complicated for our purposes. Therefore, we look for simplifying approximations. For an approximation of (8.62) it can be seen that the part of the right-hand side that contains ξ and $\zeta(k)$ is equal to $1/(1-\xi)$ for $\zeta(k) = 1$ and tends to 1 as $\zeta(k) \to \infty$. A numerical investigation showed that a very good approximation with these properties is the following:

$$\mu \approx P_1^{(k)}\lambda(k)\left(1 + \frac{1}{(1-\xi)\left(\zeta(k)\right)^2} - \frac{1}{\left(\zeta(k)\right)^{\sqrt{2}}}\right) \tag{8.66}$$

For an approximation of (8.63), we initially examine the case $P_1^{(k)} = 1$, for which:

$$C_{\mathrm{v}}^2\left[\underline{x}^{(k)}\right] = \frac{2\mathrm{B}\left(\frac{2}{\zeta(k)}, \frac{1}{\xi} - \frac{2}{\zeta(k)}\right)}{\zeta(k)\,\mathrm{B}\left(\frac{1}{\zeta(k)}, \frac{1}{\xi} - \frac{1}{\zeta(k)}\right)^2} - 1 =: C\big(\zeta(k)\big) \tag{8.67}$$

It is easily shown that for $\zeta(k) = 1$, $C\big(\zeta(k)\big)$ equals $1/(1-2\xi)$. Furthermore, it can be shown analytically (the proof is omitted) that, as $\zeta(k) \to \infty$ (which happens when $k \to \infty$), the LLD of C as a function of ζ is $C^{\#}\big(\zeta(k)\big) = -2$. Also, numerical investigation shows that the slope of -2 is virtually constant for the entire domain of $\zeta \geq 1$. This observation, combined with the value of $C_{\mathrm{v}}^2\left[\underline{x}^{(k^*_{\max})}\right]$, enables the very simple approximation:

$$C\big(\zeta(k)\big) \approx \frac{1}{(1-2\xi)(\zeta(k))^2} \tag{8.68}$$

Hence, in the general case:

$$C_{\mathrm{v}}^2\left[\underline{x}^{(k)}\right] = \frac{\gamma(k)}{\mu^2} = \frac{C\big(\zeta(k)\big) + 1}{P_1^{(k)}} - 1 \tag{8.69}$$

which yields:

$$\frac{\gamma(k)}{\mu^2} = \frac{1}{P_1^{(k)}(1-2\xi)(\zeta(k))^2} + \frac{1 - P_1^{(k)}}{P_1^{(k)}} \tag{8.70}$$

Once $p_0^{(k)}, \gamma(k)$ and μ are known, the unknown $\zeta(k)$ and $\lambda(k)$ are easily found to be given by equations (8.11) and (8.12).

Chapter 9. Streamflow maxima and minima

9.1 Streamflow extremes compared to rainfall extremes

While rainfall databases have been publicly available for a few decades, and this enabled the study of rainfall extremes and extraction of generalized results over the globe, streamflow databases with publicly available data are a more recent—and partial—development. Therefore, general results have not been obtained yet. The study of streamflow involves many more difficulties than that of rainfall. The measurement of streamflow is a demanding task and needs sophisticated equipment and analyses, and observational experience. In addition, what we measure does not necessarily represent the natural streamflow process. Several large-scale control structures on rivers, such as dams, levees, intakes and diversions, have seriously modified the natural process. To reconstruct the natural regime from the measurements needs appropriate knowledge of the modified hydrosystem and its control, and simultaneous processing of several data sets. The so derived reconstructed time series, usually called "naturalized", are rare and even more rarely have become available online as yet. To understand the natural regime, it is much easier to analyse data from pristine catchments.

In this chapter we provide several representative examples using long time series from the database of the US Geological Survey's (USGS) National Water Information System. From this database, Hirsch and Ryberg (2012) selected 200 stream gauges in the coterminous USA, of at least 85 years length through water year 2008, from basins with little or no reservoir storage or urban development (less than 150 persons per km^2 in 2000). These stations are an ideal source of the examples given here. The data retrieved[*] extend up to 2020 and are free of missing values (some "provisional" values for the most recent months were not used in the analyses). In Europe, a set of 224 stream gauges was studied in Iliopoulou et al. (2019) but in this case the time series are shorter; thus, only the longest of them (Po river with 90 years of observations, Montanari, 2012[†]) is analysed.

As we will see in the examples and contrary to rainfall, where at the lowest scales the Pareto distribution, with lower-tail index $\zeta = 1$, is appropriate, the streamflow often exhibits a higher value of ζ even at small scales; this implies a bell-shaped density function. Therefore, a candidate distribution for streamflow is the PBF. While the lower-tail index ζ is most important for the low extremes, the higher-tail index ξ is most important for the high extremes. The relationship of ξ in streamflow with that of rainfall is discussed in Digression 9.A.

Intermittence is a prominent behaviour in rainfall, which has been modelled in Chapter 8 through the probability wet (P_1) or dry (P_0). This can also be the case in streamflow but only in small streams which often, during the summer months, become dry. Large rivers

---

[*] Data retrieved on 2020-08-22 from https://nwis.waterdata.usgs.gov/nwis/inventory; the discharge values were converted from ft^3/s to m^3/s.

[†] Data made available by Alberto Montanari, https://www.albertomontanari.it/sites/default/files/uploadedfiles/po-pontelagoscuro.txt, retrieved on 2020-08-22. The data set is affected by the Italian tactic to remove the values of 29 February at leap years.

are very rarely dry, yet intermittence appears in a different mode. Specifically, there are two states, one in which the river is fed merely by groundwater (baseflow) and one dominated by flood. As we will see, a simple technique to model this type of intermittence is to set a positive lower limit to the PBF distribution.

As discussed in Chapter 8, in rainfall it is most often necessary to model extremes at multiple scales—whence the need to construct ombrian relationships. In streamflow this is not usually the case. Instead, often there is a need to construct an operational stochastic simulation model. Several tasks can only be studied by means of stochastic simulation. For example, in studying design floods of dams, it does not suffice to determine the value of river discharge for the design return period. Rather, we should determine the value of the outflow discharge of the dam spillway. In terms of low events, again it does not suffice to determine the river discharge for a design return period or reliability. Rather, we need to establish a relationship between reservoir policy (e.g. reliable yield of the reservoir) and probability of emptying of reservoir.

All these tasks are easily dealt with by stochastic simulation. A theoretical description of the related concepts can be found in Koutsoyiannis and Economou (2003) while a simplified application of a simulation methodology for the design of a reservoir spillway can be found in Koutsoyiannis (1994). Here we will not give detailed applications of simulation, as this is not the focus of this text. However, all concepts necessary for simulation are contained in Chapter 7. We note that in a simulation focusing on minima (droughts), e.g. in determining the reliable reservoir yield, a monthly time step usually suffices and the preservation of seasonality by means of a cyclostationary model becomes important (e.g. Koutsoyiannis, 2000, 2001). Instead, in a simulation focusing on maxima (floods), e.g. in determining a spillway's design discharge, a subdaily simulation step is necessary and, in this case, it becomes important to preserve the time irreversibility of the streamflow (see section 7.5 and Koutsoyiannis 2019b, 2020a), which however becomes negligible on the monthly time scale.

While a stochastic model is constructed for the specific time scale of interest (subdaily, daily, monthly or even annual, depending on the particulars of the application) a study of some multi-scale characteristics of the process is necessary (e.g. to characterize the time dependence and possibly the persistence). This is most effectively done through the process climacogram. In some applications (particularly in studies of large hydrosystems with many reservoirs) stochastic models at two or more time scales are often involved. It becomes thus imperative to make the models consistent to each other. This is done by techniques called model coupling or disaggregation (Koutsoyiannis and Manetas, 1996; Koutsoyiannis, 2001).

Digression 9.A: Does the tail index of streamflow differ from that of rainfall?

The tail index is particularly important for characterizing the extraordinary extreme values (e.g. of return periods of the order 1000 years). Underestimation of the tail index may markedly underestimate the value of such events in the design phase or overestimate the return period of extraordinary events that have occurred. As streamflow results from rainfall, it is interesting to examine whether or not the transformation of rainfall to runoff preserves the tail index. Here we

provide some hints about what we may expect, rather than a consistent analysis of the problem, which would not be easy.

In hydrology, we usually model this transformation within a catchment using the conceptual analogy of one or more connected reservoirs. Here we consider just one reservoir, which could be one segment in the catchment, with input $I(t)$ the upstream flow discharge and output $Q(t)$ the downstream flow discharge; alternatively, it could represent the entire catchment with the inflow being the precipitation. The outflow is assumed to be the reservoir spill. We wish to see whether or not the transformation of input to output preserves the tail index.

Inflow and outflow are connected by the continuity equation (conservation of mass), i.e.:

$$\frac{\mathrm{d}S}{\mathrm{d}t} + Q = I \tag{9.1}$$

where S is the reservoir storage. We need one more equation to fully describe the transformation of I to Q which we construct by means of a stage-discharge and a stage-storage relationships. We assume these to be of power type:

$$Q = k\, z^l, \qquad S = S_0 + m\, z^n \tag{9.2}$$

where z is the water elevation and k, l, m, n, S_0 are constants. From these we get $z = (Q/k)^{1/l}$and $S = S_0 + m(Q/k)^{n/l} = S_0 + \alpha Q^\beta$, where $\beta \coloneqq n/l$ and $\alpha \coloneqq m/k^{n/l}$. Eliminating S from the differential equation we obtain:

$$\alpha\beta Q^{\beta-1}\frac{\mathrm{d}Q}{\mathrm{d}t} + Q = I \tag{9.3}$$

The value of the exponent $\beta - 1$ determins the behaviour of Q. To get an idea of what the value of the exponent could be, we recall from hydraulics that a typical value of l in spillways is $3/2$. The exponent n is determined by the topography of the reservoir. For a prismatic reservoir, $n = 1$, while for a pyramidal or conic one $n = 3$ and hence β becomes $2/3$ and 2 for these two cases, respectively. Therefore, the exponent $\beta - 1$ would be –1/3 and 1, respectively. The parameter α (whose dimension is such that $\alpha Q^{\beta-1}$ has dimension of time) is also useful to interpret. A large inundation plain would imply a large α to express the fact that in a large plain Q would be less sensitive to S (as for, say, $\beta = 1, \Delta Q = \Delta S/\alpha$).

The differential equation (9.3) admits a closed solution only if $\beta - 1 = 0$. This case is indeed reasonable according to the above discussion and results in a first-order linear differential equation with general solution (see section 3.11):

$$Q(t) = Q_0 \mathrm{e}^{-t/\alpha} + \frac{1}{\alpha}\int_0^t \mathrm{e}^{-(t-s)/\alpha} I(s)\,\mathrm{d}s \tag{9.4}$$

where $Q_0 = Q(0)$ (the initial condition). This has been very popular in hydrology and is known as the linear catchment model. Evidently, if we multiply the input by a constant A, the output will also be multiplied by the same constant. Furthermore, treating the input as a stationary stochastic process $\underline{I}(t)$ and taking a large t, so that $e^{-t/\alpha} \approx 0$, we see that the output is also a stationary stochastic process. Now, temporarily assuming that the inflow is independent in time, we see from equation (9.4) that if for some q the moment $\mathrm{E}[\underline{I}(t)^q]$ diverges to infinity, then $\mathrm{E}[\underline{Q}(t)^q]$ will diverge too; also if $\mathrm{E}[\underline{I}(t)^q]$ is finite, then $\mathrm{E}[\underline{Q}(t)^q]$ will be finite too. Therefore, as the inverse of the tail index, $1/\xi$, is the threshold value determining whether or not the moments diverge ($q \geq 1/\xi$) or are finite ($q < 1/\xi$), we infer that if the tail index of $\underline{I}(t)$ is ξ, then that of $\underline{Q}(t)$ will also be ξ. This result can be extended to $\underline{I}(t)$ dependent in time, taking into account that again any $\underline{I}(t)$ is linearly equivalent to white noise (cf. Wold's decomposition; Digression 3.E).

However, the behaviour changes if $\beta \neq 1$ as the differential equation (9.3) turns to nonlinear. To illustrate the behaviour in this case, as there is no closed solution, we consider as an example an input segment (surge) with mathematical form $I(t) = Ate^{-t}$, whose total volume is A. We solve the equation numerically to find $Q(t)$ for several values of A, and investigate the ratio of the peaks of outflow to inflow, as the peak flows are the most representative for the behaviour in the distribution tails.

Some results have been plotted in Figure 9.1 for several values of the exponent β and the scale parameter α of the storage-discharge relationship $S = S_0 + \alpha Q^\beta$. It can be seen that, while in the linear case the ratio Q/I of peak values is constant (as expected based on the discussion of the linear case), for $\beta \neq 1$ the results roughly support a relationship of power type, $Q/I \propto I^\gamma$, where $\gamma \neq 0$ is the slope in the doubly logarithmic plot of Figure 9.1. Thus, $Q^{1/(1+\gamma)} \propto I$. If ξ is the tail index of inflow, which means that $\mathrm{E}\left[\underline{I}(t)^{1/\xi}\right] = \infty$, then $\mathrm{E}\left[\underline{Q}(t)^{1/(1+\gamma)\xi}\right] = \infty$. Hence, we conclude that the tail index of the outflow will be $\xi_Q = (1+\gamma)\xi$. As γ (i.e. the slope of Figure 9.1) can be either positive, zero or negative, ξ_Q will be either greater than, equal to, or smaller than ξ. In particular, the presence of large plains within the catchment will signify a decrease of ξ.

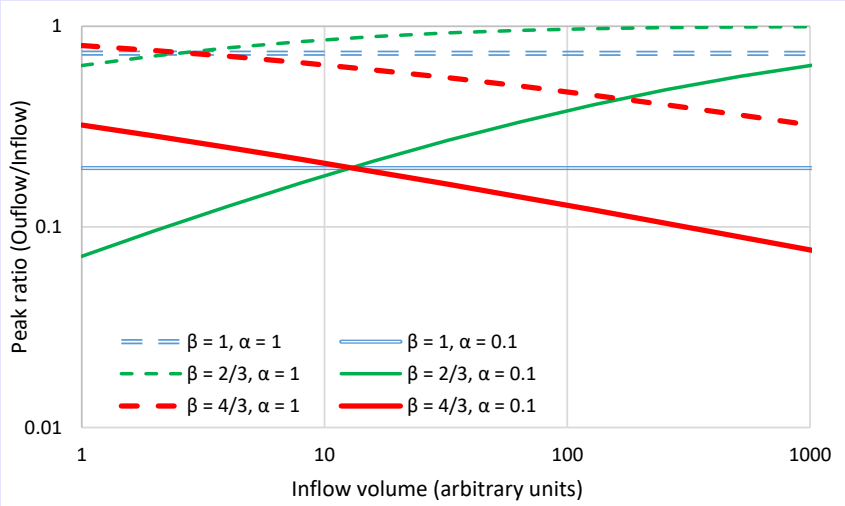

Figure 9.1 Ratio of peak outflow to peak inflow in a reservoir with inflow $I(t) = Ate^{-t}$, as a function of the inflow total volume of A, for several values of the exponent β and the scale parameter α of the storage-discharge relationship $S = (Q/\alpha)^\beta$. Values of $\beta = 1$, < 1 and > 1 correspond to a linear reservoir, a reservoir with roughly prismatic shape, and one with roughly pyramidal shape, respectively. A large α corresponds to a large reservoir area (e.g. the inundation of a plain).

Of course, the whole runoff process on the entire catchment includes other types of routing in addition to that across the flow in a river. Some of them, like snowmelt, tend to increase the tail index, while other, such as retention and infiltration, tend to decrease it.

In conclusion, only in catchments in which an assumption linearity ($\beta = 1$) is justified, we can expect a tail index of streamflow equal to that of rainfall. In large catchments which include large flood plains it would not be a surprise if we estimated a tail index equal to zero (a light-tailed distribution) even if the tail index of rainfall is positive.

9.2 PBF distribution fitting on streamflow

The extreme-oriented fitting of probability distributions has been already discussed in section 6.15, and implemented and further investigated in Digression 6.E. Here we will give a more general algorithm which can be used for fitting with emphasis either on high extremes or low extremes, or even on the body of the distribution. The algorithm uses noncentral K-moments of orders $(p, 1)$ (and for low extremes tail K-moments), with p

being within the feasible range $(1, n)$, where n is the sample size. The steps of the algorithm are the following.

1. We choose a number $m + 1$ of moment orders, i.e., $p_i = n^{i/m}$, $i = 0, \dots, m$, with $p_0 = 1, p_m = n$. While, when dealing, for instance, with daily flows, the sample size n is usually of the order of several thousands, the number m could be chosen much smaller, e.g. of the order of 100, to speed up calculations without compromising accuracy. The orders p_i need not be natural numbers.
2. We estimate the noncentral K-moments K'_{p_i} of orders p_i (and for low extremes the tail K-moments $\overline{K}'_{p_i}$) using equations (6.65) (or (6.71) for tail moments) and (6.67).
3. We construct the climacogram of the time series and estimate the Hurst parameter. If the Hurst parameter is large, say $H \geq 0.8$, we replace in the following steps the orders p_i with p'_i, where the latter are given by equations (6.93) and (6.95).
4. Using default values of the parameters ζ and ξ we estimate the Λ-coefficients Λ_1 and Λ_∞ (or $\overline{\Lambda}_1$ and $\overline{\Lambda}_\infty$ for tail moments); furthermore, for all p_i (or p'_i if they are different) we estimate the empirical return period $\hat{T}\left(K'_{p_i}\right)$ from equation (6.112) (or $\hat{\overline{T}}\left(\overline{K}'_{p_i}\right)$ from equation (6.123)).
5. Assuming default values of the scale parameter λ and of the lower bound x_{L}, we estimate the theoretical return period of each noncentral (or tail) K-moment as:

$$\frac{T\left(K'_{p_i}\right)}{D} = \left(1 + \zeta\xi\left(\frac{x - x_{\mathrm{L}}}{\lambda}\right)^{\zeta}\right)^{\frac{1}{\xi\zeta}}, \qquad \frac{\overline{T}\left(\overline{K}'_{p_i}\right)}{D} = \frac{1}{1 - \left(1 + \zeta\xi\left(\frac{x - x_{\mathrm{L}}}{\lambda}\right)^{\zeta}\right)^{-\frac{1}{\xi\zeta}}} \tag{9.5}$$

6. We form an expression for the total fitting error as the sum of the logarithmic deviations of empirical and theoretical return periods, i.e.:

$$\begin{aligned} e(\zeta, \xi, \lambda, x_{\mathrm{L}}) &:= \sum_i w_T \left(\ln\left(\hat{T}\left(K'_{p_i}\right)\right) - \ln\left(T\left(K'_{p_i}\right)\right)\right)^2, \\ \overline{e}(\zeta, \xi, \lambda, x_{\mathrm{L}}) &:= \sum_i \overline{w}_T \left(\ln\left(\hat{\overline{T}}\left(\overline{K}'_{p_i}\right)\right) - \ln\left(\overline{T}\left(\overline{K}'_{p_i}\right)\right)\right)^2 \end{aligned} \tag{9.6}$$

   where w_T and $\overline{w}_T$ denote weighting coefficients. These errors are functions of the chosen parameters $(\zeta, \xi, \lambda, x_{\mathrm{L}})$ and we evaluate them for the chosen parameter set.
7. We repeat the calculations of steps 4-6 for different sets of parameter $(\zeta, \xi, \lambda, x_{\mathrm{L}})$ until the fitting error becomes minimal.

The repetitions of steps 4-6 are executed by a solver (such as those in spreadsheet software) using as objective function (to be minimized) the error $e(\zeta, \xi, \lambda, x_{\mathrm{L}})$ (or $\overline{e}(\zeta, \xi, \lambda, x_{\mathrm{L}})$ in the case of tail moments). The procedure is typically very fast, almost immediate.

A default value of the weights is $w_T = \overline{w}_T = 1$. However, if we wish to focus the fitting on a particular part of distribution, we can use different weight; for example, if we wish to neglect all values with return periods smaller than time d (e.g. $d = 1$ year) we may use:

$$w_T = \begin{cases} 0, & x < d \\ 1, & x \geq d \end{cases} \tag{9.7}$$

If we wish to give more emphasis on fitting to high return periods, we may use

$$w_T = \left(\hat{T}\left(K'_{p_i}\right)/d\right)^b \tag{9.8}$$

where b is a positive number, e.g. $b = 0.5$. Likewise for the weights $\overline{w}$.

Obviously, instead of the approximations of the Λ-coefficients by equations (6.93) and (6.95), the more accurate nonlinear approximations of section 6.14 could be used. However, as we will see in the applications that follow, the difference in the fitting is negligible.

A first application of the methodology is given in Digression 9.B while additional applications will be seen in subsequent sections.

Digression 9.B: Fitting of a single PBF distribution on the entire domain

We have seen in Digression 6.E that the PBF distribution with a single parameter set provided a good fit for a 206-year long record of daily rainfall in Bologna. Now we investigate whether this is feasible for streamflow data. We use as an example the streamflow data of the French Broad River at Asheville, NC, USA (USGS station 03451500, 35.609°N, 82.578°W, drainage area 2 447.5 km²). The data cover the period from October 1895 to March 2020 (more than 124 calendar years, uninterrupted; daily values $n = 45\,468$).

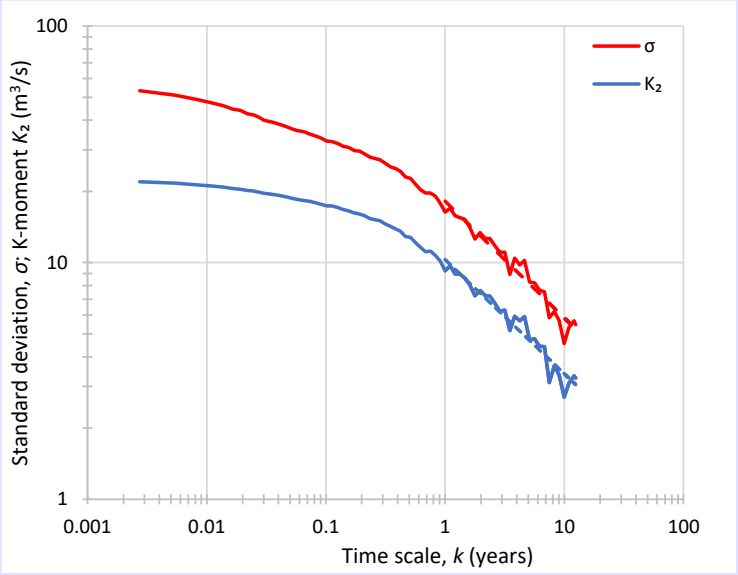

Figure 9.2 Empirical σ-climacogram and K_2-climacogram of daily data of the French Broad River at Asheville, NC, USA. The power-laws, fitted by regression and plotted as dashed lines, have slopes –0.49 and –0.48, respectively. The Hurst parameter, estimated from the annual series. is $H = 0.58$.

The climacogram of the time series is shown in Figure 9.2. Namely we give the σ-climacogram, which is the (doubly logarithmic) plot of the standard deviation, σ, vs. the time scale, k. In addition, we show in the same graph the K_2-climacogram, which is the plot of the central K-moment of orders (2,1), K_2, vs. the time scale, k. Both quantities have units of m³/s and their plots become virtually parallel straight lines for scales $k > 1$ year; their slope is $H - 1$ where H is the Hurst parameter. Here the Hurst parameter is estimated by the algorithm by Koutsoyiannis (2003) at $H = 0.58$. The relatively small value of H suggests that the autocorrelation effect can be neglected in estimating K-moments, which do not need any adaptation.

The K-moments are estimated for $m + 1$ moment orders with $m = 80$, so that $p_i := n^{i/m}$ ($p_0 = 1, p_{80} = n = 45\,468$) using the unbiased estimators. Both noncentral and tail moments are estimated. In a following step, for each K'_{p_i} or $\overline{K}'_{p_i}$ the return period of each $T\left(K'_{p_i}\right)$ or $\overline{T}\left(\overline{K}'_{p_i}\right)$ is estimated through the Λ-coefficients using both the linear and nonlinear approximations.

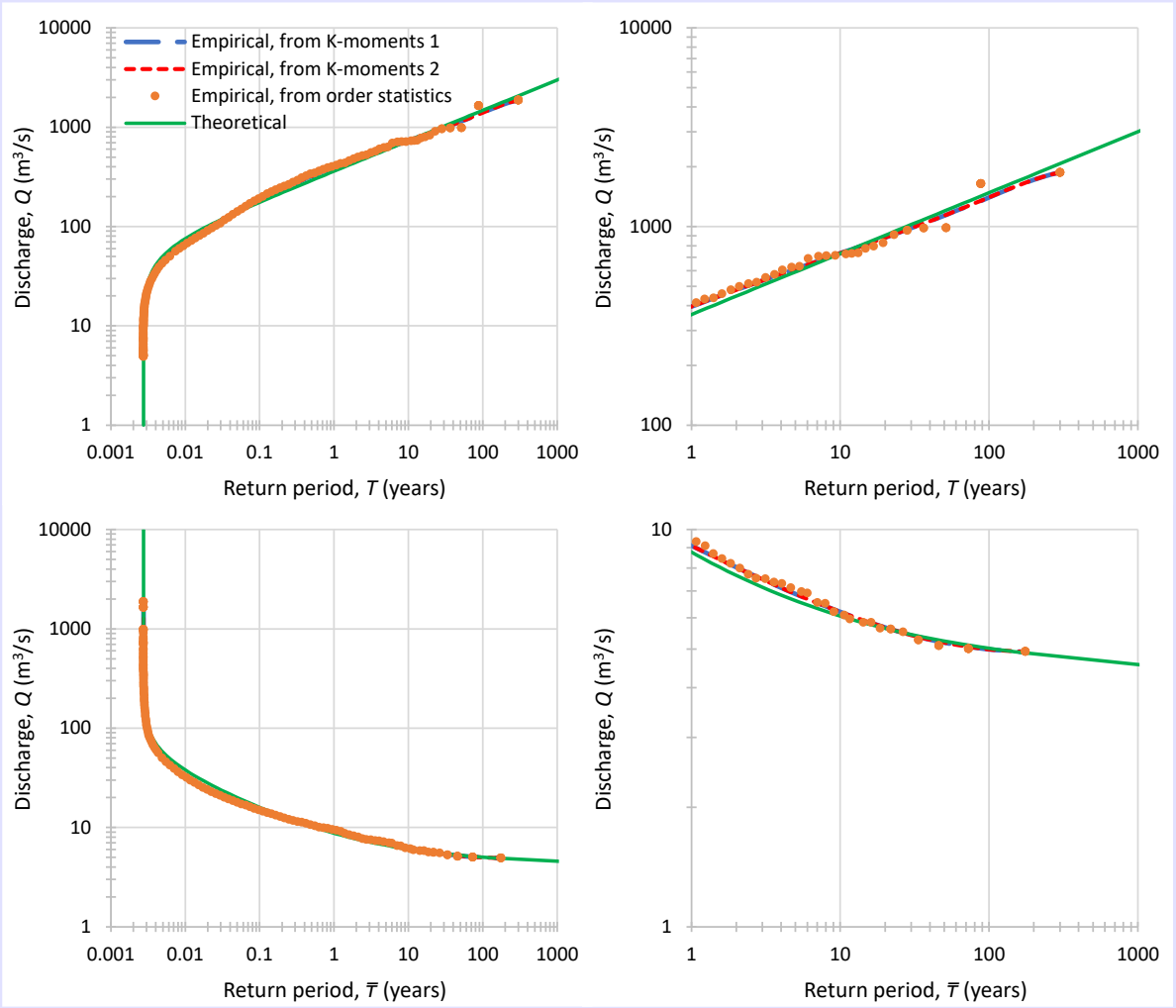

Figure 9.3 Comparison of empirical and theoretical PBF distribution fitted on daily data of the French Broad River at Asheville, NC, USA. The fitting was done based on noncentral and tail K-moment using the linear approximation (denoted in the figure "K-moments 1"). For comparison the nonlinear approximation ("K-moments 2") is also shown but it is indistinguishable from the former case. Furthermore, empirical return periods assigned from order statistics are also shown. The distribution function is depicted in terms of return periods T or $\overline{T}$ (**upper** and **lower** row, respectively), either on their entire range **(left column)** or focusing on those greater than 1 year **(right column)**. The parameters are: $\xi = 0.307, \zeta = 2.40, \lambda = 51.4\ \mathrm{m}^3/\mathrm{s}, x_L = 4.38\ \mathrm{m}^3/\mathrm{s}$.

Figure 9.3 shows the empirical estimates of these K-moments against their return periods. Both K'_{p_i} and $\overline{K}'_{p_i}$ have been used for each of the plots; note that given the return period with reference to minima $\overline{T}\left(\overline{K}'_{p_i}\right)$, the return period with reference to maxima is:

$$\frac{T\left(\overline{K}'_{p_i}\right)}{D} = \frac{1}{\left(1 - D/\overline{T}\left(\overline{K}'_{p_i}\right)\right)} \tag{9.9}$$

and thus, we use $2m + 1 = 161$ points in each of the plots. Plots are given in terms of both T and $\overline{T}$ in Figure 9.3.

In all plots, the series of points calculated by the linear and the nonlinear approximations of the return periods of K-moment fall one over the other and are not distinguishable. In addition, the empirical return periods based on order statistics are also plotted and agree very well with the K-moments plots, except a scatter for the largest return periods.

The fitting of the PBF distribution was made with the algorithm described in section 9.2 minimizing the sum $e + \overline{e}$ with default weights $(= 1)$. The parameters of the distribution are shown in the caption of Figure 9.3. The theoretical PBF distribution fitted is also indistinguishable from the curves of the K-moments. Overall, the plot shows a satisfactory global fit of a single PBF distribution on return periods spanning 5 orders of magnitude.

However, if we focus on the high return periods, as seen on the right panels of Figure 9.3, we will conclude that this global fit is not perfect. We may thus decide to perform two different fittings, with different parameter sets, separately for the low and the high flows. To do this it suffices to use the weights of equation (9.7). The resulting plots are shown in Figure 9.4, with the parameter values given in the figure caption. Now the two fittings are perfect.

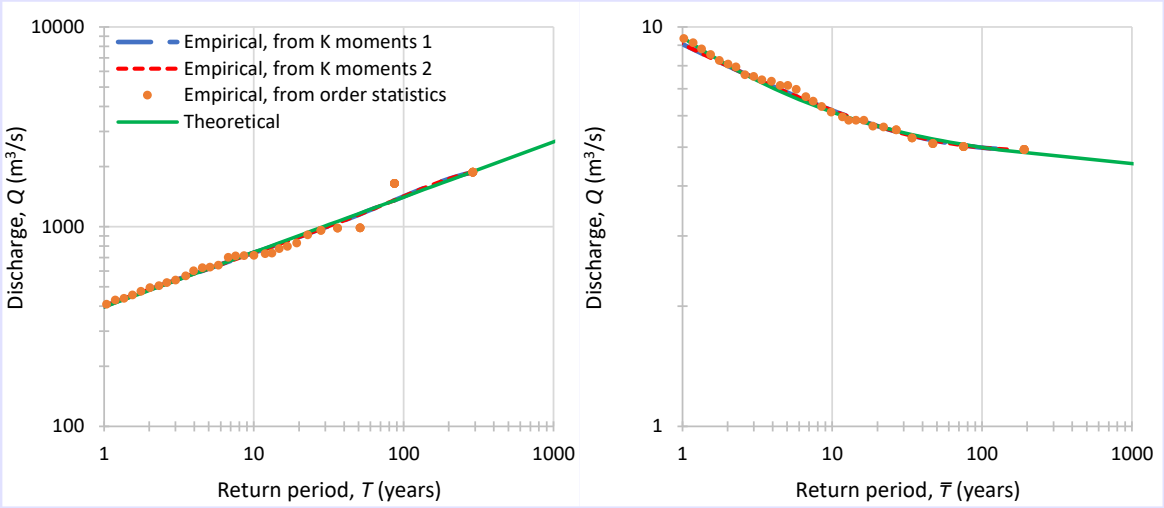

Figure 9.4 Comparison of empirical and theoretical PBF distribution fitted on daily data of the French Broad River at Asheville, NC, USA, as in Figure 9.3 but with a fitting focusing on $T > 1$ year **(left)** or $\overline{T} > 1$ year **(right)**. The parameters are, for the left panel: $\xi = 0.277, \zeta = 5.06, \lambda = 81.7$ m^3/s, $x_L = 4.87$ m^3/s, and for the right panel: $\xi = 0, \zeta = 2.16, \lambda = 77.3$ m^3/s, $x_L = 4.40$ m^3/s.

9.3 Fitting a distribution on the distribution body

While the fitting methodology described in section 9.2 is extreme-oriented and thus good for the distribution tails, it is also quite general and can easily provide a fit to the body of the distribution. The easiest way to do that is by using only low-order moments. As the PBF distribution, in the form used in this chapter, contains four parameters, a simple technique is to fit so that the first four theoretical K-moments match the corresponding empirical ones. This can be done applying again the framework of section 9.2 but with weights:

$$w\left(\hat{T}\left(K'_{p_i}\right)\right) = \begin{cases} 1, & p_i = 1,2,3,4 \\ 0, & \text{otherwise} \end{cases} \tag{9.10}$$

The advantage of using this framework, in addition to its generality, is that we bypass the theoretical calculation of the K-moments per se, in essence replacing it with the

calculation of Λ-coefficients which are easily and accurately approximated. An illustration is given in Digression 9.C.

Digression 9.C: An example of fitting a PBF distribution on its body

Here we use another example, the streamflow data of the Susquehanna River at Danville, PA, USA (USGS station 01540500, 40.958°N, 76.619°W, drainage area 29 059.7 km²). The data cover the period October 1905 to June 2020 (more than 114 calendar years, uninterrupted; daily values $n = 42\,081$). The σ-climacogram and the K_2-climacogram of daily data are shown in Figure 9.5. The Hurst parameter estimate is $H = 0.61$ and its effect on return periods can be neglected.

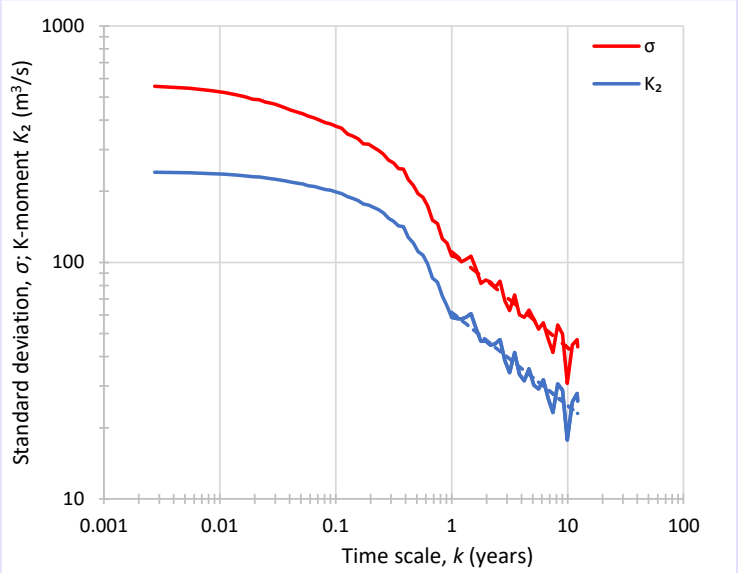

Figure 9.5 Empirical σ-climacogram and K_2-climacogram of daily data of the Susquehanna River at Danville, PA, USA. The power-laws, fitted by regression and plotted as dashed lines, have slopes –0.41 and –0.39, respectively. The Hurst parameter estimate is $H = 0.61$.

Initially, we try a global fitting following the procedure of Digression 9.B. This is shown in Figure 9.6, where it can be seen that this fitting is unsatisfactory for both distribution tails, as well as for the body of the distribution. Good fittings on the tails are shown in Figure 9.7, performed in the same manner as in Digression 9.B.

For a fitting on the body of the distribution we can follow the procedure of section 9.3. This is illustrated in Figure 9.8. The resulting theoretical distribution is good for the body, i.e., for small return periods, but totally inappropriate for either of the tails.

Thus, in this case we have three different fittings, with different parameters sets, each of which are good for a part of the distribution, the upper tail, the lower tail and the body. Assuming that we perform simulation for a particular technical problem, which of the three should we use? For problems related to floods, it is reasonable to use the fitting on the upper tail. But for problems related to low flows the answer is not that easy. At first glance it appears that the fitting on the lower tail is pertinent. However, simulation of low flows is performed when studying regulation structures such as reservoirs and in this case is not the quantity of natural low flow that matters, but the succession of flows. Therefore, it may be more appropriate to use the fitting on the body of the distribution.

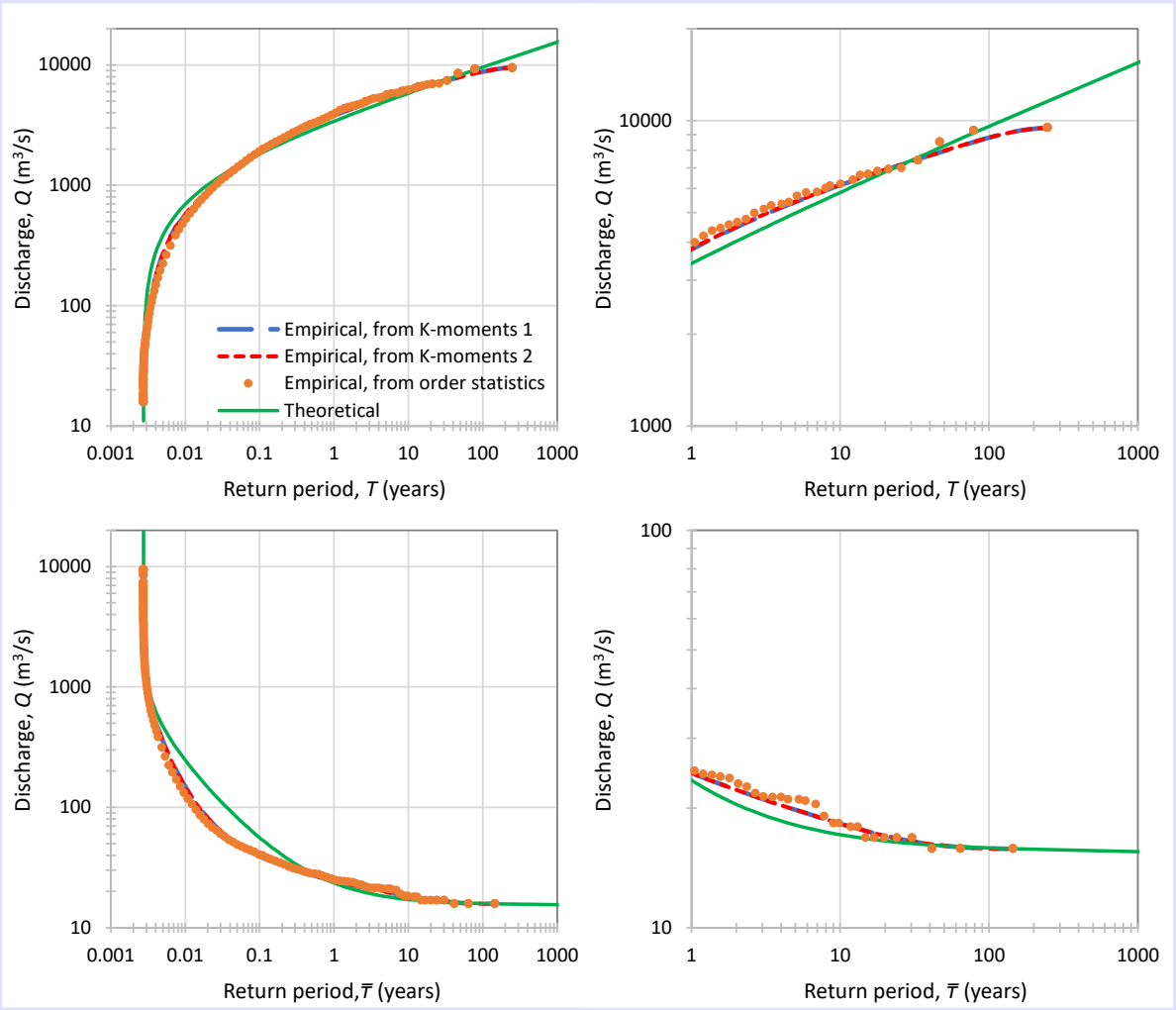

Figure 9.6 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Susquehanna River at Danville, PA, USA. The distribution function is depicted in terms of return periods T or $\overline{T}$ (**upper** and **lower** row, respectively), either on their entire range **(left column)** or focusing on those greater than 1 year **(right column)**. The parameters are: $\xi = 0.202, \zeta = 1.43, \lambda = 495.6\ \mathrm{m^3/s}, x_L = 15.56\ \mathrm{m^3/s}$. For further explanations see caption of Figure 9.3.

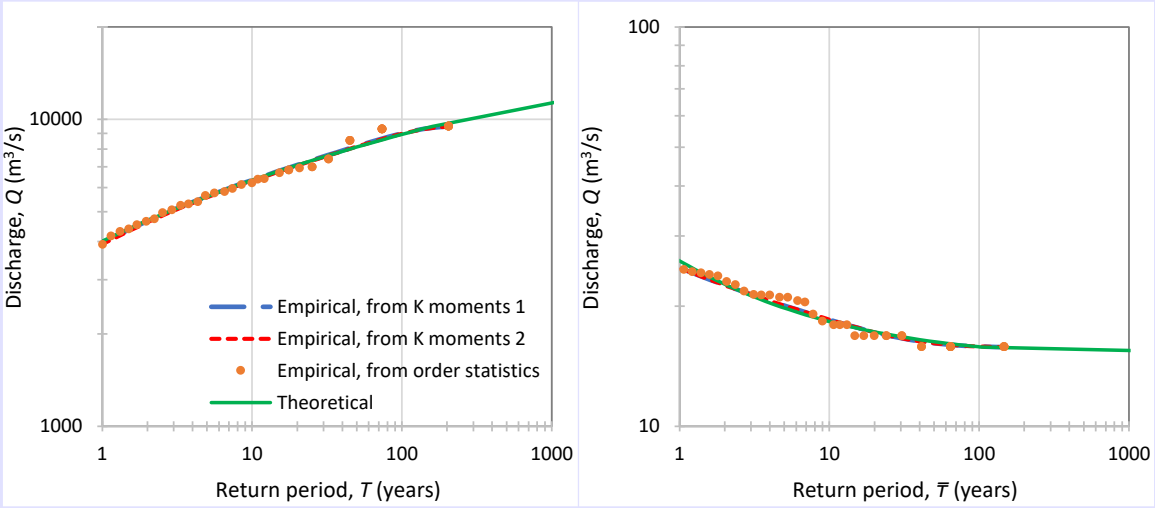

Figure 9.7 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Susquehanna River at Danville, PA, USA, as in Figure 9.6 but with a fitting focusing on $T > 1$ year **(left)** or $\overline{T} > 1$ year **(right)**. The parameters are, for the left panel: $\xi = 0.0, \zeta = 0.721, \lambda = 339.1\ \mathrm{m^3/s}, x_L = 15.64\ \mathrm{m^3/s}$, and for the right panel: $\xi = 0.5, \zeta = 2.06, \lambda = 199.4\ \mathrm{m^3/s}, x_L = 14.60\ \mathrm{m^3/s}$.

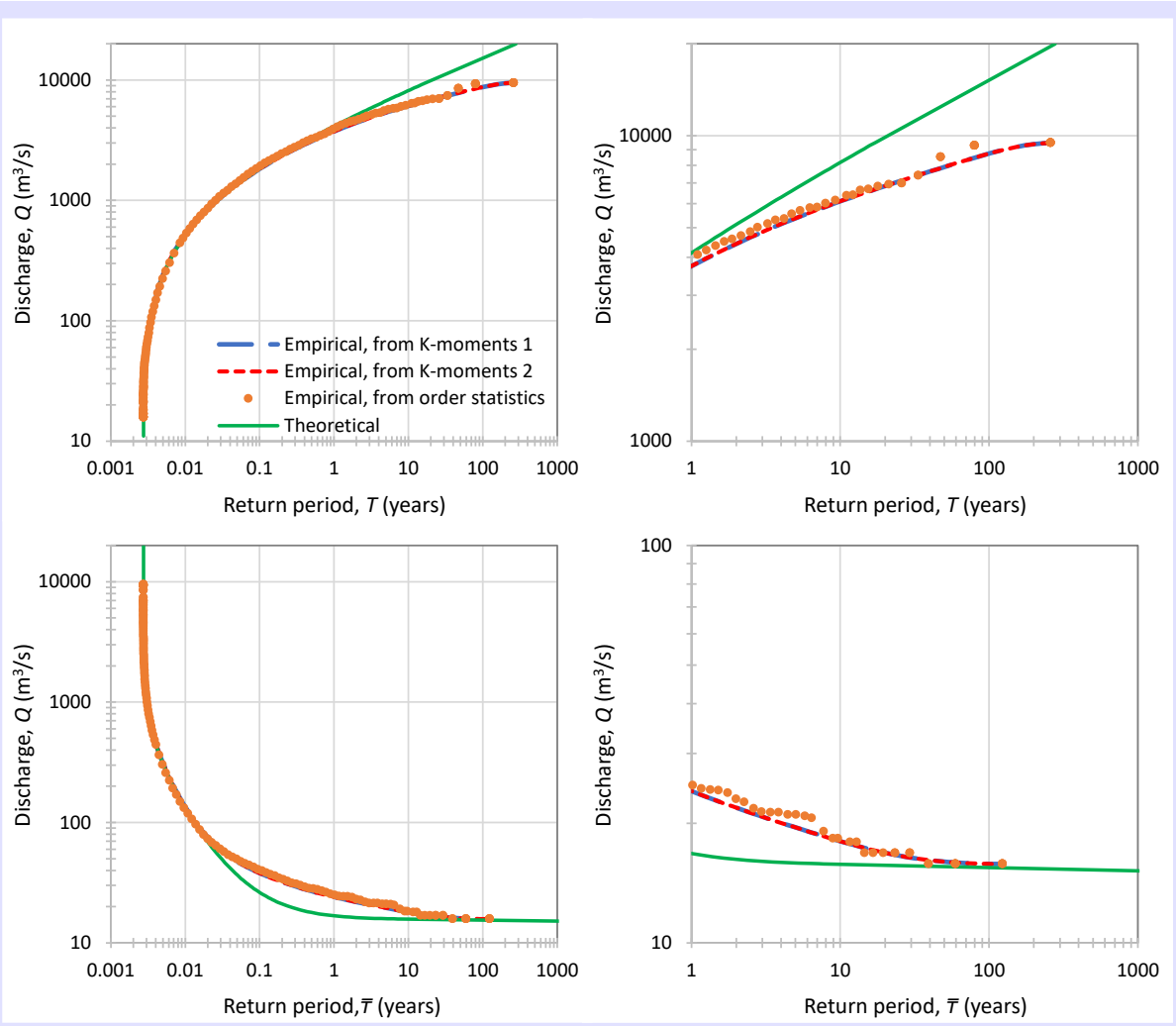

Figure 9.8 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Susquehanna River at Danville, PA, USA, as in Figure 9.6 but with a fitting focusing on the body of the distribution performed by fitting the first four noncentral K-moments. The distribution function is depicted in terms of return periods T or $\overline{T}$ (**upper** and **lower** row, respectively), either on their entire range **(left column)** or focusing on those greater than 1 year **(right column)**. The parameters are: $\xi = 0.247, \zeta = 1.04, \lambda = 329.8\ \mathrm{m}^3/\mathrm{s}, x_L = 15.64\ \mathrm{m}^3/\mathrm{s}$. For further explanations see caption of Figure 9.3.

Obviously, one may think of using a more complex model, such as the sum of two PBF distributions. This is feasible and only needs a slight modification of the methodology, but it is out of our scope as here model parsimony is a strong desideratum.

9.4 Distribution fitting in the presence of persistence

Persistence is quite frequent in streamflow. However, in the above illustrations, its intensity was moderate as the Hurst parameter was around 0.60. If it becomes large, around 0.80 or greater, then it affects the estimates of K-moments and should be taken into account. The fitting methodology described in section 9.2 can deal with such cases readily. An illustration is given in Digression 9.D.

Digression 9.D: An example of distribution fitting in the presence of persistence

We use as an example of a streamflow record with persistence the Red River of the North at Grand Forks, ND, USA (USGS station 05082500, 47.927°N, 97.029°W, drainage area 77 958.6 km²). The data cover the period April 1882 to November 2019 (more than 136 full calendar years, uninterrupted; daily values $n = 50\,269$). The σ-climacogram and the K_2-climacogram of daily data are shown in Figure 9.9. The Hurst parameter estimate is $H = 0.91$ and should have a marked effect, i.e., bias, on return periods. We note that, with this value, the slope of the theoretical σ-climacogram is –0.09. However, the empirical σ-climacogram in Figure 9.9 has a slope –0.18. This is not an inconsistency or an error; it is a result of the fact that the empirical climacogram is affected by bias, which was taken into account in the estimation of $H = 0.91$.

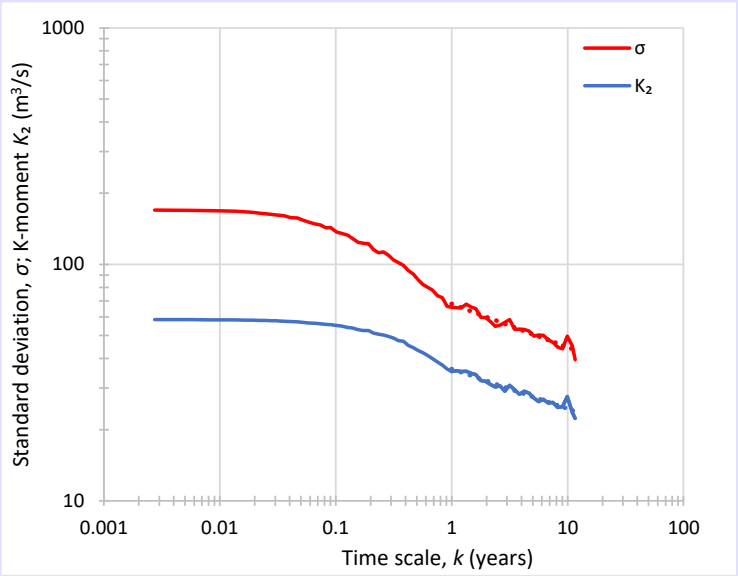

Figure 9.9 Empirical σ-climacogram and K_2-climacogram of daily data of the Red River of the North at Grand Forks, ND, USA. The power-laws, fitted by regression and plotted as dotted lines, have slopes –0.18 and –0.17, respectively. The Hurst parameter, estimated from the annual time series, is $H = 0.91$.

The bias correction factor is $\Theta \approx 2H(1-H)/(n-1) - 1/2(n-1)^{2-2H} = -0.071$. Using $p' \approx 2\Theta + (1-2\Theta)p^{((1+\Theta)^2)}$ we transform each p to p'. For $p = 1, p' = 1$ and for $p = n = 50\,269, p' = 12\,977$ (a big reduction, almost to 1/4). The return periods are then calculated as $\hat{T}(K'_{p_i}) = \Lambda_\infty p' + \Lambda_1 - \Lambda_\infty$. The remaining steps are the same as in Digression 9.B. The global fitting is shown in Figure 9.10. The empirical return periods based on K-moments, as plotted in the figure, are adapted for the bias due to persistence while those based on order statistics are not. While adaptation of the latter is possible, as described in section 8.5, we deliberately avoided it to illustrate the difference. Fittings on the tails are shown in Figure 9.11 performed in the same manner as in Digression 9.B.

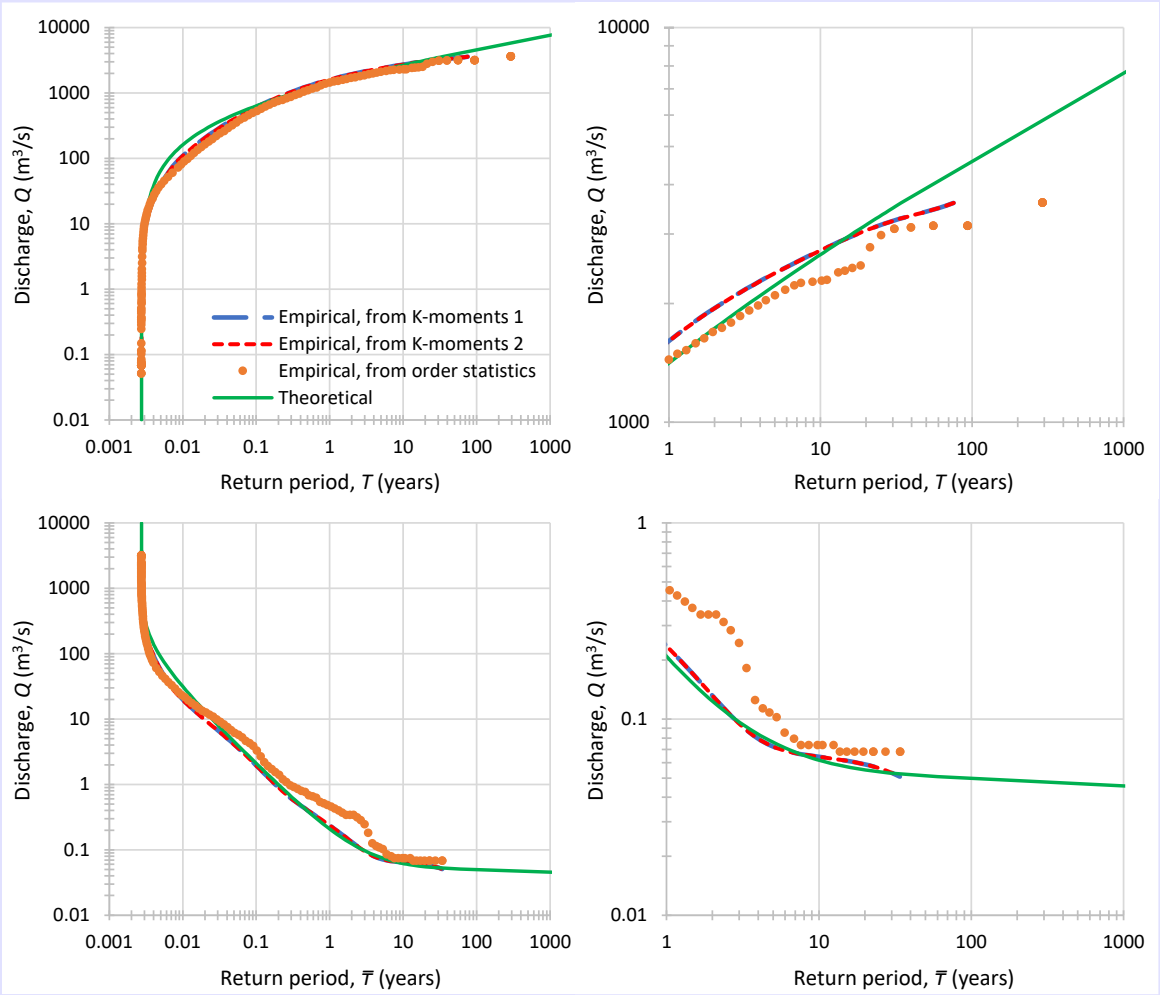

Figure 9.10 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Red River of the North at Grand Forks, ND, USA. The distribution function is depicted in terms of return periods T or $\overline{T}$ (**upper** and **lower** row, respectively), either on their entire range **(left column)** or focusing on those greater than 1 year **(right column)**. The empirical return periods based on K-moments are adapted for the bias due to persistence while those based on order statistics are not. The parameters are: $\xi = 0.194, \zeta = 0.906, \lambda = 106.8\ \mathrm{m^3/s}, x_L = 0.049\ \mathrm{m^3/s}$. For further explanations see caption of Figure 9.3.

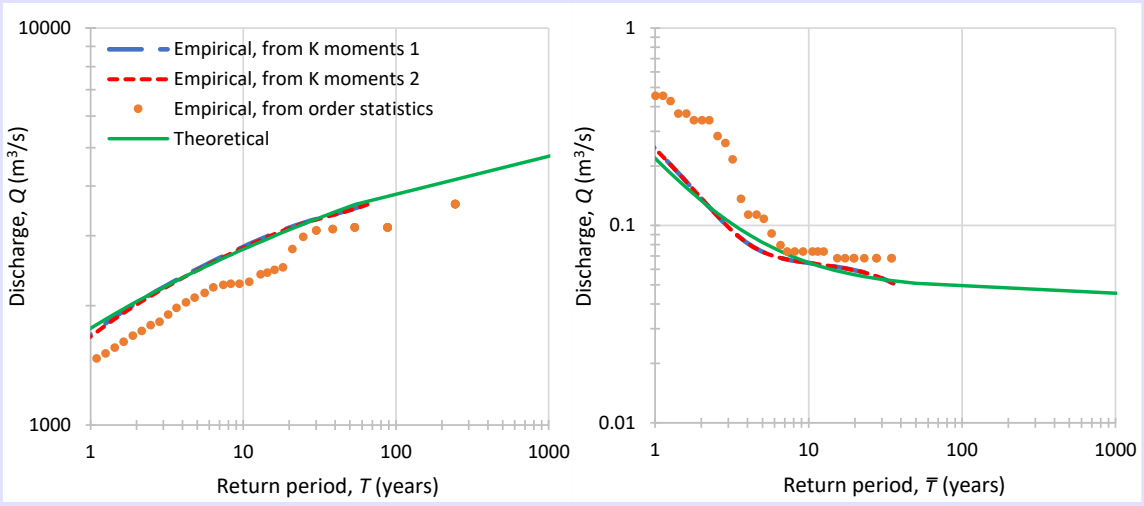

Figure 9.11 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Red River of the North at Grand Forks, ND, USA, as in Figure 9.3 but with a fitting focusing on $T > 1$ year **(left)** or $\overline{T} > 1$ year **(right)**. The parameters are, for the left panel: $\xi = 0, \zeta = 0.720, \lambda = 148.5\ \mathrm{m^3/s}, x_L = 0.033\mathrm{m^3/s}$, and for the right panel: $\xi = 0.5, \zeta = 1.008, \lambda = 59.7\ \mathrm{m^3/s}, x_L = 0.047\ \mathrm{m^3/s}$.

9.5 Distribution fitting in the presence of extraordinary extremes

Sometimes the streamflow time series contain an extraordinary extreme, which may differ significantly from other extremes. This is usually called an outlier and needs particular attention (Grubbs, 1969). An outlier may indicate measurement error but usually it just reflects the high variability of the streamflow—the so-named Noah effect (Mandelbrot and Wallis, 1968). Many view an outlier as a cause of serious problems in statistical analyses and exclude it from the data set to make the analysis easier and the model fitting more elegant.

However, this is not a proper tactic. When studying extremes, excluding the most extreme observation is rather irrational. The suggestion here is to check the measurement conditions to see if it reflects a measurement error. Once this possibility is excluded, the outlier should be kept in the data set and the fitting methodology described in section 9.2 should be kept unchanged. The K-moments framework and in particular the use of K-moments for order $q = 1$ is the most robust in the presence of outliers. This does not mean that the model would not be affected by outliers—this would not be logical. An illustration is given in Digression 9.E, where we also discuss the difference by considering or not the highest observation.

Digression 9.E: An example of distribution fitting in the presence of extraordinary extremes

In this example we use the streamflow record of the Tenmile Creek near Rimini, MT, USA (USGS station 06062500, 46.524°N, 112.257°W). The drainage area is small, 80 km^2). The data cover the period October 1914 to February 2004 with a gap from October 1994 to April 1997 (85 full calendar years, daily values $n = 31\,706$). The daily flow time series is plotted in Figure 9.12, from where it can be seen that on 1981-05-22 an extraordinary extreme flood occurred with a discharge of 53.24 m^3/s. This is by an order of magnitude higher than the usual flood discharges.

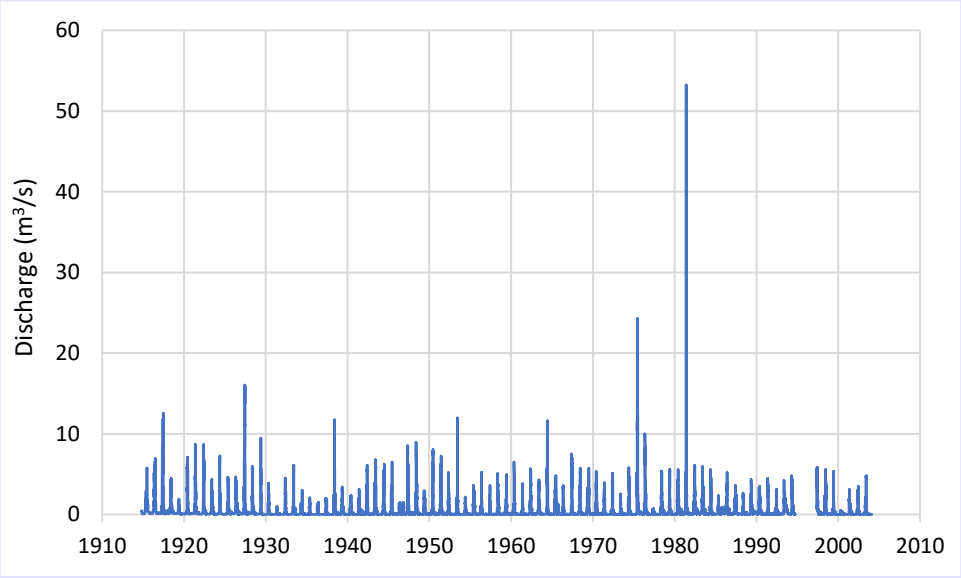

Figure 9.12 Plot of the discharge time series at the Tenmile Creek near Rimini, MT, USA.

The σ-climacogram and the K_2-climacogram of data are shown in Figure 9.13. The climacograms show an erratic behaviour at scales between 1 and 2 years which could be the result of the annual periodicity (prominent also in Figure 9.12). Therefore, the climacograms derived by

the annual time series are also shown, whose plot is smooth. The Hurst parameter, estimated from the annual series is $H = 0.79$ and should have a slight effect, i.e., bias, on return periods.

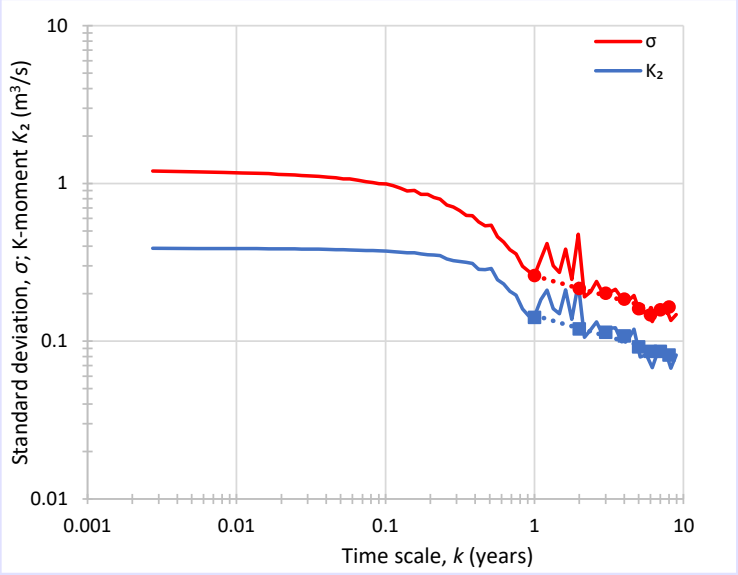

Figure 9.13 Empirical σ-climacogram and K_2-climacogram of daily data of the Tenmile Creek near Rimini, MT, USA. The lines have been constructed from the daily series and the points from the annual series. The power laws, fitted by regression on the points of the annual series and plotted as dotted lines, have slopes −0.26 and −0.27, respectively. The Hurst parameter, estimated from the annual series, is $H = 0.79$.

Contrary to the earlier investigated time series, the Tenmile Creek series contains zero values (dry condition), namely, 48 zero values out of 31 706 total values. This suggests a small probability dry, $P_0 = 0.00151$ at the daily scale. Yet this small value makes the analysis of low extremes unnecessary. Indeed, the return period of the zero value will be $\overline{T}_0/D = 1/0.00151 = 660.5$, or $\overline{T}_0 = 660.5$ d $= 1.81$ years. Therefore, for $\overline{T} \geq \overline{T}_0 = 1.81$ years all distribution quantiles will be zero. Hence, we only study the high extremes here.

Even though the bias is expected to be small in this case ($H = 0.79 < 0.80$) we take it into account for illustration. The upper and lower panels of Figure 9.14 depict fitting without and with bias adaptation, respectively. In the latter case, the bias correction factor is $\Theta \approx 2H(1-H)/(n-1) - 1/2(n-1)^{2-2H} = -0.0064$. As we have done in Digression 9.D, using $p' \approx 2\Theta + (1-2\Theta)p^{((1+\Theta)^2)}$ we transform each p to p'. For $p = 1, p' = 1$ and for $p = n = 31706, p' = 28\,122$ (an 11% reduction). The return periods are then calculated as $\hat{T}(K'_{p_i}) = \Lambda_\infty p' + \Lambda_1 - \Lambda_\infty$. The fitting in the Figure 9.14 was made with weights $w = 1$. It can be seen that the bias adaptation has only slight effects on the parameters, shown in the caption of the figure.

Fittings on the tails, with $w = 0$ for $T < 0$, are shown in Figure 9.15. Its two panels compare the two fitting cases (a) considering all data, including the highest value (left panel), and (b) considering all data but the highest value (right panel). The differences in the two cases are dramatic, both in the distribution parameters (shown in the figure caption), particularly in the tail index ξ and the resulting return periods. In case (a) the return period of the highest value is 213 and 383 years if estimated empirically and theoretically, respectively. The difference is reasonable for an outlier. But in case (b) the theoretical return period becomes 8600 years, 22 times higher. Empirical return period cannot be assigned in this case as the value has been excluded from the analysis. One can further observe that the fitting in case (b) looks better as the agreement between model and (censored) observations is perfect. In case (a) there is difference between theoretical and empirical estimates and also between empirical estimates derived by K-moments and order statistics. However, given the huge difference of the fittings in the two cases, and adopting an engineering point of view, we should clearly prefer the fitting of case (a) and abandon any temptation to dismiss the extraordinary extreme for the sake of modelling elegance.

This case also emphasizes the better performance of K-moments against order statistics.

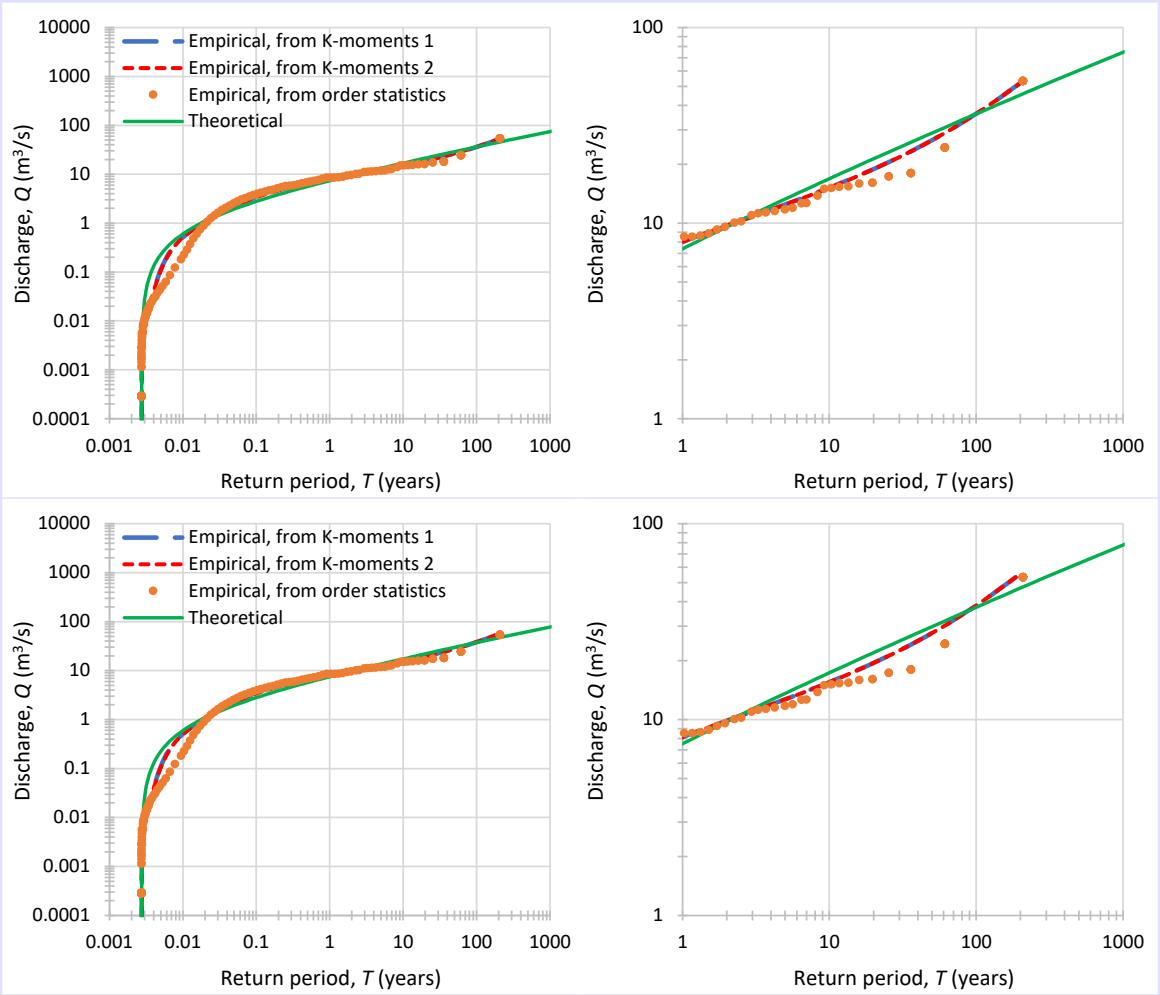

Figure 9.14 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Tenmile Creek near Rimini, MT, USA, without **(upper)** and with **(lower)** bias adaptation (the points based on order statistics are not adapted in either case). The distribution function is depicted in terms of return periods T, either on their entire range **(left column)** or focusing on those greater than 1 year **(right column)**. The parameters are for the upper panels: $\xi = 0.305, \zeta = 0.896, \lambda = 0.37$ m^3/s, $x_L = 0$, and for the lower panels: $\xi = 0.308, \zeta = 0.887, \lambda = 0.36$ m^3/s, $x_L = 0$. For further explanations see caption of Figure 9.3.

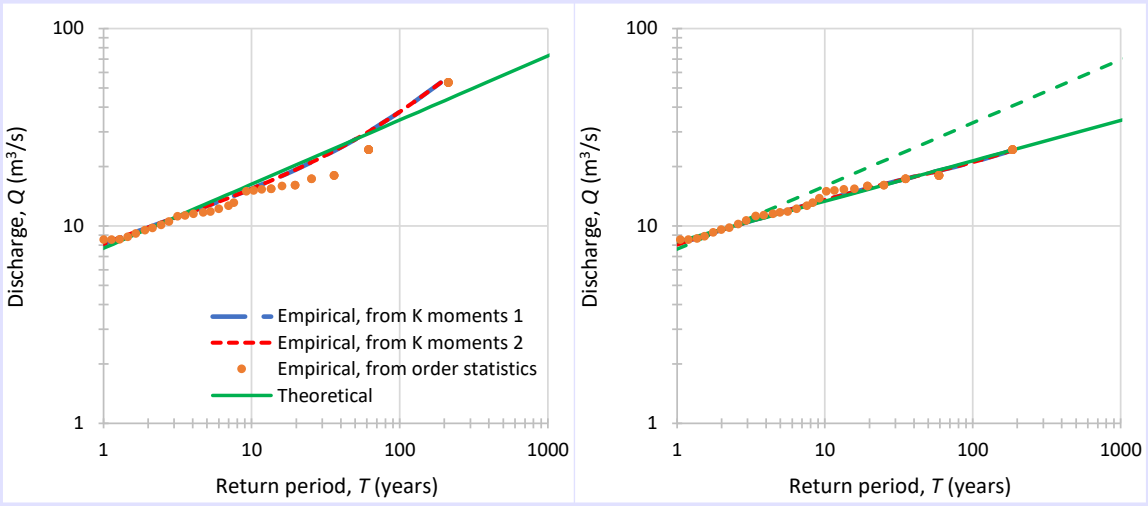

Figure 9.15 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Tenmile Creek near Rimini, MT, USA, as in Figure 9.14 but with a fitting focusing on $T > 1$ year by considering the highest value **(left)** or not **(right)**. The parameters are, for the left panel: $\xi = 0.325, \zeta = 5.08, \lambda = 1.25$ m^3/s, $x_L = 0$, and for the right panel: $\xi = 0.205, \zeta = 20, \lambda = 2.67$ m^3/s, $x_L = 0$.

9.6 Some general remarks

The above framework, implemented in several case studies, suggests that the PBF distribution on one hand and the K-moments approach on the other hand, provide acceptable means to deal with the distribution of streamflow and particularly its tails. However, in most of the cases a single parameter set is not enough to cover the entire span of discharge variation, which can reach 6 orders of magnitude. We generally need different parameter sets for the high and low extremes, and perhaps a third one for the body of the distribution. Notably, the methodology is exactly the same in all three cases. Only the objective function (weighted fitting error) changes.

The case studies examined in the Digressions cover different conditions with varying characteristics. An additional case study is given in Digression 9.F for a large European basin which is not pristine and has a high baseflow, much higher than all basins examined. As already stated in section 9.1, we cannot draw generalized conclusions from a few catchments. However, some indications can be seen in Table 9.1 which gathers the fitted parameter sets of all case studies. The tail index ξ can take large values, but not higher than 1/3, which means that skewness is generally finite, even though the kurtosis could be infinite. For large basins the tail index tends to become zero, which agrees with the indicative theoretical analysis of Digression 9.A. The lower-tail index ζ varies considerably; in most cases it is higher than 1 (bell-shaped density function) but it also can be smaller than 1 (decreasing density function). The lower bound x_L is generally >0 but can be zero for small ones, where a probability dry >0 can emerge.

Table 9.1 Fitted parameter sets of all streamflow records studied.

River and stream gauge	Area (km²)	H (-)	Optimization case	ξ (-)	ζ (-)	λ (m³/s)	x_L (m³/s)
Tenmile Creek near Rimini	80.0	0.79	$e+\overline{e}$ $(w=1)$	0.308	0.887	0.36	0
			e $(T>1$ year$)$	0.325	5.08	1.25	0
French Broad River at Asheville	2 447.5	0.58	$e+\overline{e}$ $(w=1)$	0.307	2.40	51.4	4.38
			e $(T>1$ year$)$	0.277	5.06	81.7	4.87
			$\overline{e}$ $(\overline{T}>1$ year$)$	0	2.16	77.3	4.40
Susquehanna River at Danville	29 059.7	0.61	$e+\overline{e}$ $(w=1)$	0.202	1.43	495.6	15.56
			e $(T>1$ year$)$	0	0.721	339.1	15.64
			$\overline{e}$ $(\overline{T}>1$ year$)$	0.5	2.06	199.4	14.60
			Four moments	0.247	1.04	329.8	15.64
Po River at Pontelagoscuro	70 091	0.61	$e+\overline{e}$ $(w=1)$	0.088	2.03	1879.8	158.9
			e $(T>1$ year$)$	0	1.71	2419.1	0
			$\overline{e}$ $(\overline{T}>1$ year$)$	0.088	2.39	1582.4	148.1
Red River of the North at Grand Forks	77 958.6	0.91	$e+\overline{e}$ $(w=1)$	0.194	0.906	106.9	0.049
			e $(T>1$ year$)$	0	0.720	148.5	0.033
			$\overline{e}$ $(\overline{T}>1$ year$)$	0.5	1.008	59.7	0.047

Digression 9.F: An example for a large European river

In a final example we use the streamflow record of the Po River at Pontelagoscuro, Italy (near the city of Ferrara; 44.888°N, 11.607°E). The drainage area is large, 70 091 km². The Po River has 141 main tributaries and the related river network has a total length of about 6 750 km and 31 000 km for natural and artificial channels, respectively. About 450 lakes are located in the Po River

basin. The water level of the larger south-alpine lakes of glacial origin is regulated according to given management policies, therefore obtaining a regulation volume of approximately 1.3 km³. (Montanari, 2012). Therefore, because of regulation, the streamflow in this case is not natural. Yet one may assume that the highest floods would not differ substantially from natural as the regulation margin is diminished. The data cover the period January 1920 to December 2009 (90 full calendar years, uninterrupted; daily values $n = 32\,850$).

The σ-climacogram and the K_2-climacogram of data are shown in Figure 9.16. The Hurst parameter, estimated from the annual series is $H = 0.61$ (a moderate value) and its effect (bias) on return period estimation is negligible.

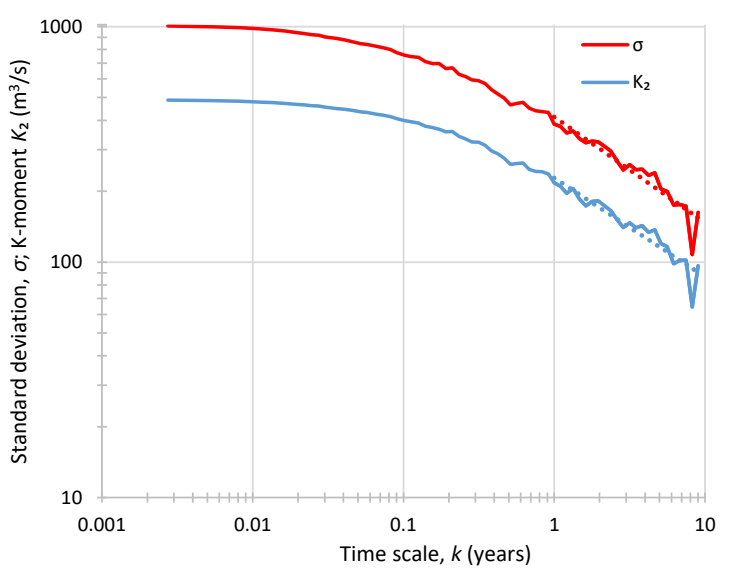

Figure 9.16 Empirical σ-climacogram and K_2-climacogram of daily data of the Po River at Pontelagoscuro, Italy. The power-laws, fitted by regression and plotted as dotted lines, have slopes –0.45 and –0.42, respectively. The Hurst parameter estimate is $H = 0.61$.

The global fitting, performed by the procedure of Digression 9.B, is shown in Figure 9.17, where it can be seen that this fitting is not quite satisfactory for both distribution tails, as well as for the body of the distribution. Good fittings on the tails are shown in Figure 9.18, again performed in the same manner as in Digression 9.B.

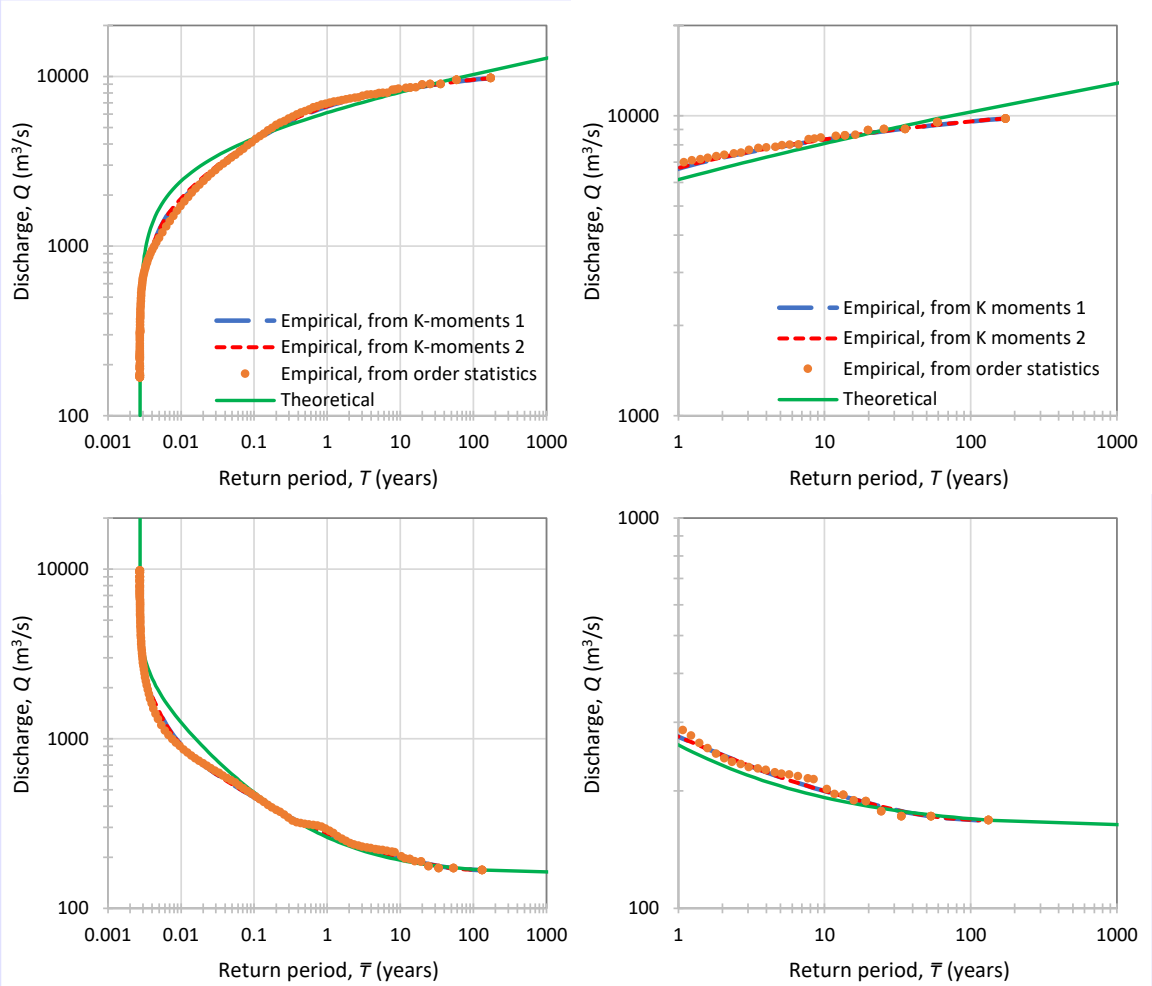

Figure 9.17 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Po River at Pontelagoscuro, Italy. The distribution function is depicted in terms of return periods T or $\overline{T}$ (**upper** and **lower** row, respectively), either on their entire range **(left column)** or focusing on those greater than 1 year **(right column).** The parameters are: $\xi = 0.088, \zeta = 2.03, \lambda = 1879.8\ \mathrm{m}^3/\mathrm{s}, x_L = 158.9\ \mathrm{m}^3/\mathrm{s}$. For further explanations see caption of Figure 9.3.

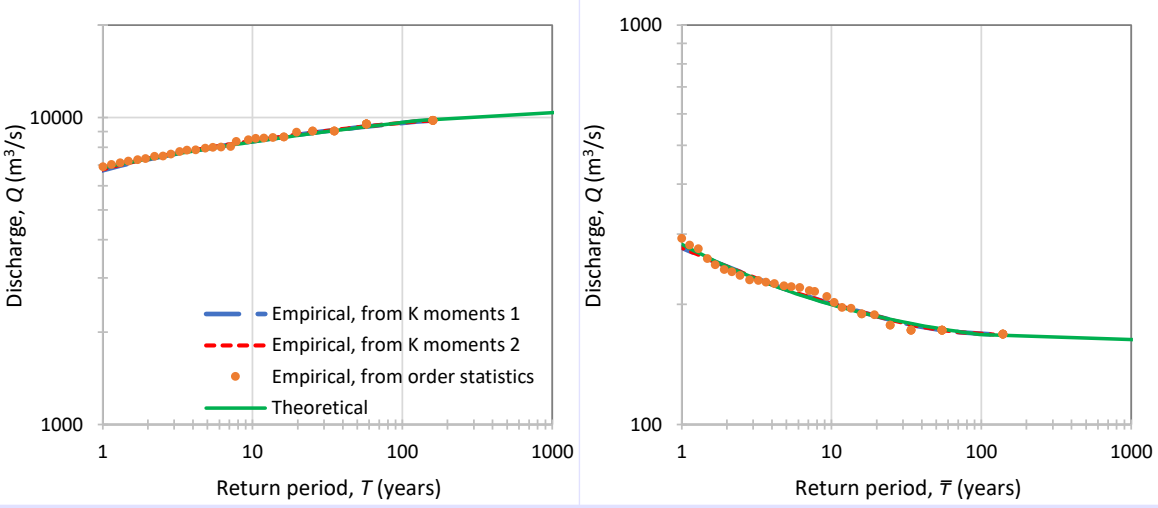

Figure 9.18 Comparison of empirical and theoretical PBF distribution fitted on daily data of the Po River at Pontelagoscuro, Italy, as in Figure 9.17 but with a fitting focusing on $T > 1$ year **(left)** or $\overline{T} > 1$ year **(right)**. The parameters are, for the left panel: $\xi = 0, \zeta = 1.71, \lambda = 2419.1\ \mathrm{m}^3/\mathrm{s}, x_L = 0$, and for the right panel: $\xi = 0.088, \zeta = 2.39, \lambda = 1582.4\ \mathrm{m}^3/\mathrm{s}, x_L = 148.1\ \mathrm{m}^3/\mathrm{s}$.

Chapter 10. Extremes of atmospheric processes

We have already studied the most demanding processes: streamflow, which varies among several orders of magnitude, and rainfall, whose study requires description at many time scales simultaneously. The study of extremes of atmospheric process is easier and no additional methodology is required. In addition, databases with observational information of atmospheric processes abound. This information includes ground data (measurements at meteorological stations), reanalyses (gridded data resulting from assimilation of meteorological measurements into weather models) and satellite data (resulting from images and observations by remote sensing instruments and incorporation of ground measurements). Here we will deal with ground data only, which are best suited for the study of extremes. The other categories are useful for studies in the global or continental scales (see examples in Koutsoyiannis, 2020b).

Atmospheric data such as wind, temperature, pressure, radiation, etc., can easily be retrieved from several databases publicly available. Among them, KNMI's Climexp system[*], in connection with the European Climate Assessment & Dataset project (ECAD; Klein Tank et al., 2002)[†] is the most convenient. Other data such as relative humidity, vapour pressure and dew point can be accessed through national databases; good examples are the Climate Data Center (CDC) of the German Meteorological Service (Deutscher Wetterdienst)[‡] and the USA NOAA's National Centers for Environmental Information (NCEI).[§]

In the following sections of this chapter we provide representative examples for the wind speed, temperature and dew point, where the latter is similar (and measured on same units) with temperature, but has higher hydrological importance as it determines the quantity of water vapour in the atmosphere. The behaviour of wind speed is not very different from that of precipitation, even though its variability is smaller. Temperature and dew point have different behaviour as their distributions are well-formed bell-shaped and not very distant from the normal. Bell-shaped is the distribution of air pressure, too.

In any of these processes the persistence can reach high levels (e.g. $H = 0.9$) and therefore, whenever this happens, the bias due to autocorrelation should be taken into account. Negligence (and often ignorance) of persistence is common among scientists and practitioners, less in the engineering community and more in the climatological community as well as in the insurance industry, which provides services related to extreme events. Not only does this negligence affect a distribution fitting but may also have negative consequences in the design, operation and management of structures. A relevant example is in wind farms, which are typically designed without studying

---

[*] https://climexp.knmi.nl/.

[†] This also provides access to data (more than 20 000 meteorological stations): https://www.ecad.eu/; see also: https://data.europa.eu/euodp/en/data/dataset/jrc-tmy-tmy-download-service.

[‡] https://cdc.dwd.de/portal/.

[§] https://www.ncei.noaa.gov/access/search/dataset-search; http://www.ncdc.noaa.gov/isd; ftp://ftp.ncdc.noaa.gov/pub/data/noaa/. We note though that this data base is not as easy to use as the other ones.

persistence. In the operation phase, the operators are often surprised to see the installation deliver less power than the average for long periods. However, this is a normal behaviour, related to persistence; it is exactly the clustering of low wind periods.

Most of the processes are characterized by double cyclostationarity, daily and annual. This should be taken into account when a full stochastic model of the process of interest is constructed (e.g. Dimitriadis and Koutsoyiannis, 2015b; Deligiannis et al., 2016). However, here we only deal with the marginal distributions with emphasis on extremes and we will omit the study of cyclostationarity.

10.1 Wind

As already mentioned, the behaviour of wind speed is not very different from that of precipitation. However, the variability of the former is smaller. As seen in Chapter 8, intermittence is a prominent behaviour in rainfall, which has been modelled through the probability wet (P_1) or dry (P_0). On the other hand, as seen in Chapter 9, in streamflow intermittence can appear either in a manner similar to rainfall (in small ephemeral streams), or in two-state wet mode, one in which the river is fed merely by groundwater (baseflow) and one dominated by flood. In the case of wind in principle there is no intermittence in the sense of a state where air ceases to move. However, the very small wind speeds are often registered as zero, because of imperfection of the measuring equipment. This case emerges particularly at hourly or finer time scales.

Long time series of wind speed suggest long-term persistence and, most often, heavy-tailed distributions (Tsekouras and Koutsoyiannis, 2014; Koutsoyiannis et al., 2018), even though the light-tailed Weibull distribution has been the dominant model in the literature.

The stochastic analysis of wind speed has recently become very important because of its relevance with wind energy generation. In renewable energy design and management, it is the body of the distribution, rather than its tails, which matters. Indeed, values at either of the tails are not relevant to energy production because the production ceases when the wind speed is too low or too high. The lower tail per se is not quite important, apart from the probability that the wind speed is below the threshold at which the production ceases. However, the upper tail is very important for the safe design of the wind turbine and the entire construction, as the extreme winds determine the turbine loads. The calculations of the load (see e.g. Dai et al., 2011) are apparently out of our scope.

Below we provide representative examples for studies of daily (Digression 10.A) and hourly (Digression 10.B) wind speed data.

Digression 10.A: An example of fitting a PBF distribution on mean daily wind speed

To illustrate the behaviour of wind on daily scale we use one of the longest data sets, that of the Hoofdplaat station in Netherlands (51.38°N, 3.67°E, 0.0 m). The data cover the period April 1908 to December 2019 (111 calendar years, with an interruption of 10 months starting in December 1944; daily values $n = 40\,720$, all of which but four are >0). Characteristic plots of the time series are given in Figure 10.1.

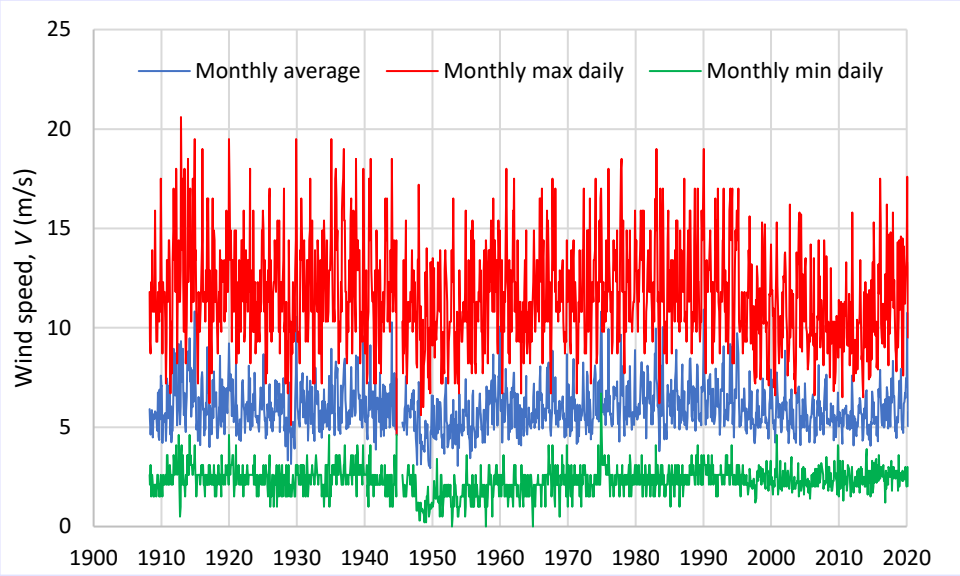

Figure 10.1 Monthly plot of the daily wind speed time series at the Hoofdplaat station, Netherlands. For each month, the average, the minimum and the maximum daily values are plotted.

The σ-climacogram and the K_2-climacogram of data are shown in Figure 10.2. The Hurst parameter, estimated from the annual series is $H = 0.9$ and its effect (bias) on return period estimation is substantial.

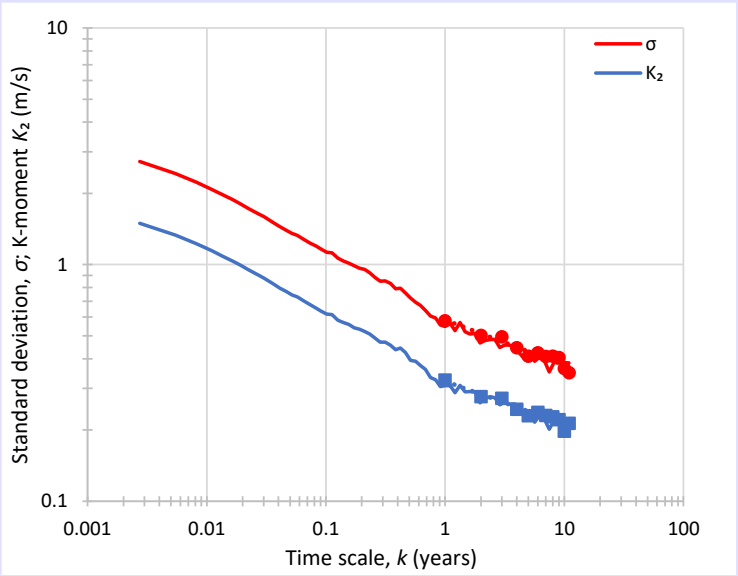

Figure 10.2 Empirical σ-climacogram and K_2-climacogram from daily wind speed time series at the Hoofdplaat station, Netherlands. The lines have been constructed from the daily series and the points from the annual series. The power laws, fitted by regression on the points of the annual series and plotted as dotted lines, have slopes –0.19 and –0.18, respectively. The Hurst parameter estimate is $H = 0.90$.

The fitting of the PBF distribution, performed by the procedure of Digression 9.B, is shown in Figure 10.3. It can be seen that this fitting is quite satisfactory for the body and tails of the distribution, even though it was made with focus on $T > 1$ year tails. The optimal parameter ξ is 0 and thus the PBF distribution in this case switches to the Weibull distribution.

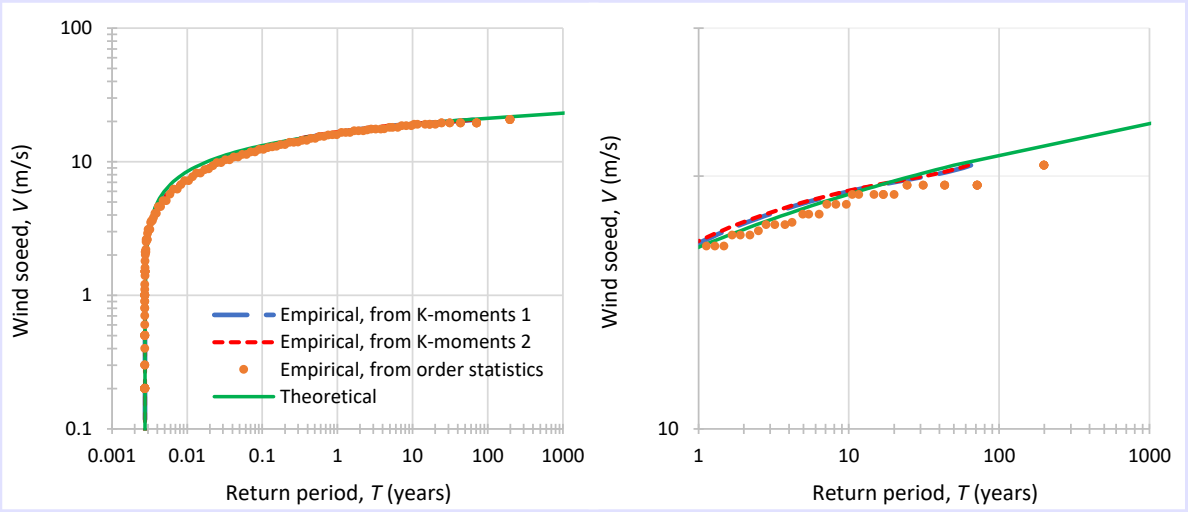

Figure 10.3 Comparison of empirical and theoretical PBF distribution fitted on daily wind speed data of the Hoofdplaat station, Netherlands. The fitting was done based on noncentral and tail K-moment using the linear approximation (denoted in the figure "K-moments 1"). For comparison the nonlinear approximation ("K-moments 2") is also shown but it is indistinguishable from the former case. Furthermore, empirical return periods assigned from order statistics are also shown. The points for K-moments have been adapted for the bias effect of autocorrelation but the points based on order statistics have not. The distribution function is depicted in terms of return period T, either on their entire range **(left)** or focusing on those greater than 1 year **(right)**. The parameters are: $\xi = 0, \zeta = 2.28, \lambda = 7.53$ m/s, $x_L = 0$ and were fitted with focus on $T > 1$ year.

Digression 10.B: An example of fitting a PBF distribution on mean hourly wind speed

Hourly or sub-hourly wind data are rarer than daily in publicly available databases. However, they are quite useful for operational use in diverse tasks such as simulation of wind energy generation, design of wind turbines and estimation of wind load on buildings. Here we use one of the longest hourly data sets, that of the MIT station in Boston, Ma., USA (42.367°N, 71.033°W, 9.0 m).

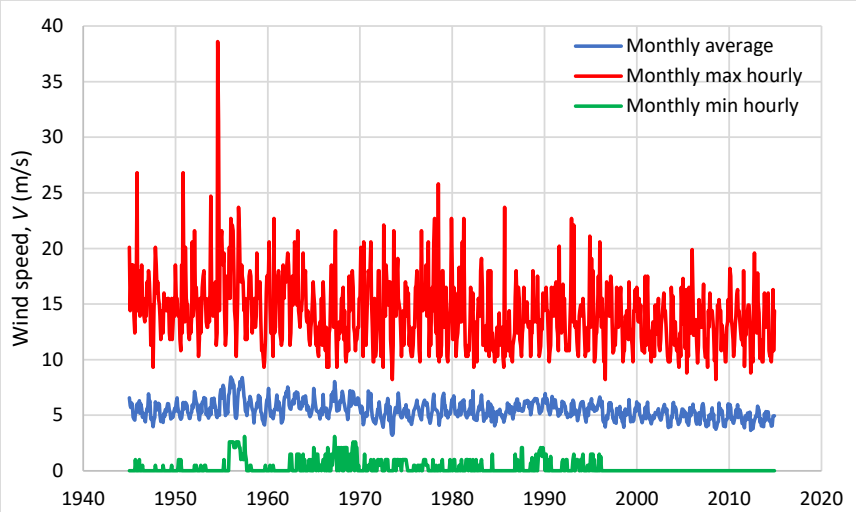

Figure 10.4 Monthly plot of the hourly wind speed time series at the MIT station in Boston, Ma., USA. For each month, the average, the minimum and the maximum hourly values are plotted.

The data are available from the NOAA system but a great deal of effort was needed to convert them to hourly time series, which was undertaken in the study by Dimitriadis and Koutsoyiannis (2018). They cover, with minor gaps, the period January 1945 to December 2014 (70 calendar

years with a few missing values; hourly values $n = 589\,551$, 1% of which are zero). Characteristic plots of the time series are given in Figure 10.4.

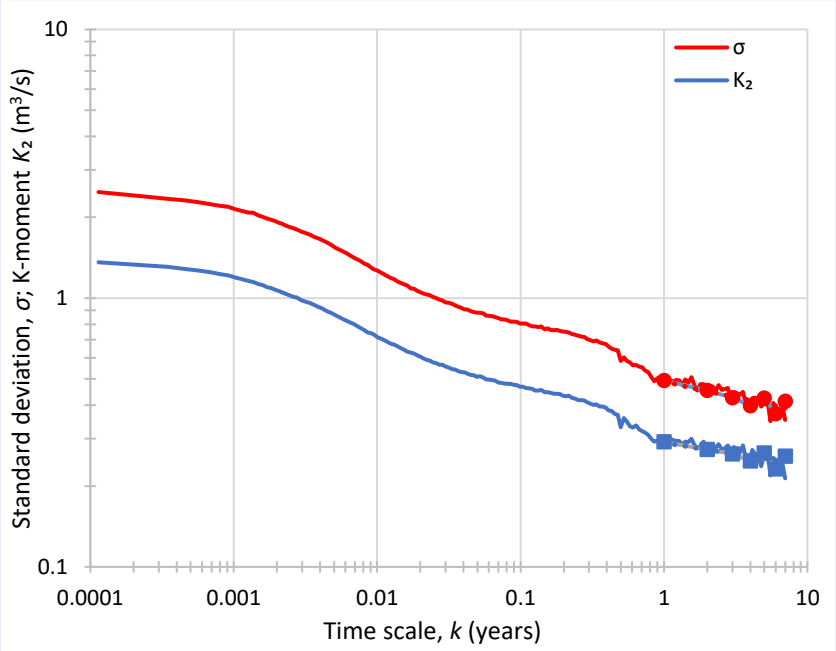

Figure 10.5 Empirical σ-climacogram and K_2-climacogram from hourly wind speed time series at the MIT station in Boston, Ma., USA. The lines have been constructed from the daily series and the points from the annual series. The power laws, fitted by regression on the points of the annual series and plotted as dotted lines, –0.12 and –0.09, respectively. The Hurst parameter estimate is $H = 0.92$.

The fitting of the PBF distribution, performed by the procedure of Digression 9.B, is shown in Figure 10.3. It can be seen that this fitting is quite satisfactory for the body and tails of the distribution, even though the it was made with focus on $T > 1$ year tails. In this case the tail index is substantially different from 0 (0.12 to 0.15, depending on the focus of the fitting, similar to the typical values in rainfall).

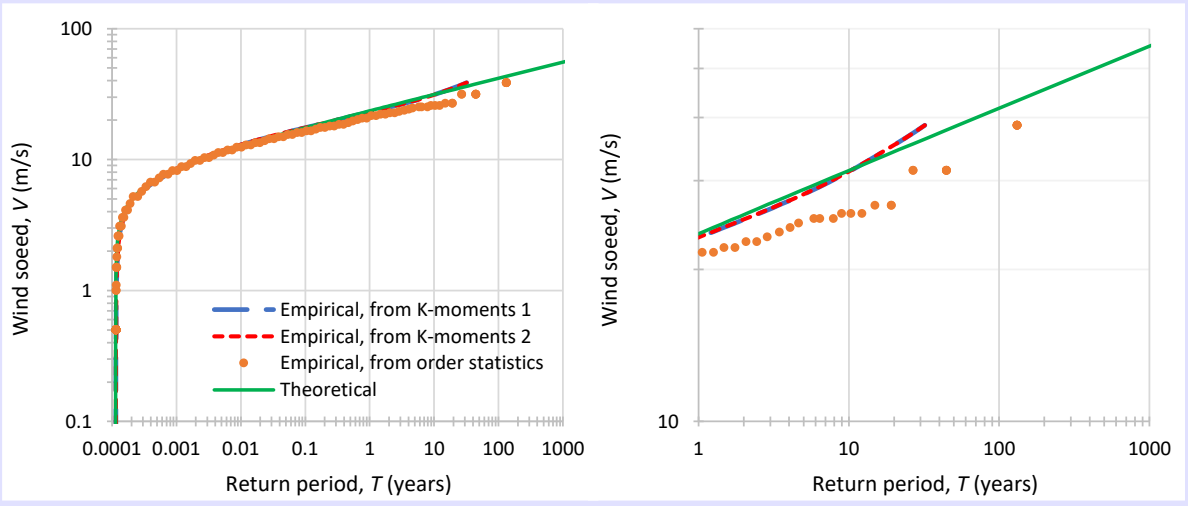

Figure 10.6 Comparison of empirical and theoretical PBF distribution fitted on the hourly wind speed time series at the MIT station in Boston, Ma., USA. The distribution function is depicted in terms of return period T, either on their entire range **(left)** or focusing on those greater than 1 year **(right)**. The parameters are: $\xi = 0.122, \zeta = 3.08, \lambda = 5.71$ m/s, $x_L = 0$ and were fitted on the entire domain of wind speed (a fit with focus on T>1 year increases the tail index to ξ=0.148). For further explanations see caption of Figure 10.3.

In the above analysis we have used all values in the time series, including the zeros. If we exclude the zero values ($n_0 = 6086$), we can study also the resulting lower tail, using the $n =$

583 465 nonzero values. To visualize both tails in a single graph we use a probability plot in terms of the excess return period (see Digression 5.A). This is seen in Figure 10.7 where the fitting of the PBF distribution is excellent (compare the theoretical curve with the empirical ones based on K-moments), spanning 13 orders of magnitude of excess return period.

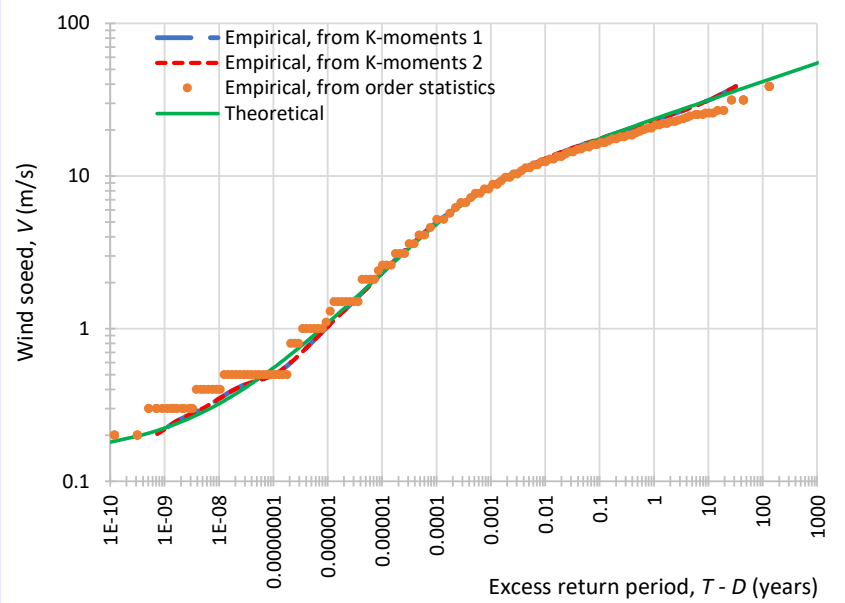

Figure 10.7 Comparison of empirical and theoretical PBF distribution fitted on the hourly wind speed time series at the MIT station in Boston, Ma., USA, excluding the zero values. The distribution function is depicted in terms of the excess return period $T - D$, on its entire range. The parameters are: $\xi = 0.120, \zeta = 2.73, \lambda = 5.35$ m/s, $x_L = 0.15$ m/s and were fitted on the entire domain of wind speed. The fitting was done based on noncentral and tail K-moment using the linear approximation (denoted in the figure "K-moments 1"). For comparison the nonlinear approximation ("K-moments 2") is also shown but it is indistinguishable from the former case. Furthermore, empirical return periods assigned from order statistics are also shown. The points for K-moments have been adapted for the bias effect of autocorrelation but the points based on order statistics have not.

10.2 Temperature

The study of temperature has been a very hot topic due its direct relationship with "global warming", "climate change" and other similar slogans. Certainly, temperature is connected to the concentration of carbon dioxide in the atmosphere. Despite another heralded slogan, "science is settled", this relationship still is not clear, while in a recent study Koutsoyiannis and Kundzewicz (2020) have shown that the causality relationship is not clear. They suggested that the relationship of atmospheric CO_2 and temperature (T) may qualify as belonging to the category of "hen-or-egg" problems, where it is not always clear which of two interrelated processes is the cause and which the effect. Examining modern data (1980–2019) of T and CO_2 concentration, they concluded that, while both causality directions exist, the dominant direction is $T \rightarrow CO_2$ and not $CO_2 \rightarrow T$ as commonly perceived. Paleoclimatic data are even more categorical that changes in temperature precede those in CO_2.

At the global level, the temperature is meaningfully influenced by the rhythm of the major ocean-atmosphere fluctuations, such as the ENSO and IPO in the Pacific as well as the AMO in the Atlantic (Kundzewicz et al., 2020). At the local level, the global behaviour certainly plays a role on local changes, but other factors, most profoundly urbanization,

may also influence the temperature substantially. This is exemplified in Digression 10.C, referring to the longest instrumental temperature record in Milano, Italy. The behaviours seen in that record are characteristic for temperature time series. Among these are the high persistence and the light-tailed distributions close to Normal, which entail much lower variability than in other variables such as rainfall, runoff and wind.

Digression 10.C: Study of the longest instrumental temperature record (Milano, Italy)

According to both the Global Historical Climate Network (GHCN) – Daily and the European Climate Assessment & Dataset project (ECAD), the meteorological station with the longest temperature record is that of Milan (Milano) in Italy (45.47°N, 9.19°W, 150.0 m). The measurements have started in 1763 and the data are available up to 2008 from KNMI (and other data bases). They cover, with minor gaps, the period January 1763 to November 2008 (246 years with a few missing values; daily values $n = 89\,686$ for daily maximum temperatures and $n = 89\,697$ for daily minimum temperatures). Here we study both the daily maximum and the daily minimum temperatures, which are typically original (raw) measurements, instead of the daily mean temperatures which are processed data and thus subject to the changes in processing methodologies.

For better understanding of the conditions in the wider area and also for seeing the evolution after 2008, we also study two adjacent stations with long records. The first is in Lugano, Switzerland (46.00°N, 8.97°E, 273.0 m, ~60 km from Milano), a small town with population of 55 359 in 2018,[1] which has not changed substantially in the last 50 years, while the geomorphology of the area does not favour urban expansion. The measurements have started in 1901 and the data are available up to date from KNMI (and other data bases). They cover, with minor gaps, the period January 1901 to June 2020 (more than 119 years with a few missing values; total values $n = 43\,573$ for daily maximum temperatures and $n = 43\,585$ for daily minimum temperatures).

The second is in Monte Cimone, Italy (44.20°N, 10.70°E, 2165.0 m, ~185 km from Milano), a mountainous area without settlements. The measurements have started in 1950 and the data are available up to date from KNMI (and other data bases). They cover, with minor gaps, the period January 1951 to November 2018 (67 years, in some of which there are missing values; total values $n = 24\,283$ for daily maximum temperatures and $n = 24\,263$ for daily minimum temperatures).

The daily time series for Milano and Lugano are plotted in Figure 10.9, along with the running, on 10-year windows, maximum and minimum values for return period of 2 years. The latter were estimated on the basis of the K-moment for the appropriate moment order which corresponds to return period of 2 years. A separate plot is included for the running maxima and minima of 2-year return period for all three stations. All plots indicate upward and downward fluctuations with the upward ones prevailing for the minimum temperature in the period 1940 to today. At the same period, the maximum temperature in Milano is also increasing. However, by comparing the temporal evolution of maxima in Milano for the 2-year return period and those of the two nearby stations, Lugano and Monte Cimone, in which there is no increasing trend, it appears that urbanization might be the principal factor causing temperature increase, rather than global effects. On the other hand, the increase of the minimum temperatures in the recent decades seems to be a more general behaviour (Glynis, 2020). Apparently, this is a favourable behaviour as it reduces the incidence and impacts of extreme cold (see section 11.3).

A typical "modern" interpretation of the situation would be to attribute the upward segments to global warming, which in turn would be attributed to human CO_2 emissions, etc., and eventually to rely on climate models for future projections. However, here we prefer to investigate the stochastic properties of the time series and try to build a consistent stationary stochastic representation with long term persistence. Indeed, as shown in Figure 10.10, the σ-climacograms and the K_2-climacograms of all series suggest high values of the Hurst parameter, estimated from the annual series, between $H = 0.91$ and $H = 0.94$. Even the Lugano series of maximum daily temperature, which does not show a warming trend, suggests $H = 0.91$.

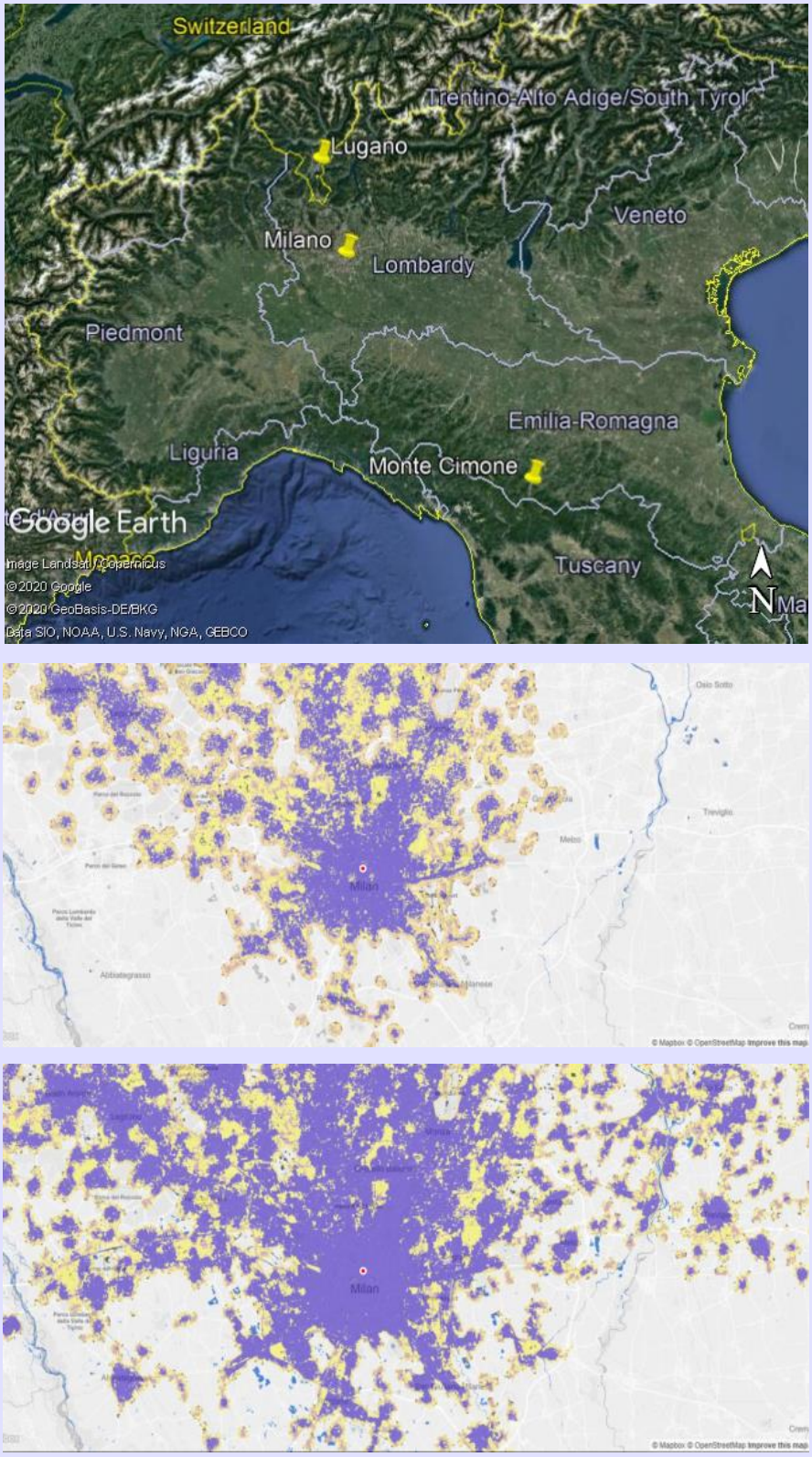

Figure 10.8 (upper) Location of three meteorological stations (source: Google Earth). **(middle and lower)** Depiction of urbanization in Milano in 1988 (population 3 506 838, urban extent 88 417 ha) and in 2013, respectively (population 6 402 051, urban extent 277 177 ha) (source: Glynis, 2019, from data provided by the Atlas of Urban Expansion[2]).

Notably, the oscillations in the climacograms constructed from the daily series in Figure 10.10 reflect the strong annual periodicity of temperature (cf. Koutsoyiannis, 2017). Yet, these have not influenced the estimation of H as this has been based on the climacograms of the annual series.

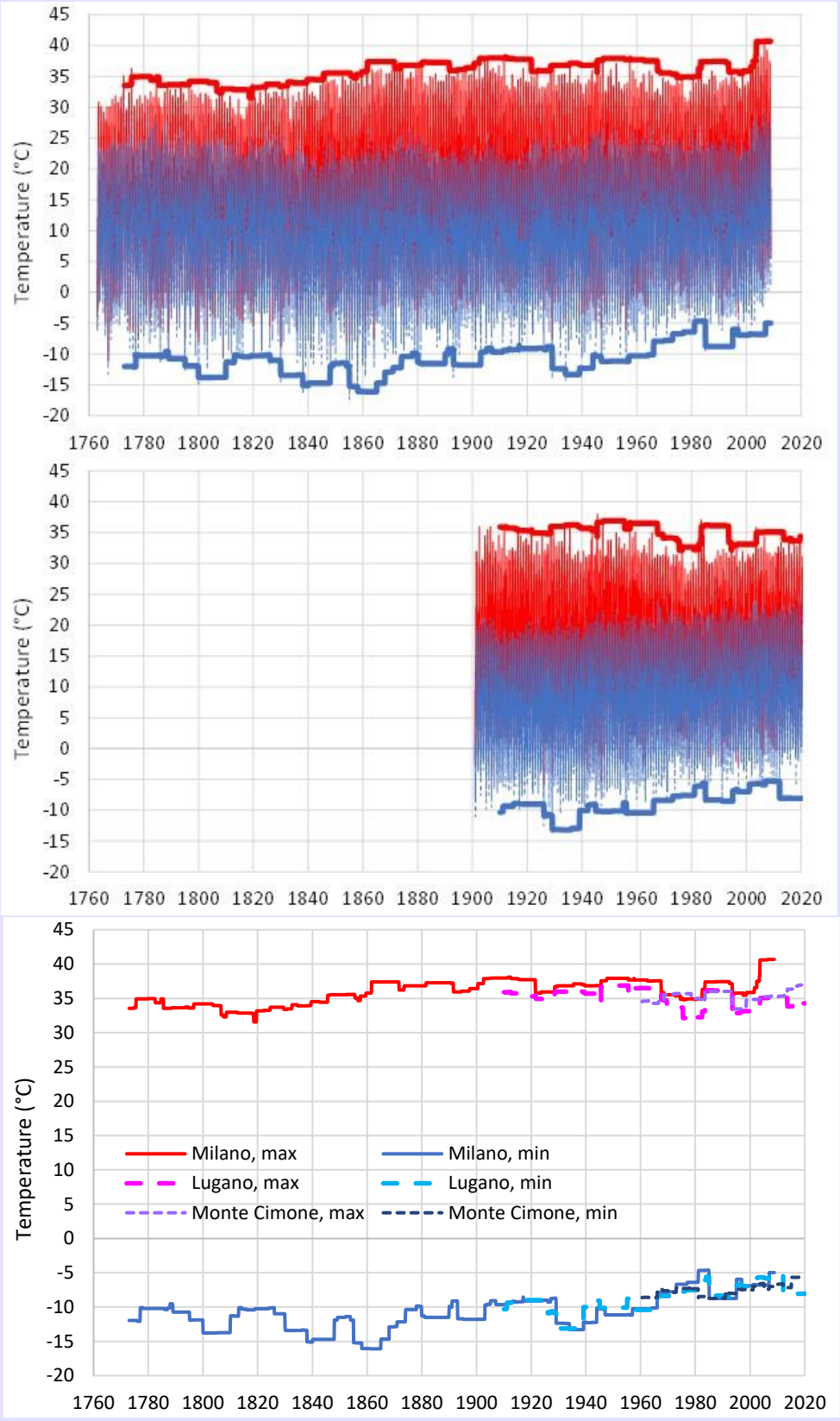

Figure 10.9 (upper) Time series of daily maximum (continuous lines) and minimum (dashed lines) temperature of Milano, along with the running, on 10-year windows, maximum and minimum values for return period of 2 years. **(middle)** As upper but for Lugano. **(lower)** Running, on 10-year windows, maximum and minimum values for return period of 2 years. for all three stations. For comparability, the temperatures of Monte Cimone were shifted up by 13 °C because of the altitude difference of 2 km, assuming a temperature gradient of 6.5 °C/km.

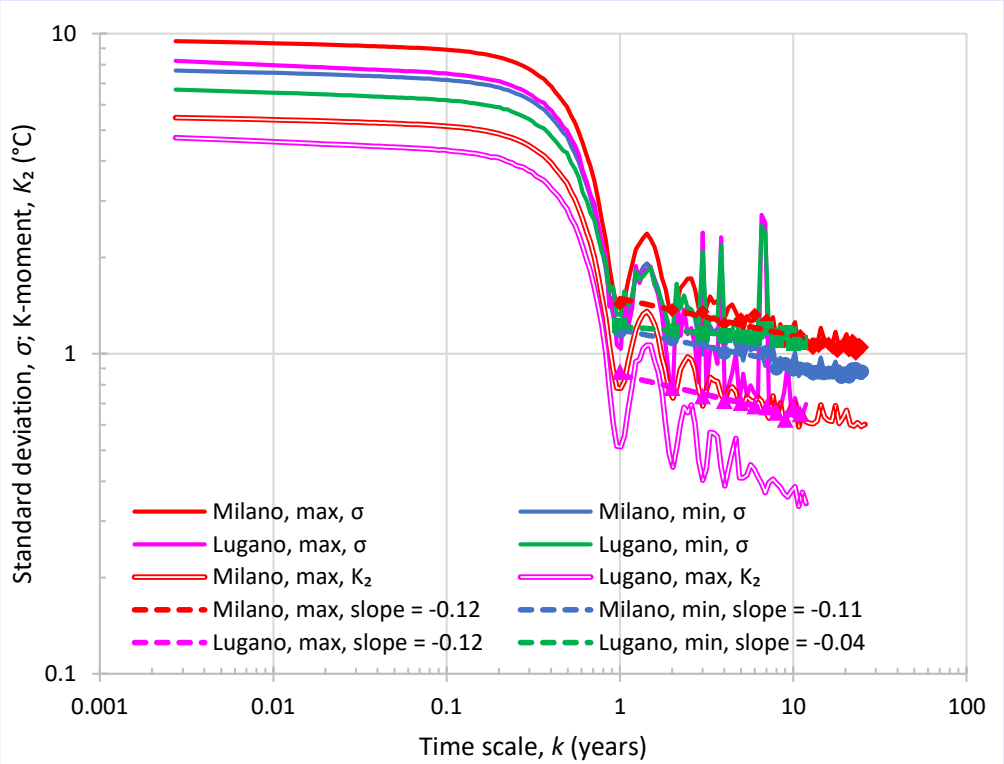

Figure 10.10 Empirical σ-climacogram (single lines) and K_2-climacogram (double lines) from daily maximum and minimum temperatures at Milano, Italy, and Lugano, Switzerland. The lines have been constructed from the daily series and the points of the same colour from the annual series. The points corresponding to the annual series suggest power laws with slopes as indicated. The Hurst parameters, estimated from the annual series, are $H = 0.93$ for both series of Milano, $H = 0.91$ for the maximum temperature at Lugano and $H = 0.94$ for the minimum temperature at Lugano.

The fitting of the Weibull distribution on the daily maximum temperatures of Milano, performed by the procedure described in Digression 9.B, is shown in Figure 10.12. The fitting is not ideal for the entire range of temperature (left panel) but if we focus on the upper tail, for $T >$ 1 year, the fitting becomes perfect (right panel).

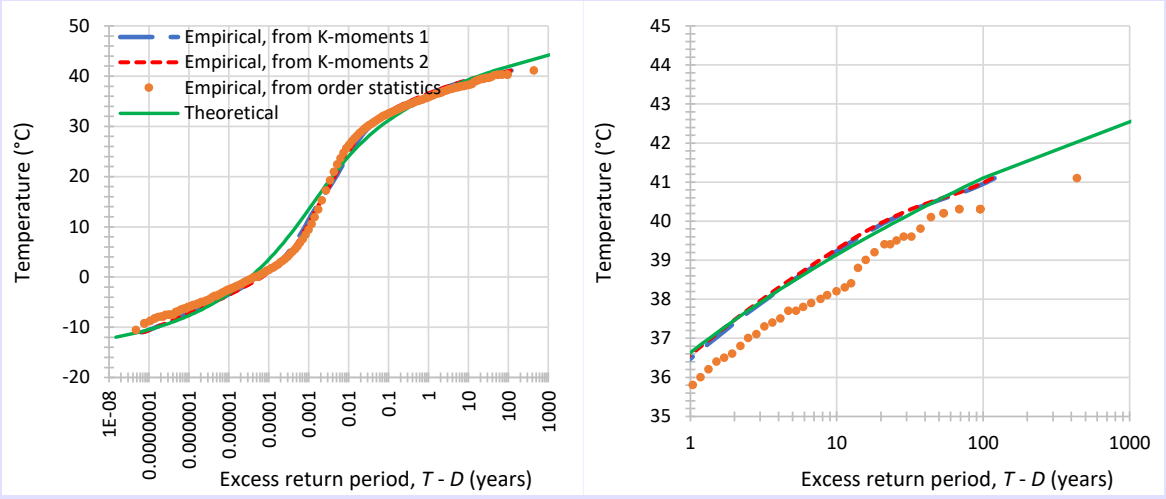

Figure 10.11 Comparison of empirical and theoretical Weibull distribution fitted on the daily maximum temperatures at Milano, Italy. The distribution function is depicted in terms of excess return period $T - D$, either on their entire range **(left)** or focusing on those greater than 1 year **(right)**. The parameters are: $\zeta = 5.11, \lambda = 35.20$ °C, $x_L = -15.37$ °C for the left panel and $\zeta = 6.97, \lambda = 40.23$ °C, $x_L = -15.26$ °C for the right panel. For further explanations see caption of Figure 10.3.

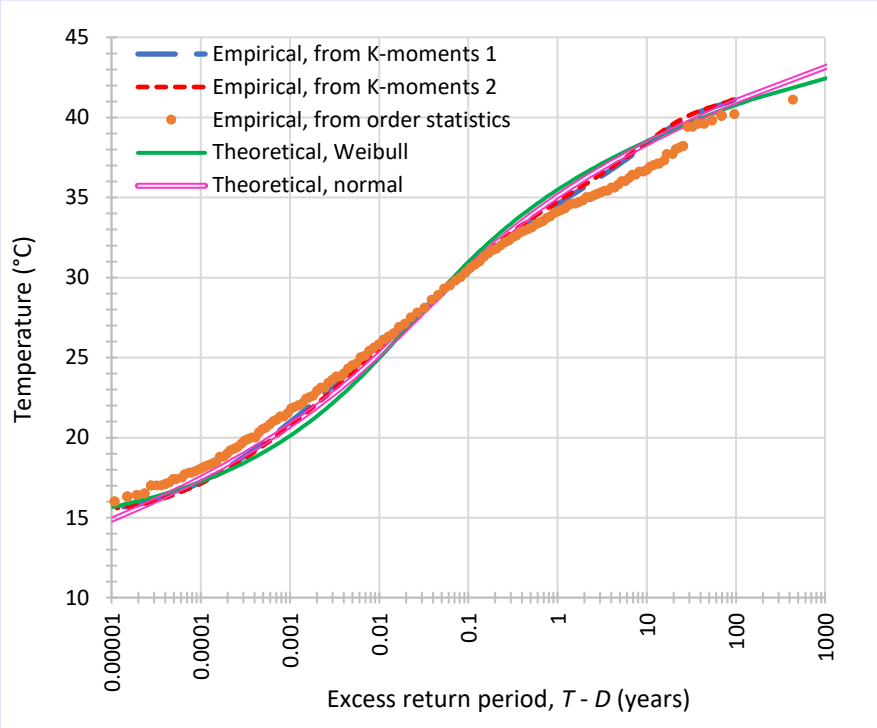

Figure 10.12 Comparison of empirical and theoretical Weibull and Normal distributions fitted on the daily maximum temperatures in August at Milano, Italy. The distribution function is depicted in terms of excess return period $T - D$. The parameters of the Weibull distribution are: $\zeta = 3.87, \lambda = 15.79$ °C, $x_L = 13.71$ °C and those of the fitted Normal distribution $\mu = 27.86$ °C, $\sigma = 3.83$ °C (the sample statistics are $\hat{\mu} = 28.13$ °C, $\hat{\sigma} = 3.33$ °C). For further explanations see caption of Figure 10.3.

The fitted Weibull distribution is in fact very close to the Normal. This is illustrated in Figure 10.12, which is similar to Figure 10.12 (left) but refers to the temperatures of the data of the hottest month, August, only. In Figure 10.12 both the Weibull and the Normal distributions have been fitted and they are very close to each other.

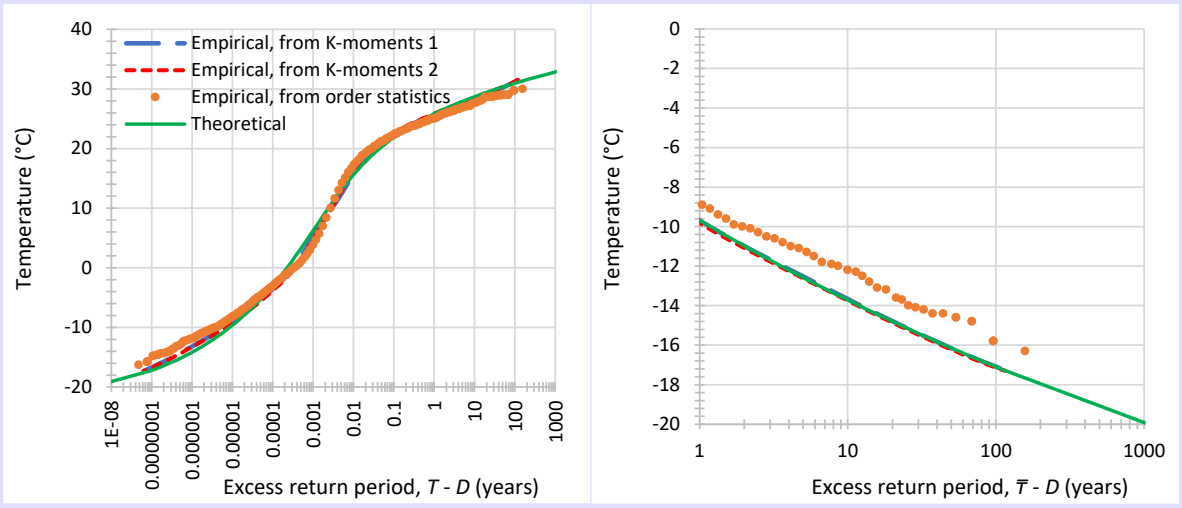

Figure 10.13 Comparison of empirical and theoretical Weibull distribution fitted on the daily minimum temperatures at Milano, Italy. The distribution function is depicted in terms of excess return period $T - D$, on their entire range **(left)** and in terms of the excess return period of minima $\overline{T} - D$, when focusing on $\overline{T} > 1$ year **(right)**. The parameters are: $\zeta = 5.80, \lambda = 36.28$ °C, $x_L = -23.45$ °C for the left panel and $\zeta = 12.93, \lambda = 39.06$ °C, $x_L = -34.43$ °C for the right panel. For further explanations see caption of Figure 10.3.

In turn, the fitting of the Weibull distribution on the daily minimum temperatures of Milano, is shown in Figure 10.13. Here again the distribution shape is close to Normal. The fitting for the

entire range of temperature (left panel) is better than that of maximum temperature (Figure 10.12, left). If we focus on the upper tail, for $T > 1$ year, the fitting again becomes perfect (right panel).

_____

[1] https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=urb_cpopcb&lang=en
[2] http://atlasofurbanexpansion.org/

10.3 Dew point

When studying the extreme events related to the water cycle, a variable more useful than atmospheric temperature is the dew point, defined to be the temperature at which the air must be cooled to become saturated with water vapour. The dew point is measured in the same units as temperature, and depends on the temperature on the one hand and on the presence of atmospheric moisture on the other hand. The relationship of these quantities is provided by the Clausius-Clapeyron equation, i.e. the law determining the equilibrium of liquid and gaseous phase of water, which maps temperatures to saturation vapour pressures. This law in essence describes an entropy-maximizing state, that is, a state where the uncertainty at a microscopic level becomes maximum, interestingly yielding a virtually deterministic law at the macroscopic level. While probability is typically used inductive reasoning, using data and statistics, this law serves as an example in which probability can also be used for deductive reasoning, with the impressive result described in Digression 10.D.

The dew point influences both the evaporation rate, which increases with an increasing departure thereof from temperature, as well storm intensity, which may increase with a larger dew point. It is natural to expect that, since temperature has been increasing in the recent decades, the dew point should have increased too. However, global reanalysis data (specifically from the ERA5 reanalysis*) show a much slower increase in the dew point (Koutsoyiannis, 2020b). Interesting related information about the zonal variation of the increase of temperature and dew point is provided by Figure 10.14, which depicts the difference of the Earth's temperature and dew point from their averages in the period 1980-99. A positive difference corresponds to an increase after 1999. It is important to note that the greater increases are located in the northern polar area. In the tropical zone, which is hydrologically most important as the main source of evaporated water, the temperature increase is half the global average, while there is no increase at all in the dew point. The latter point is of highest hydrological significance.

While this information is useful for global and zonal studies, local studies should be based on local data. Their analysis is not different from that of temperature and the general behaviours reported for rainfall in section 10.2 are generally valid also for the dew point. This is illustrated in Digression 10.E.

---

* The ERA5 (Copernicus Climate Change Service, 2017) is the fifth-generation atmospheric reanalysis of the European Centre for Medium-Range Weather Forecasts (ECMWF), where the name ERA refers to *ECMWF ReAnalysis*. It spans the modern observing period from 1979 onward, with daily updates continuing forward in time, with fields available at a horizontal resolution of 31 km on 139 levels, from the surface up to 0.01 hPa (around 80 km). It has been produced as an operational service and its fields compare well with the ECMWF operational analyses.

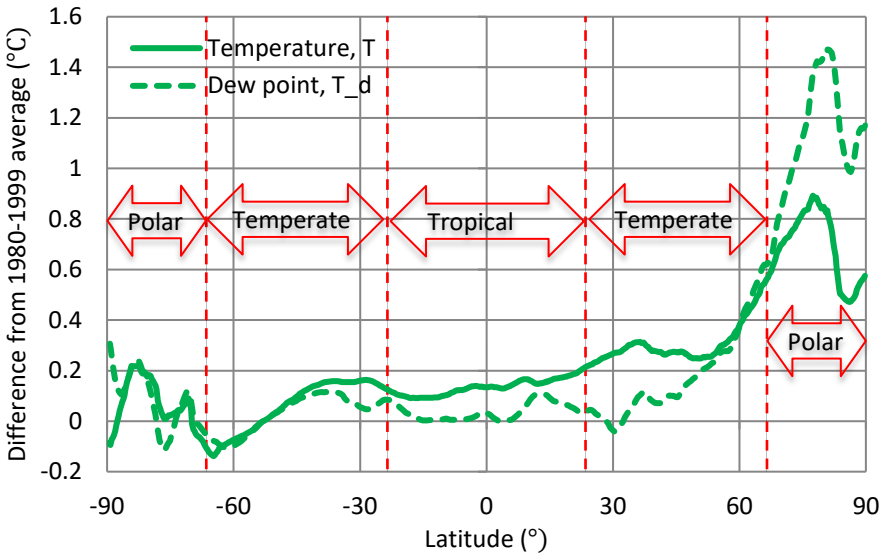

Figure 10.14 Zonal distribution of the difference of the earth temperature and dew point from their averages in the period 1980-99. Note that the graph represents averages for the entire 40+ year period, rather than differences between two periods (the latter are about twice the former). (Source: Koutsoyiannis, 2020b.)

Digression 10.D: How entropy maximization at a microscopic level results in a macroscopic deterministic law

Koutsoyiannis (2014a) has highlighted the probabilistic nature of the law that determines the equilibrium of liquid and gaseous phase of water by deriving it purely by maximizing probabilistic entropy, i.e. uncertainty. In particular, the law was derived by studying a single molecule (Figure 10.15) and maximizing the combined uncertainty of its state related to:

(a) its phase (whether gaseous, denoted as A, or liquid, denoted as B);
(b) its position in space; and
(c) its kinetic state, i.e., its velocity and other coordinates corresponding to its degrees of freedom and making up its thermal energy.

The partial entropies of the two phases, i.e., the entropies conditional on the particle being in the gaseous (A) or liquid (B) phase, are:

$$\varphi_{\mathrm{A}} = c_{\mathrm{A}} + (\beta_{\mathrm{A}}/2)\ln\varepsilon_{\mathrm{A}} + \ln V_{\mathrm{A}}\,, \qquad \varphi_{\mathrm{B}} = c_{\mathrm{B}} + (\beta_{\mathrm{B}}/2)\ln\varepsilon_{\mathrm{B}} + \ln V_{\mathrm{B}} \tag{10.1}$$

with c_i $(i = \mathrm{A}, \mathrm{B})$ denoting a constant (incorporating several physical and mathematical constants), β_i the degrees of freedom of a water molecule, ε_i the (thermal) energy of the water molecule and V_i the volume available for the motion of the water molecule in the specified phase. As the water molecule has a 3-dimensional (not linear) structure, the rotational energy is distributed into three directions, so that the total number of degrees of freedom (translational and rotational) is $\beta_{\mathrm{A}} = 6$. The number of degrees of freedom in the liquid phase is greater than 6 because of the "social behaviour" of water molecules. Specifically, in addition to the translational and rotational degrees of freedom of individual molecules, there are local clusters with low energy vibrational modes that can be thermally excited. The average number of degrees of freedom per molecule (individual and collective involving more than one water molecules) is very high, $\beta_{\mathrm{B}} = 18$.

The total entropy is:

$$\varphi = \pi_{\mathrm{A}}\varphi_{\mathrm{A}} + \pi_{\mathrm{B}}\varphi_{\mathrm{B}} + \varphi_{\pi} \tag{10.2}$$

where π_i is the probability that the molecule is at phase i, with corresponding entropy:

$$\varphi_\pi \coloneqq -\pi_\mathrm{A} \ln \pi_\mathrm{A} - \pi_\mathrm{B} \ln \pi_\mathrm{B} \tag{10.3}$$

Thus, the total entropy can be written as:

$$\varphi = \pi_\mathrm{A}(\varphi_\mathrm{A} - \ln \pi_\mathrm{A}) + \pi_\mathrm{B}(\varphi_\mathrm{A} - \ln \pi_\mathrm{A}) \tag{10.4}$$

The two phases are in open interaction and the constraints are:

$$\pi_\mathrm{A} + \pi_\mathrm{B} = 1, \qquad \pi_\mathrm{A}\varepsilon_\mathrm{A} + \pi_\mathrm{A}(\varepsilon_\mathrm{A} - \xi) = \varepsilon \tag{10.5}$$

where ξ is the amount of energy required for a molecule to move from the liquid to gaseous phase (i.e. to break its bonds with other molecules, the phase change energy).

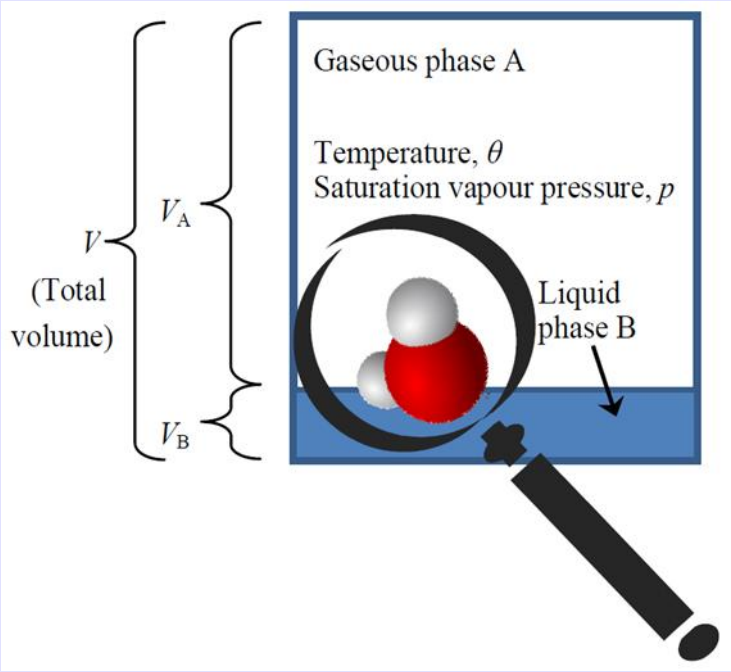

Figure 10.15 Explanatory sketch indicating basic quantities involved in the equilibrium of the water vapour with liquid water, with zoom on a single molecule which "tries to hide itself" by maximizing the combined uncertainty related to its phase (being either gaseous or liquid with probabilities π_A and π_B, respectively), position and kinetic state.

By defining the *natural temperature*, θ, with units of energy (joules) rather than temperature (kelvins), in accordance to the probabilistic principle that entropy is a dimensionless quantity φ, as:

$$\frac{1}{\theta} \coloneqq \frac{\partial \varphi}{\partial \varepsilon} \tag{10.6}$$

Denoting e the partial pressure of the N_A water molecules being in the gaseous phase and maximizing the entropy in that phase, we obtain the law of ideal gases in the form (Koutsoyiannis, 2014a):

$$e = \frac{N_\mathrm{A}\theta}{V_\mathrm{A}} = \frac{\theta}{v} \Leftrightarrow ev = \theta \tag{10.7}$$

where $v \coloneqq V_\mathrm{A}/N_\mathrm{A}$.

Furthermore, by maximizing the combined entropy of the two phases, as given in equation (10.4), we obtain the law of the equilibrium of the two phases as (Koutsoyiannis, 2014a):

$$e = e_0 \exp\left(\frac{\xi}{\theta_0}\left(1 - \frac{\theta_0}{\theta}\right)\right)\left(\frac{\theta_0}{\theta}\right)^{\beta_\mathrm{B}/2 - \beta_\mathrm{A}/2 - 1} \tag{10.8}$$

where (θ_0, e_0) are the coordinates of the triple point of water (specifically, θ_0 = 37.714 yJ corresponding to T_0 = 273.16 K, e_0 = 6.11657 hPa).

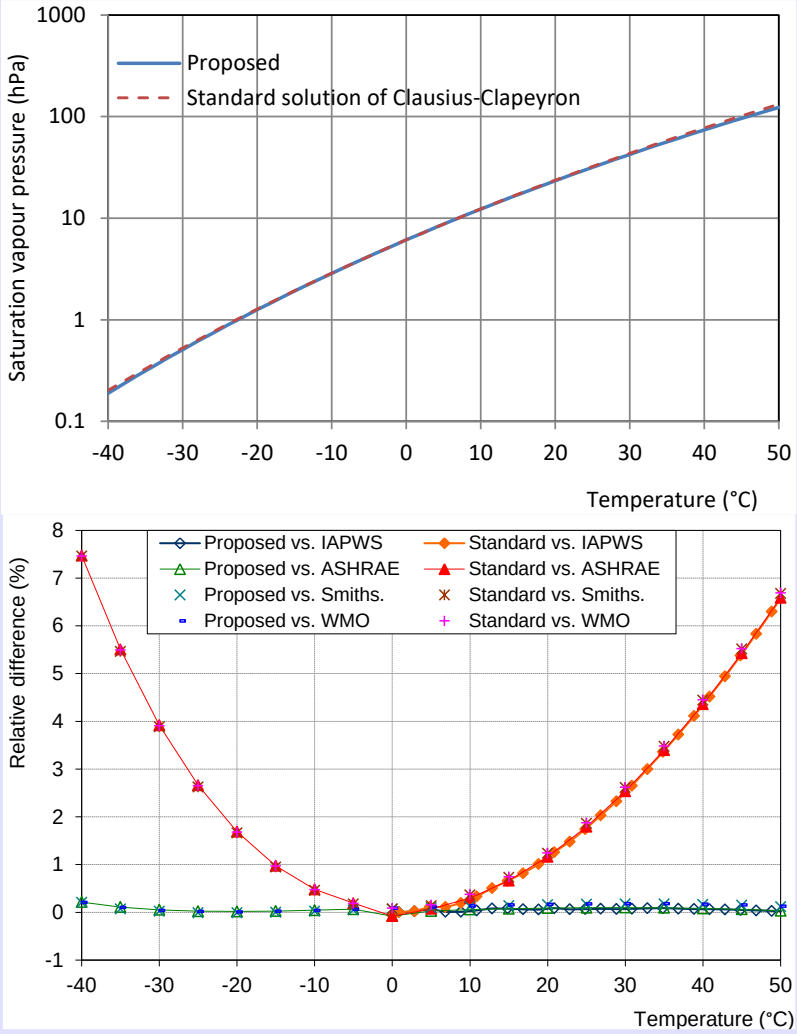

Figure 10.16 (upper) Comparison of saturation vapour pressure obtained by the proposed equation in either of the forms (10.8) or (10.10) and by a standard equation of the literature, namely, $e = e_0 \exp(19.84(1 - T_0/T))$. **(lower)** Comparison of relative differences of the saturation vapour pressure obtained by the proposed and the standard equations with accurate measurement data of different origins, as indicated in the legend and detailed in Koutsoyiannis (2012).

The same law (also known as Clausius-Clapeyron equation in integrated form) can be written in more customary notation, in terms of absolute temperature in kelvins and using macroscopic quantities, as (Koutsoyiannis, 2012):

$$e = e_0 \exp\left(\frac{\alpha}{RT_0}\left(1 - \frac{T_0}{T}\right)\right)\left(\frac{T_0}{T}\right)^{(c_L - c_p)/R} \tag{10.9}$$

where (T_0, e_0) are again the coordinates of the triple point of water, R is the specific gas constant of water vapour (R = 461.5 J kg^{-1} K^{-1}, $\alpha := \xi R/k = \xi N_a$ (with k the Boltzmann's constant and N_a the Avogadro constant), c_p is the specific heat at constant pressure of the vapour and c_L is the specific heat of the liquid water. By substitution of the various constants we end up with the following form of the equation (Koutsoyiannis, 2012):

$$e := e(T) = e_0 \exp\left(24.921\left(1 - \frac{T_0}{T}\right)\right)\left(\frac{T_0}{T}\right)^{5.06} \tag{10.10}$$

This form is both convenient and accurate (more accurate than other customary forms, theoretical or empirical, as illustrated in Figure 10.16).

A state in which the vapour pressure e_a is lower than the saturation pressure $e(T)$ is characterized by the relative humidity:

$$U := \frac{e_a}{e(T)} = \frac{e(T_d)}{e(T)} \tag{10.11}$$

which serves as a formal definition of both the relative humidity U and the dew point T_d. In particular the dew point is calculated by the following equation, direct result of (10.11):

$$T_d = e^{-1}(U\, e(T)) \tag{10.12}$$

where $e^{-1}(\,)$ is the inverse function of $e(\,)$; its numerical handling is discussed in Koutsoyiannis (2012). As in equilibrium the maximum U is 1, it results that T is an upper threshold of T_d.

Digression 10.E: An example of fitting a distribution on daily dew point data

The dew point can be derived by equation (10.12) from measurements of temperature and relative humidity, but can also be measured by devices called hygrometers. Data from raw measurements are not quite frequent, yet the KNMI database provides data of daily maximum dew points in 35 stations in the Netherlands. The station with the longest record is De Bilt (52.101°N, 5.177°E, 2.0 m). Its data cover the period January 1951 to May 2018 (66 calendar years with the entire 1960 missing and a few missing values in other years). The number of daily values is $n = 24\,561$, 2.6% of which are below 1 °C, while no negative values are contained. Characteristic plots of the time series are given in Figure 10.17.

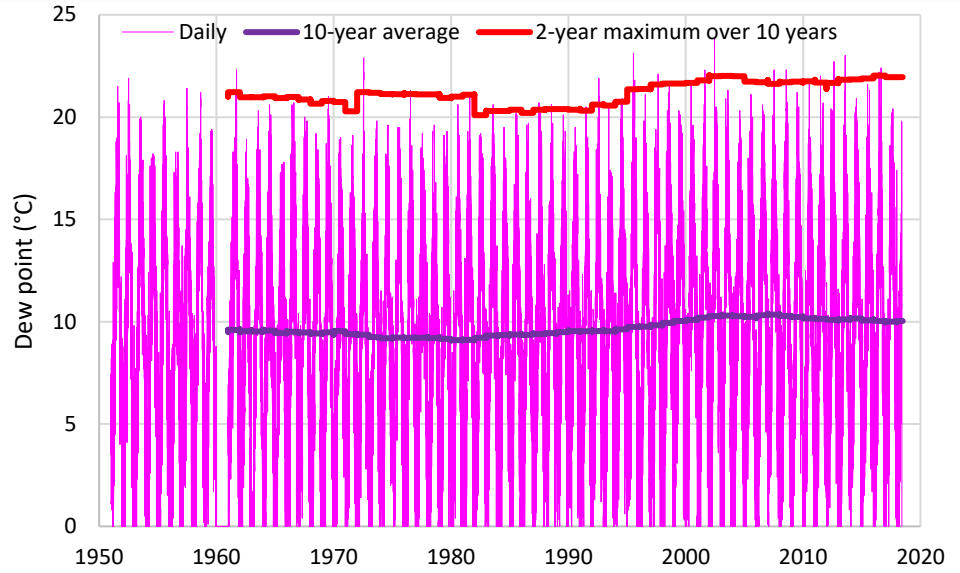

Figure 10.17 Time series of daily maximum dew point at De Bilt, Netherlands, along with the running, on 10-year windows, average and maximum value for return period of 2 years.

According to equation (10.12) and Figure 10.16, negative values would have been expected, but they cannot be directly measured. Therefore, the smallest of the values (<1 °C), while were kept in the climacogram analysis, were not used in the distribution fitting.

Focusing on the running maximum for return period of 2 years, which is also plotted in Figure 10.17, we may observe some fluctuation with a decreasing trend before 1990 and an increasing one thereafter. On may thus speculate that the global warming caused temperature increase in De Bilt which in turn drifted the dew point, because of its positive correlation with temperature. On the other hand, one should not forget the possible effect of urbanization on temperature, also having in mind the fact that The Netherlands constitute one of the most urbanized areas in Europe and in the world (Figure 10.18).

Figure 10.18 Depiction of urbanization in Europe by comparing satellite images of night lights of **(upper)** 2000 and **(lower)** 2012. Image source: https://www.nightearth.com/. The 2000 image was created by NASA using data from the Defense Meteorological Satellite Program (DMSP)'s Operational Linescan System (OLS), originally designed to view clouds by moonlight. The 2012 image was captured by NASA using the Suomi National Polar-orbiting Partnership (Suomi NPP) satellite during April and October 2012 (from the "day-night band" of the Visible Infrared Imaging Radiometer Suite—VIIRS, which detects light in a range of wavelengths from green to near-infrared).

Therefore, avoiding attribution attempts, here we merely investigate the stochastic properties of the time series and try to build a consistent stationary stochastic representation with

persistence. Indeed, as shown in Figure 10.19, the climacograms suggest a high value of the Hurst parameter, estimated from the annual series at $H = 0.89$.

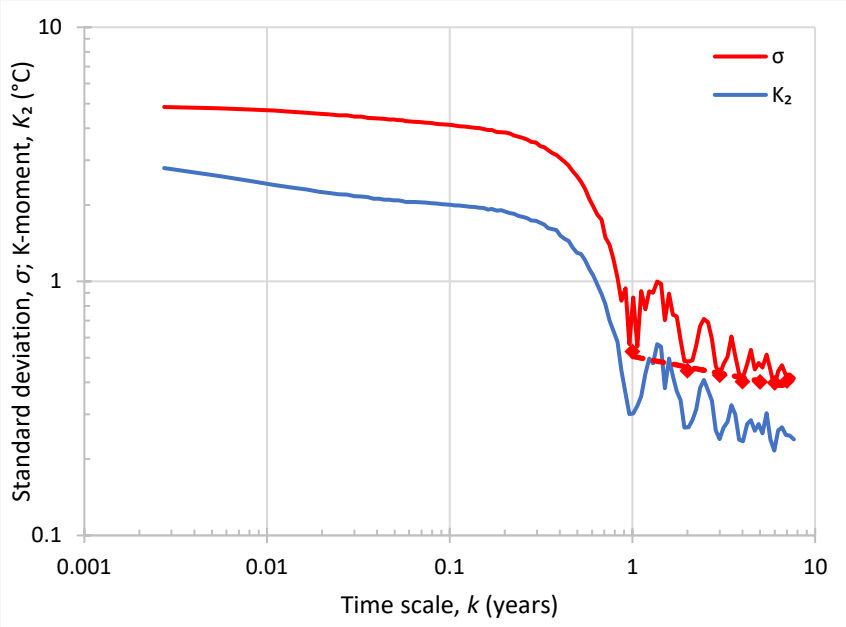

Figure 10.19 Empirical σ-climacogram and K_2-climacogram of daily maximum dew point at De Bilt, Netherlands. The lines have been constructed from the daily series and the points from the annual series. The points corresponding to the annual series suggest a power law with exponent –0.14. The Hurst parameter, estimated from the annual series, is $H = 0.89$.

The fitting of the Weibull distribution on the daily maximum dew point, performed by the procedure described in Digression 9.B, is shown in Figure 10.20. Specifically, a on the entire domain of dew point except for the values < 1°C for the reasons explained above. The high value of the shape parameter ζ (close to 10) suggests a distribution shape close to Normal.

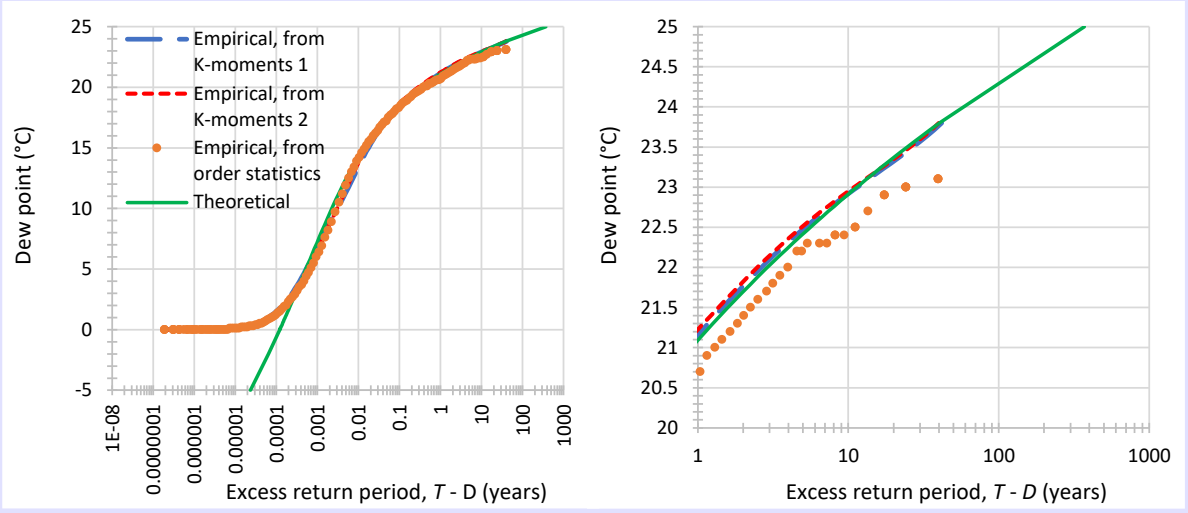

Figure 10.20 Comparison of empirical and theoretical Weibull distribution fitted on the daily dew point at De Bilt, Netherlands. The distribution function is depicted in terms of the excess return period $T - D$, on its entire range **(left)** and focusing on $T - D > 1$ year **(right)**. The parameters are: $\zeta = 9.98, \lambda = 45.42$ °C, $x_L = -34.43$ °C and were fitted on the entire domain of dew point except for the values < 1°C. For further explanations see caption of Figure 10.3.

Chapter 11. Epilogue: Technology for risk reduction

As already mentioned (section 5.3), the notion of the risk incorporates three factors: the probability of occurrence of a dangerous event, the exposure and the vulnerability (Kron et al. 2019). More formally, the risk is usually defined as the product of three variables:

$$R = H\,E\,V \tag{11.1}$$

which have the following meaning:

- H is the *hazard*, i.e., the occurrence probability of a dangerous event (unit: dimensionless).
- E denotes the *exposure*, i.e. the "values" that are exposed to a dangerous event. These values are usually expressed in monetary terms (unit: e.g. $ or €) representing the economic value of the objects that are present at the location involved. In its severest form, E represents human lives (unit: dimensionless).
- V is the vulnerability, i.e. the lack of resistance to damaging or destructive forces, expressed as a value between 0 and 1 (unit: dimensionless), with the highest value 1 representing full damage of the exposed values.

There is a single means to control any of this variables: technology. Considering as an example the flood risk at a specific location, we can reduce the hazard by several technological solutions, e.g., by building a dam upstream of that location—this, however, is not possible for other types of hazards, e.g. earthquakes. We can also reduce the vulnerability by modelling the flood extent, delignating flood-prone zones and then implementing urban planning to prohibit or discourage human settlements at those zones (this applies to all types of hazards). Finally, we can reduce vulnerability, e.g., by developing early warning systems (Di Baldassarre et al., 2010).

The risk is an objective quantity that should be distinguished from its perception. The latter is determined by other factors or interests: political, economic or social. For example, as a result of intensification of reporting of occurring disasters and projections of future catastrophes, people think that the risk from hydroclimatic extremes has been radically increasing. As we will see in the next subsections, this is just a social perception, in fact opposite to reality.

11.1 Is hydroclimatic hazard increasing?

A first indication supporting a negative reply to the question in the subsection's title is provided by a list of world record point precipitation measurements compiled by Koutsoyiannis and Papalexiou (2017) for various time scales ranging from 1 min to 2 years. As can be seen in Figure 11.1, reproduced from that study, the fact is that the highest frequency of record rainfall events occurred in the period 1960-80; later the frequency was decreased remarkably.

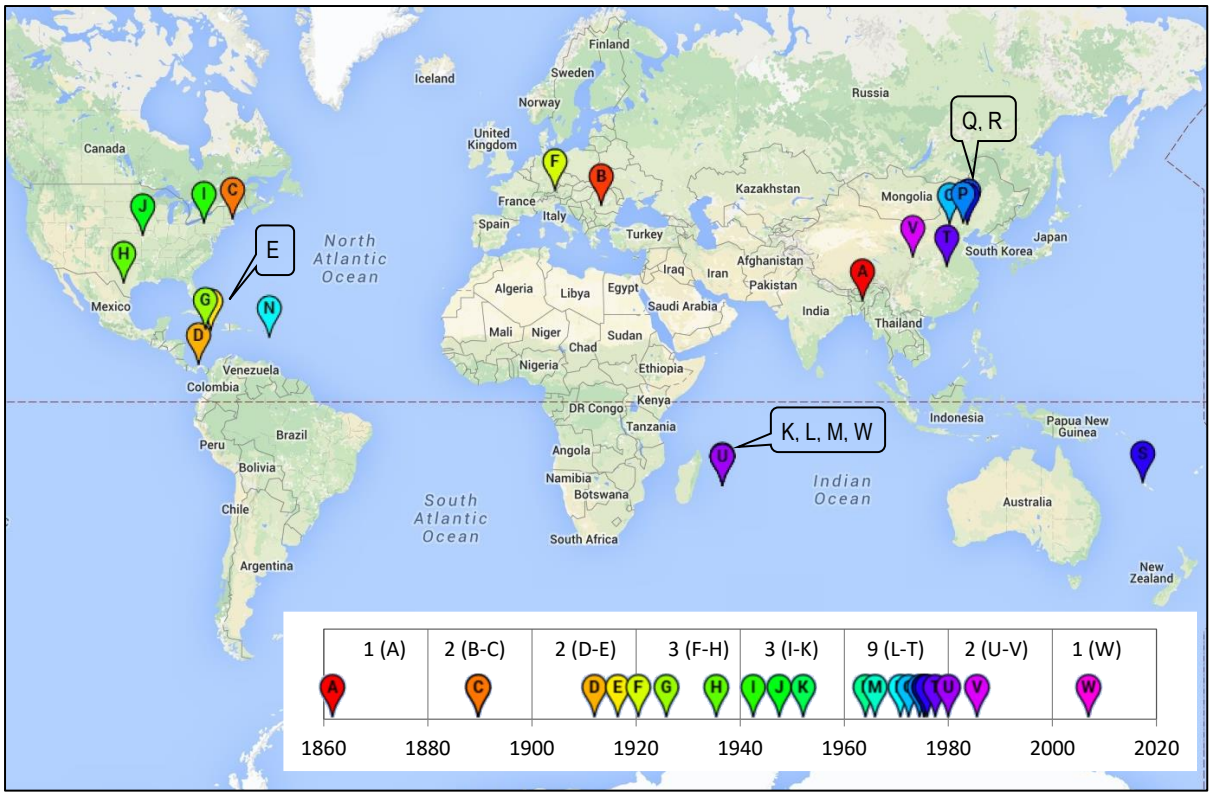

Figure 11.1 World record point precipitation measurements (locations and time stamps of the events producing record rainfall) for time scales ranging from 1 min to 2 years, compiled in Koutsoyiannis and Papalexiou (2017). The time scales on which the different events have given the record rainfall are as follows: A – 1 month to 2 years; B – 20 min; C – 2.17 h; D – 5 min; E – 15 min; F – 8 min; G – 3 min; H – 2.75 h; I – 3 h; J – 42 min; K – 1 d; L – 9 h; M – 18 h; N – 1 min; O – 2.5 h; P – 30 min; Q – 1 h; R – 2 h; S – 2 d; T – 6 h; U – 9 d to 15 d; V – 72 min; W – 2 d to 7 d.

A more detailed analysis, again on global basis, has been provided by Koutsoyiannis (2020a), based on reanalysis and satellite data of daily rainfall. Analyses of precipitation maxima have been made to test the allegation by IPCC (2013a) about an intensification of the hydrological cycle and the related extremes. Notably, the intensification allegation, if quantified, turns out to be of the order of 1%– 5% (IPCC, 2013a; Koutsoyiannis, 2020b,c). Such percentages of change are negligible and rather non detectable, given the high variability of precipitation articulated in previous sections. Nonetheless, here we reproduce some of the results of Koutsoyiannis (2020b) for the incredulous reader. The analyses have been performed separately for each continent and their results are presented graphically. Figure 11.2 shows the temporal evolution of the monthly maximum daily precipitation areally averaged over the continents. None of the sources of data in none of the continents provides support on the intensification allegation. In particular, the observational data (CPC and GPCP) could support the opposite hypothesis, that of extreme rainfall deintensification.

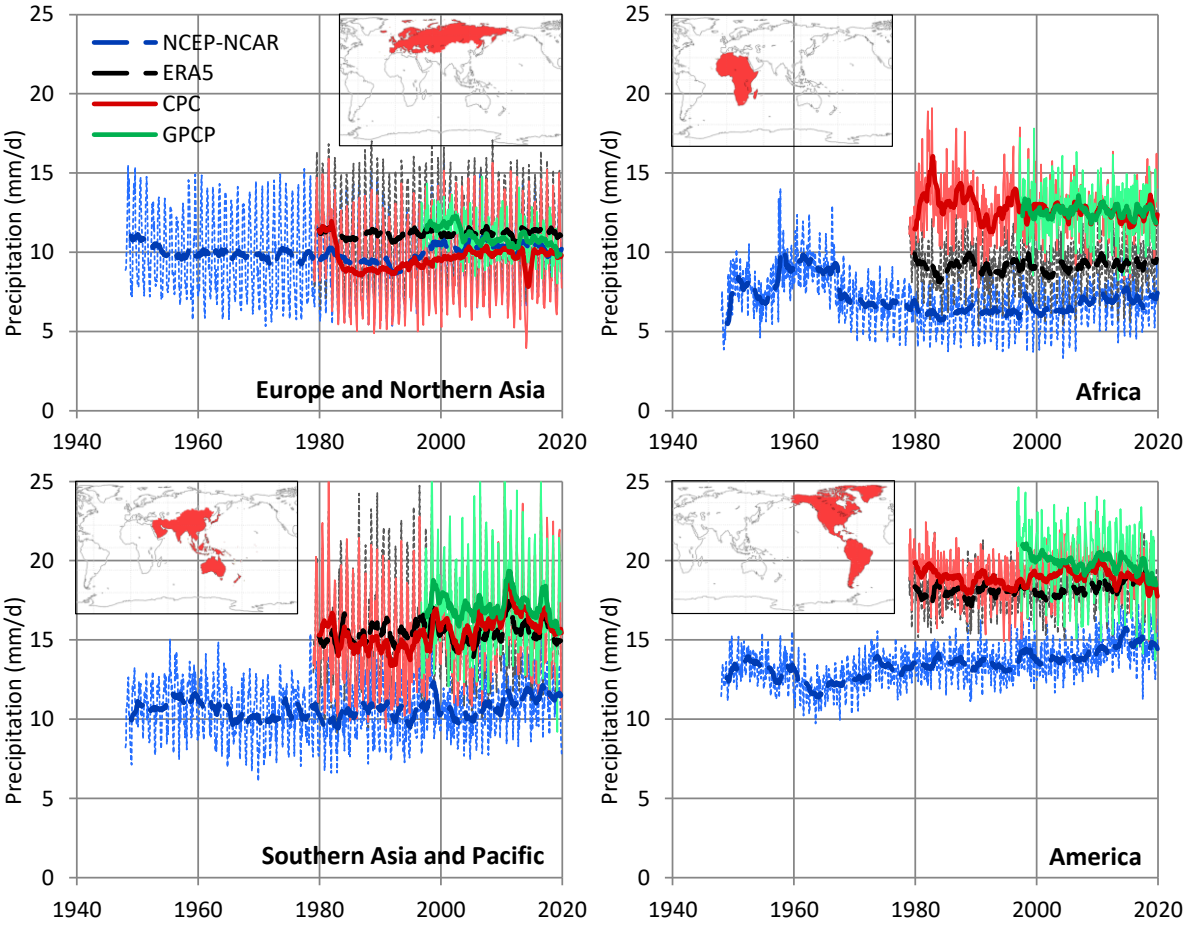

Figure 11.2 Variation of the monthly maximum daily precipitation areally averaged over the continents. Thin and thick lines of the same colour represent monthly values and running annual averages (right aligned), respectively. Dashed lines are for reanalyses and continuous lines for observations. (Source: Koutsoyiannis, 2020b; NCEP-NCAR: reanalysis data; ERA5: reanalysis data; CPC: unified gauge-based daily precipitation gridded over land; GPCP: precipitation data set combining gauge and satellite precipitation data over a global grid.)

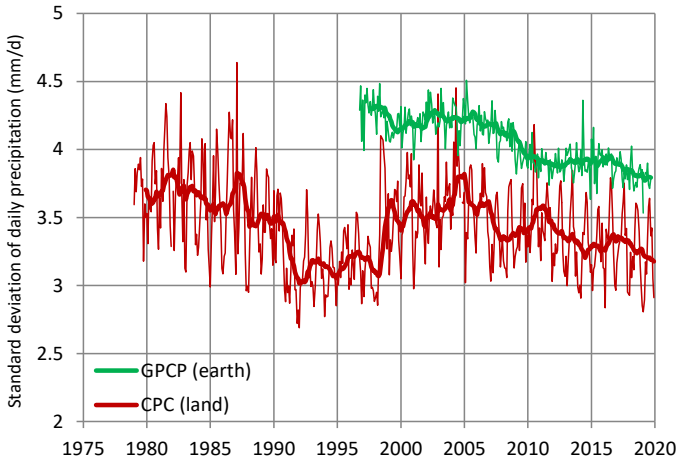

Figure 11.3 Variation of the standard deviation of daily precipitation in each month, areally averaged. Thin and thick lines of the same colour represent monthly values and running annual averages (right aligned), respectively. (Source: Koutsoyiannis, 2020b; GPCP: precipitation data set combining gauge and satellite precipitation data over a global grid; CPC: unified gauge-based daily precipitation gridded over land.)

Deintensification becomes even more evident if we examine the temporal evolution of standard deviation of daily precipitation in each month, averaged over land. In this respect, Figure 11.3, shows that deintensification, expressed as decreasing standard deviation, is evident in the 21st century both from CPC and GPCP observational data. This finding is consistent with earlier findings by Sun et al. (2012). A similar result is shown in a different manner in Figure 11.4 in terms of precipitation rate exceeding a threshold. Clearly, neither the frequency of high precipitation nor the sum of high intensity precipitation is intensifying. Rather, in most of the cases, there has been deintensification in the 21st century. Again, however, it will be more prudent to speak about fluctuations rather deintensification. This is consistent with the general approach in this book to use stationary models (with appropriate dependence structure) for extremes, a suggestion also made in other works (Koutsoyiannis, 2003, 2006b, 2011a; Montanari, and Koutsoyiannis, 2014; Koutsoyiannis and Montanari, 2015; De Luca et al., 2020).

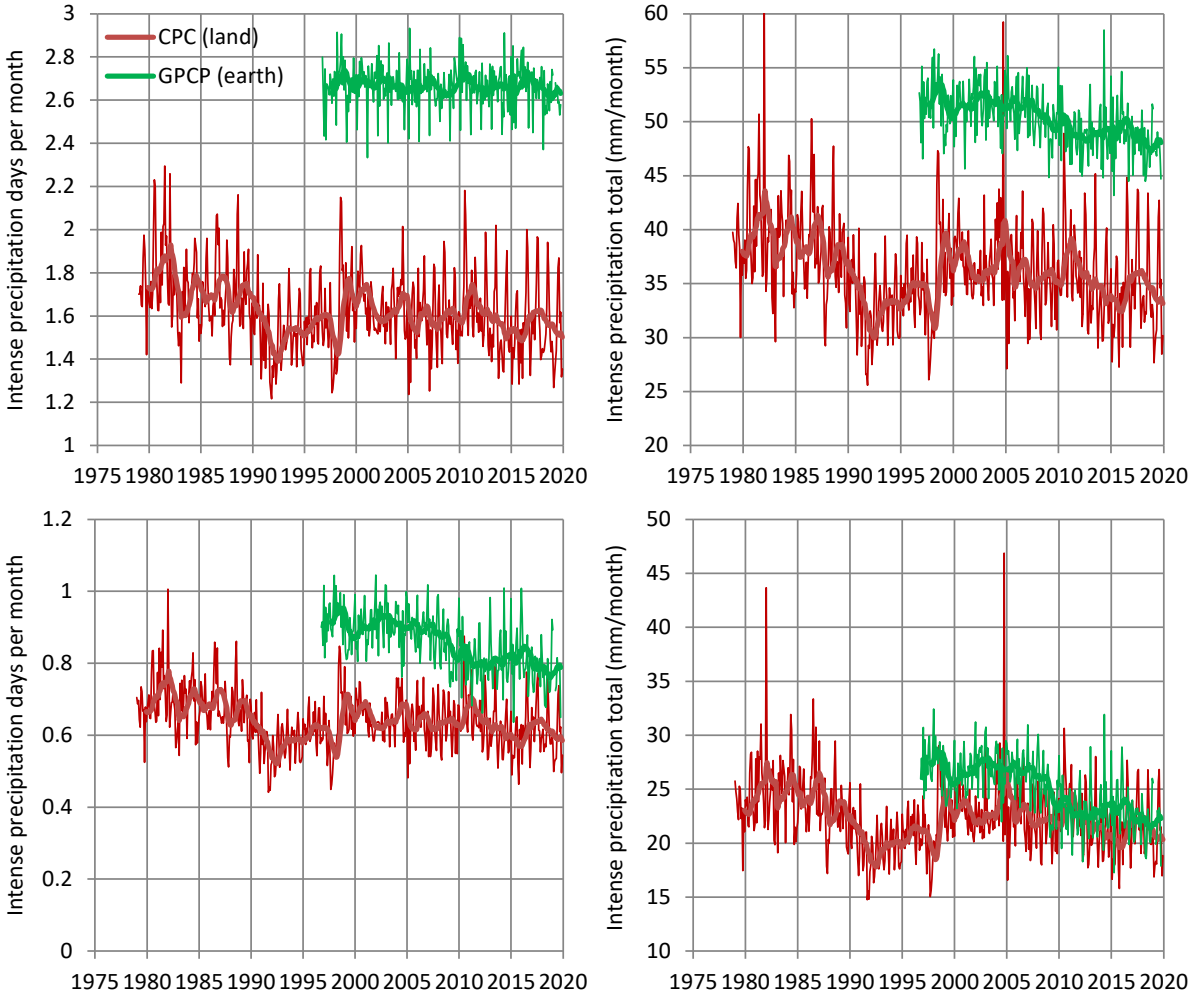

Figure 11.4 (left column) Average days per month with precipitation exceeding a threshold value, which is 10 mm/d **(upper row)** and 20 mm/d **(lower row)**; **(right column)** monthly total of daily precipitation exceeding the threshold value. Thin and thick lines of the same colour represent monthly values and running annual averages (right aligned), respectively.

For examining the floods, we use the already mentioned (section 9.1) database of the US Geological Survey's (USGS) National Water Information System and in particular the registered annual peaks by Hirsch and Ryberg (2012) in 200 stream gauges in the coterminous USA in pristine or near-pristine catchments, of at least 85 years length through water year 2008. Figure 11.5 depicts the frequency of a record high flood per decade, obtained from this database. The annual average (= 0.0109 events per year) and the confidence limits on decadal basis are also plotted in the figure. Fluctuations are evident in the time evolution, with low flood occurrences in the 1960s and high in the 1900s and 1990s, with highest ones in the 1890s. Only in the 1890s was the frequency of record high floods higher than the upper confidence limit. Even this very old event would not be an issue of concern: first because an exceedance from the 95% confidence area in one out of 13 decades is not unnatural and second because if we considered the dependence structure and performed Monte Carlo simulations to determine the confidence limits, the confidence zones would be wider.

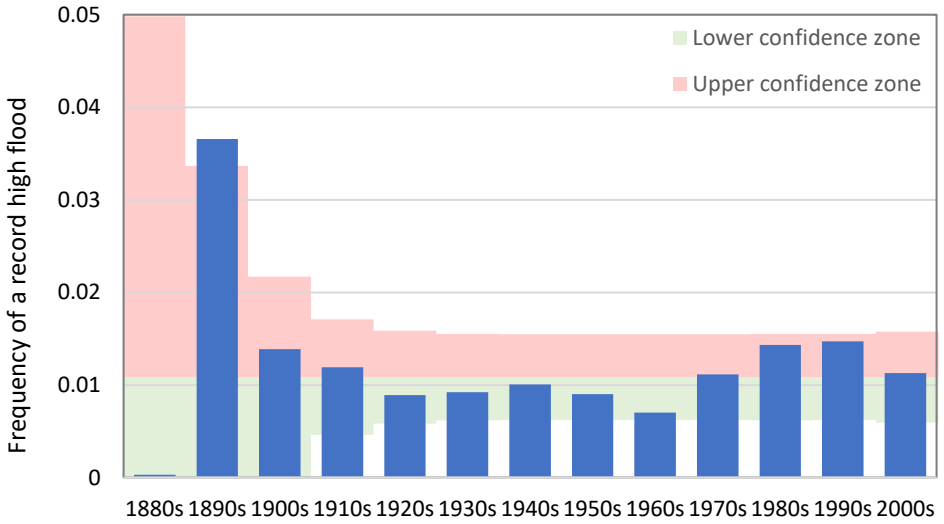

Figure 11.5 Frequency of a record high flood per decade, obtained from the database by Hirsch and Ryberg (2012) from 200 selected stream gauges in pristine or near-pristine catchments in the coterminous USA. The database contains 18 846 station years and the record highs are 205 (> 200 because of some ties) and thus the average probability of a record high is 205/18 846 = 0.0109. The confidence limits are approximate, constructed for confidence level of 95% on the basis of independence (Papoulis, 1990, p. 284).

About the possible intensification of the wind field over the globe, some information is provided in Figure 11.6 in terms of the global maximum wind speed, zonal and meridional. The plots do not show any noteworthy change (e.g. trend). Only slight fluctuations appear. Thus, the regime shown does not justify intensification of wind or of precipitation extremes that the latter could induce.

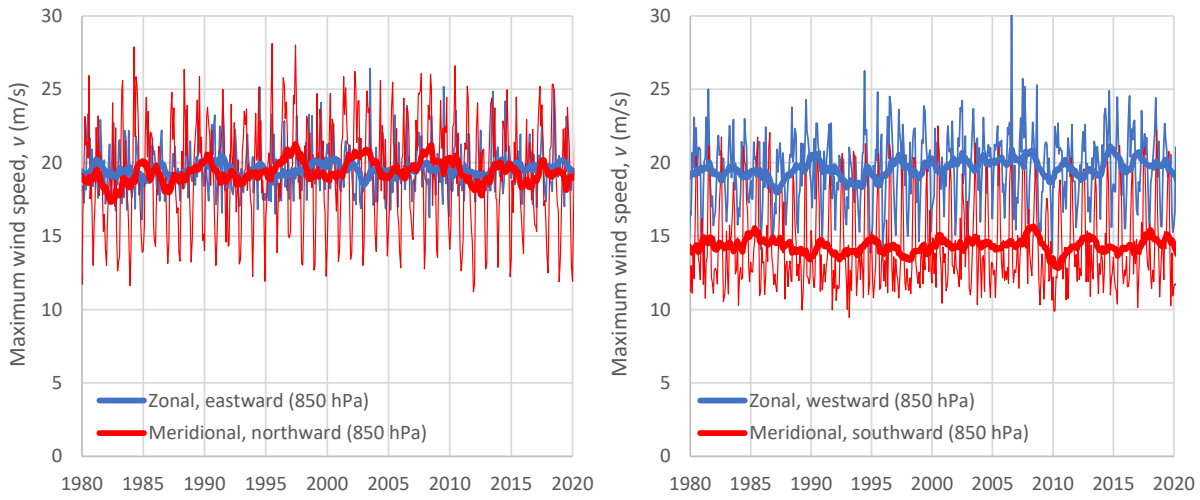

Figure 11.6 Variation of the monthly maximum daily wind speed, in each of the four directions (two zonal and two meridional) over the globe. Thin and thick lines of the same colour represent monthly values and running annual averages (right aligned), respectively. (Data source: ERA5 reanalysis data retrieved from Climexp.)

11.2 Do background conditions favour enhancement of hydroclimatic extremes?

If we focus on storms and floods and consider the wind regime as a background condition that influences them, then, as we have already seen, there are no changes that would seem to favour intensification. Another possible background condition affecting precipitation extremes is atmospheric moisture. An extensive study thereof has been recently presented in Koutsoyiannis (2020b), from which we reproduce here a small part of the results, referring to the *water vapour amount* (also known as *vertically integrated water vapour*, or *precipitable water*[*] and expressed in mm or equivalently kg/m^2). This is estimated by radiosonde data of temperature and relative humidity on a local basis, but on a global basis, which is of interest here, it can be either estimated by reanalysis data or provided by satellite data. In all cases, the water vapour amount is the mass of water vapour, integrated over the entire atmosphere, per unit area. An increased water vapour amount could potentially lead to increased storm severity.

However, the data, plotted in Figure 11.7 do not show such a case. Specifically, the data originating from two reanalyses, ERA5 and NCEP-NCAR[†], which agree impressively well to each other, indicate fluctuation over time, with no monotonic trend. One of the two satellite data sets (NVAP) also agrees on the average, indicating no trend. However, the most recent satellite data set (MODIS) suggests a decreasing trend, just the opposite of

[*] The adjective *precipitable* for the water vapour amount is a misnomer: if the total water vapour amount in the atmosphere was indeed to precipitate in its entirety, this would violate the laws of thermodynamics.

[†] The NCEP-NCAR reanalysis is jointly produced by the National Center for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR). Its temporal coverage includes 4-times daily, daily and monthly values for 1948 to present at a horizontal resolution of 1.88° (~ 210 km). It uses a state-of-the-art analysis/forecast system to perform data assimilation using observations and a numerical weather prediction model. The data assimilation and the model used are identical to the global system implemented operationally at NCEP except in the horizontal resolution.

the IPCC predictions. In addition, Figure 11.8, which provides layered information for the MODIS data, shows that the decreasing trend is more pronounced in the upper atmospheric levels (440 to 10 hPa).

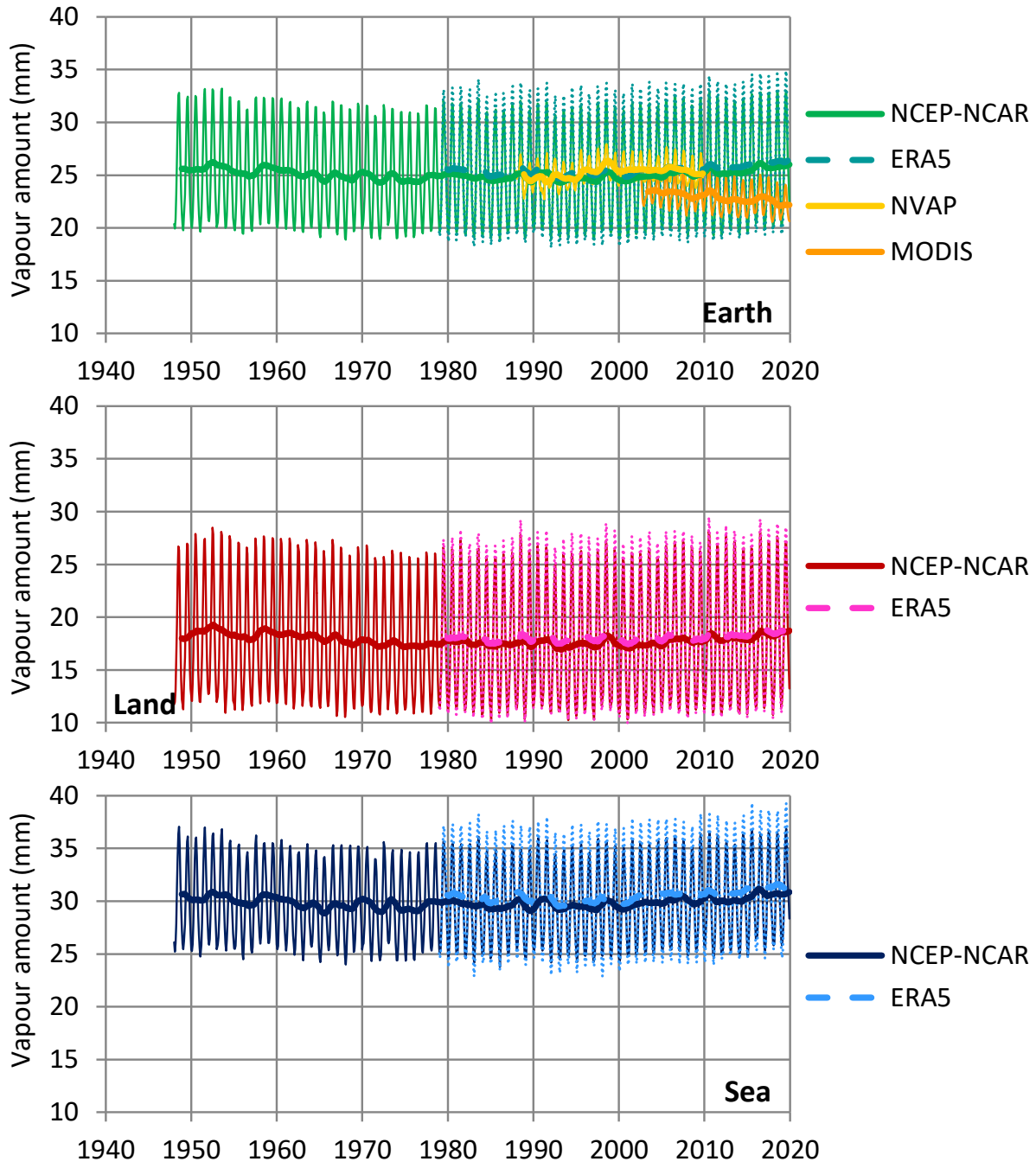

Figure 11.7 Variation of water vapour amount over the globe and the land and sea areas. Thin and thick lines of the same colour represent monthly values and running annual averages (right aligned), respectively. (Source: Koutsoyiannis, 2020b; NCEP-NCAR: reanalysis data; ERA5: reanalysis data; NVAP: satellite data from the NASA Pathfinder project; MODIS: satellite data, averages from the MODIS-Terra & MODIS-Aqua satellites.)

In other words, as far as hydrological extremes are concerned, observations do not show any changes in the background conditions that would favour occurrence of more frequent or more intense extremes.

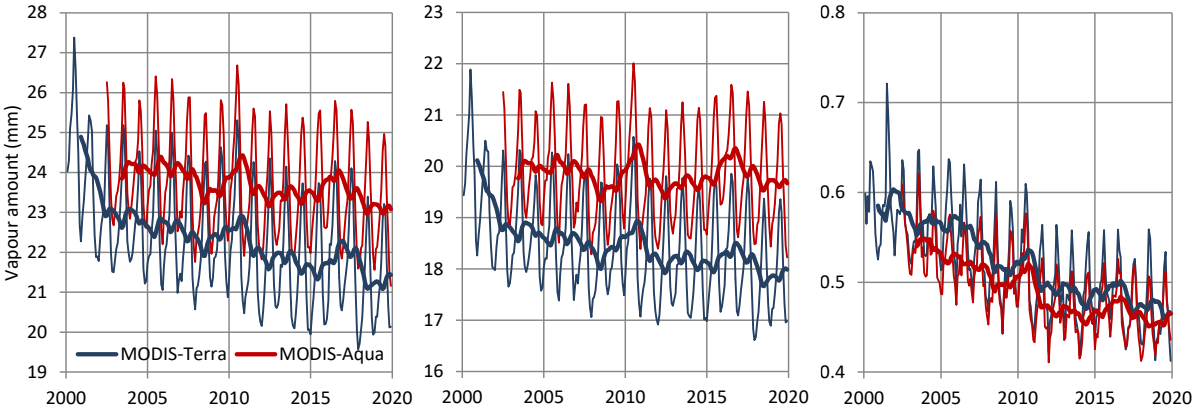

Figure 11.8 Variation of water vapour amount as in Figure 11.7 but only for the MODIS satellite data set and separately for its two platforms, Terra and Aqua: **(left)** total of the vertical column; **(middle)** from surface to 680 hPa; **(right)** from 440 to 10 hPa. Thin and thick lines of the same colour represent monthly values and running annual averages (right aligned), respectively. (Source: Koutsoyiannis, 2020b.)

11.3 Is the risk from hydroclimatic extremes increasing?

If we try to approach changes in the risk from extremes, including the influence of exposure and vulnerability, the ultimate measure of the risk is the number of deaths from natural disasters. Relevant data are shown in Figure 11.9 for all natural disasters classified into five categories, three of which are of hydroclimatic type.

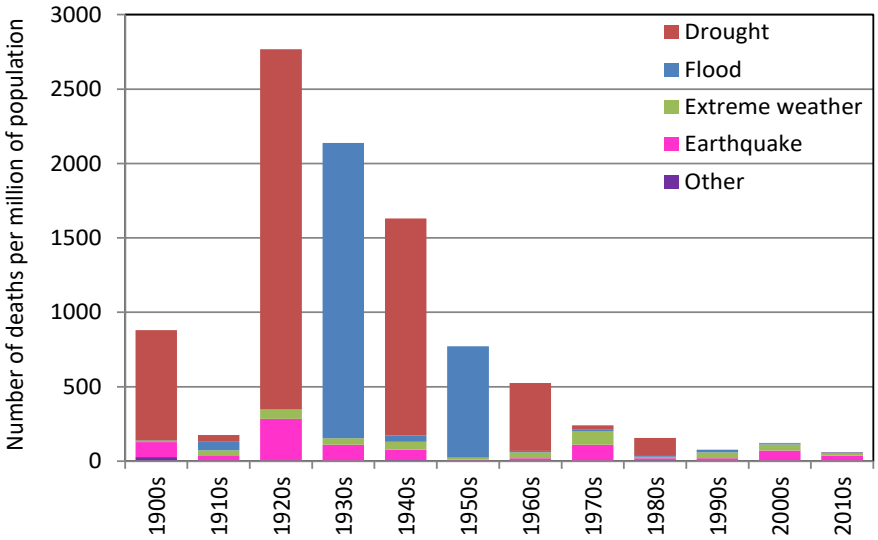

Figure 11.9 Evolution of the frequency of deaths from natural disasters per decade in the 20th and 21st century. In addition to deaths from loods and droughts, deaths from other categories of natural catastrophes are also plotted: "extreme weather" includes storms, extreme temperatures (cold- or heatwave, severe winter conditions) and fog; "earthquake" also includes tsunamis; "other" comprises landslides (wet or dry), rockfalls, volcanic activity (ash fall, lahar, pyroclastic flow and lava flow) and wildfires. (Source: Koutsoyiannis, 2020b; victim data: OFDA/CRED International Disaster Database*; population data: United States Census†.)

* https://ourworldindata.org/ofdacred-international-disaster-data

† https://www.census.gov/data-tools/demo/idb/informationGateway.php

Clearly, the impacts of the disasters of hydroclimatic type, particularly the severest of them which caused human victims, have spectacularly dropped since the beginning of the 20th century. The victims from these categories of disasters have been diminished, while other types still cause large numbers of victims. Thus, in the 2010s the primary cause is earthquake, representing 59% of the total number of victims. Obviously, the reason behind such diminishing is not that floods and droughts have become less severe or less frequent. Rather it is the improvement of technology, and risk assessment, management and reduction, along with the strengthening of the international collaboration and the economy, so that the advances could be actually implemented.

Interestingly, according to data of 2010-2017, the deaths from natural disasters represent 0.08% of the total number of deaths, as seen in Figure 11.10. This number ranks them last in the last position in Figure 11.10, with the penultimate cause being cold and heat. Deaths from cold and heat are registered together, while a multi-country analysis by Gasparrini (2015) suggests that these are mostly (at 95%) due to cold. For comparison, the contribution to deaths of respiratory diseases (belonging in the broader category of health issues) is 11.6%, about 150 times higher than natural disasters (and, apparently, this figure should have now increased due the Covid 19). The curious reader is encouraged to try to trace the reasons why the general perception of the public, informed by the media, is inverse to reality. Also, why the climate related risks, the least severe of all, have been promoted so enormously by international organizations and philanthropists.

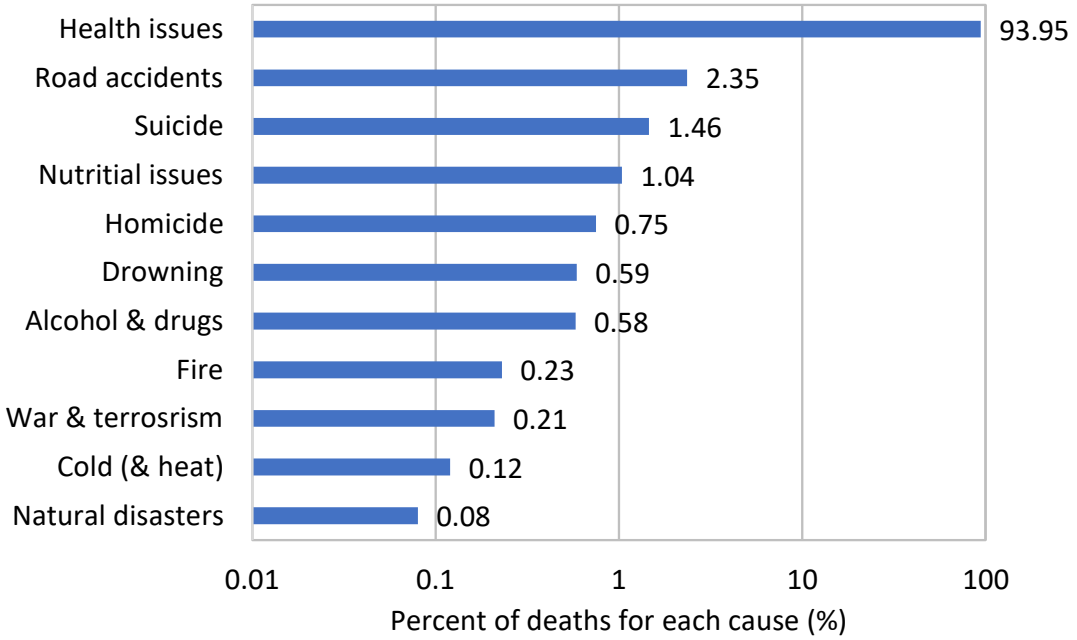

Figure 11.10 Average share of deaths per cause in the 2010s. Data from the database of Our World in Data*; note that the total is slightly greater than 100% (101.4%, perhaps suggesting that in some of the cases there are two causes).

* https://ourworldindata.org/grapher/share-of-deaths-by-cause

11.4 Gazing the future

The enormous promotion of climate related risks has been accompanied with the development of a paradigm of prophecy for the future of the planet and of humanity, based on models. There is no parsimony in the time horizon of such prophesies, which can reach the year 100 000 AD (Shaffer et al., 2009) or even 1 million years (Archer et al., 2020).

The prophetic approach is also quite pessimistic, generally predicting future disasters, more recently despising science and technology, if not attempting to deprive mankind of them, like in Aeschylus's extract from Prometheus Bound, which appears as a motto in the beginning of the book.

The book supported the more traditional historical approach, which is also stochastic, both in the modern and the ancient meaning of the term (cf. the quotation by Basilius Caesariensis in Digression 1.A). We use the scientific method to reveal hidden secrets of the past and quantify the evolution of natural processes. We use stochastics to describe that evolution in the past and to make induction for the future.

History teaches us that technology has substantially contributed to risk reduction, to the quality and length of human life, and to human life as a value. It can thus further improve the present. By improving the present, using lessons from the past, we might develop an optimistic envisagement of the future—and indeed, the information presented in this epilogue allows it.

References

Archer, D., Kite, E. and Lusk, G., 2020. The ultimate cost of carbon. Climatic Change. doi: 10.1007/s10584-020-02785-4.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, ed by Petrov, B.N., Csáki, F., Tsahkadsor, Armenia, USSR, 1971, Akadémiai Kiadó, Budapest 267–281. (Republished in *Breakthroughs in Statistics, I*, ed. by Kotz, S. and Johnson, N.L., Springer-Verlag, pp. 610–624, 1992).

Akaike, H., 1974. A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19 (6): 716–723, doi: 10.1109/TAC.1974.1100705.

Allen, R. J., and DeGaetano, A.T., 2005. Considerations for the use of radar-derived precipitation estimates in determining return intervals for extreme areal precipitation amounts, *Journal of Hydrology*, 315, 203-219.

Anagnostopoulos, G.G., Koutsoyiannis, D., Christofides, A., Efstratiadis, A., and Mamassis, N., 2010. A comparison of local and aggregated climate model outputs with observed data. *Hydrological Sciences Journal*, 55 (7), 1094–1110, doi: 10.1080/02626667.2010.513518.

Arendarczyk, M., Kozubowski, T.J., and Panorska, A.K., 2018. The joint distribution of the sum and maximum of dependent Pareto risks. *Journal of Multivariate Analysis*, 167, 136-156, doi: 10.1016/j.jmva.2018.04.002.

Atkins, P. 2007. *Four Laws that Drive the Universe*, Oxford Univ. Press, Oxford, 131 pp.

Bacro, J.-N., and Chaouche, A., 2006. Incertitude d'estimation des pluies extrêmes du pourtour méditerranéen: illustration par les données de Marseille. *Hydrol. Sci. J.*, 51(3), 389-405.

Bailey, A.L., 1929. A summary of advanced statistical methods. US Statistical Bureau, https://babel.hathitrust.org/cgi/pt?id=mdp.39015067991326&view=1up&seq=7.

Barnes, F.B., 1954. Storage required for a city water supply. *J. Inst. Eng., Australia*, 26, 198-203.

Barnett, V., 2006. Chancing an interpretation: Slutsky's random cycles revisited. *Euro. J. History of Economic Thought*, 13(3), 411-432, doi: 10.1080/09672560600875596.

Beard, L.R., 1965. Use of interrelated records to simulate streamflow. *Proc. Am. Soc. Civil Eng., J. Hydraul. Div.*, 91(HY5), 13-22.

Ben-Naim, A., 2008. *A Farewell to Entropy: Statistical Thermodynamics Based on Information*, World Scientific Pub., Singapore, 384 pp.

Bernoulli, J., 1713. Ars Conjectandi, Opus Posthumum. Accedit Tractatus de Seriebus Infinitis, et Epistola Gallicé Scripta de Ludo Pilae Reticularis. Basileae, Impensis Thurnisiorum, Fratrum [English translation: *The art Of Conjecturing, Together With Letter to a Friend on Sets in Court Tennis*. Translated with an Introduction and Notes by Edith Dudley Sylla; The Johns Hopkins University Press, Baltimore, Maryland, USA, 2006.]

Birkhoff, G. D., 1931. Proof of the ergodic theorem. *Proc. Nat. Acad. Sci.*, 17, 656–660.

Blom, G., 1958. *Statistical Estimates and Transformed Beta Variables*. Wiley, New York, USA.

Boltzmann, L. 1877. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht. *Wien. Ber.*, 76, 373-435.

Bortkiewicz, L., 1917. *Die Iterationen — Ein Beitrag zur Wahrscheinlichkeitstheorie*. Springer, Berlin.

Box, G.E., and Jenkins, G.M., 1970. *Time Series Models for Forecasting and Control*. Holden Day, San Francisco, USA.

Brill, P.H., 2017. *Level Crossing Methods in Stochastic Models*. Springer.

Camuffo, D., della Valle, A., Bertolin, C., Santorelli, E., 2017a. Temperature observations in Bologna, Italy, from 1715 to 1815: a comparison with other contemporary series and an overview of three centuries of changing climate. *Climatic Change*, 142 (7), doi: 10.1007/s10584-017-1931-2.

Camuffo, D., della Valle, A., Bertolin, C., Santorelli, E., 2017b. Online Resources 1-12 for: Temperature observations in Bologna, Italy, from 1715 to 1815: a comparison with other contemporary series and an overview of three centuries of changing climate. *Climatic Change*, 142 (7).

Castellarin, A., Vogel, R.M., and Matalas, N.C., 2005. Probabilistic behavior of a regional envelope curve. *Water Resour. Res.*, 41, W06018, doi: 10.1029/2004WR003042.

Cavanaugh, N.R., Gershunov, A., Panorska, A.K., and Kozubowski, T.J., 2015. The probability distribution of intense daily precipitation. *Geophys. Res. Lett.*, 42, doi:10.1002/2015GL063238.

Chaouche, K., 2001. Approche multifractale de la modélisation stochastique en hydrologie. Thèse, Ecole Nationale du Génie Rural, des Eaux et des Forêts, Centre de Paris, France (http://www.engref.fr/thesechaouche.htm).

Chaouche, K., Hubert, P., and Lang, G., 2002. Graphical characterisation of probability distribution tails. *Stoch. Environ. Res. Risk Assess.*, 16(5), 342–357.

Chow, V.T., 1969. Stochastic analysis of hydrologic systems. Final Report, University of Illinois. Urbana, Illinois, https://www.ideals.illinois.edu/bitstream/handle/2142/90345/Chow_1969.pdf?sequence=2.

Chow, V.T., and Kareliotis, S.J., 1970. Analysis of stochastic hydrologic systems. *Water Resources Research*, 6(6), 1569-1582.

Chow, V.T., Maidment, D.R. and Mays, L.W., 1988. *Applied Hydrology*. McGraw-Hill, New York.

Clausius, R., 1865. Über verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie. *Annalen der Physik und Chemie*, 125 (7), 353–400, http://books.google.gr/books?id=qAg4AAAAMAAJ.

Clausius, R., 1867. *The Mechanical Theory of Heat: With Its Applications to the Steam-Engine and To the Physical Properties of Bodies*. J. van Voorst, London, http://books.google.gr/books?id=8LIEAAAAYAAJ.

Clausius, R., 1872. A contribution to the history of the mechanical theory of heat. *Phil. Mag.*, 43, 106–115.

Coles, S., and Pericchi, L., 2003. Anticipating catastrophes through extreme value modelling, *Appl. Statist.*, 52, 405–416.

Coles, S., Pericchi L.R., and Sisson, S., 2003. A fully probabilistic approach to extreme rainfall modeling, *J. Hydrol.*, 273 (1–4), 35–50.

Copernicus Climate Change Service, 2017. ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, Copernicus Climate Change Service Climate Data Store (CDS), https://cds.climate.copernicus.eu/cdsapp#!/home,.

Cunnane, C., 1978. Unbiased plotting positions—a review. *Journal of Hydrology*, 37(3-4), 205-222.

Dai, J.C., Hu, Y.P., Liu, D.S., and Long, X., 2011. Aerodynamic loads calculation and analysis for large scale wind turbine based on combining BEM modified theory with dynamic stall model. *Renewable Energy*, 36 (3), 1095-1104, doi: 10.1016/j.renene.2010.08.024.

Davison, A.C., and Huser, R., 2015. Statistics of Extremes. *Annual Review of Statistics and Its Application* 2 (1), 203-235.

Dechant, A., and Lutz, E., 2015. Wiener-Khinchin theorem for nonstationary scale-invariant processes. *Physical Review Letters* 115 (8), 080603.

Deligiannis, I., Dimitriadis, P., Daskalou, O., Dimakos, Y., and Koutsoyiannis, D., 2016. Global investigation of double periodicity of hourly wind speed for stochastic simulation; application in Greece, *Energy Procedia*, 97, 278–285, doi:10.1016/j.egypro.2016.10.001.

De Luca D.L., Petroselli A., and Galasso L., 2020. Modelling climate changes with stationary models: Is it possible or is it a paradox?, in: Sergeyev Y., and Kvasov D. (eds) *Numerical Computations: Theory and Algorithms, NUMTA 2019, Lecture Notes in Computer Science*, 11974, Springer, Cham, Switzerland.

Di Baldassarre, G., Montanari, A., Lins, H., Koutsoyiannis, D., Brandimarte, L., and Blöschl, G., 2010. Flood fatalities in Africa: From diagnosis to mitigation. *Geophys. Res. Lett.*, 37, L22402, doi: 10.1029/2010GL045467.

Dimitriadis, P., and Koutsoyiannis, D., 2015a. Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes. *Stochastic Environmental Research & Risk Assessment*, 29 (6), 1649–1669, doi: 10.1007/s00477-015-1023-7.

Dimitriadis, P., and Koutsoyiannis, D., 2015b. Application of stochastic methods to double cyclostationary processes for hourly wind speed simulation, *Energy Procedia*, 76, 406–411, doi: 10.1016/j.egypro.2015.07.851.

Dimitriadis, P., and Koutsoyiannis, D., 2018. Stochastic synthesis approximating any process dependence and distribution. *Stochastic Environmental Research & Risk Assessment*, doi: 10.1007/s00477-018-1540-2.

Dimitriadis, P., Koutsoyiannis, D., and Tzouka, K., 2016. Predictability in dice motion: how does it differ from hydrometeorological processes?. *Hydrological Sciences Journal*, 61 (9), 1611–1622, doi:10.1080/02626667.2015.1034128.

Dimitriadis, P., Tzouka, K., Koutsoyiannis, D., Tyralis, H., Kalamioti, A., Lerias, E., and Voudouris, P., 2019. Stochastic investigation of long-term persistence in two-dimensional images of rocks. *Spatial Statistics*, 29, 177–191, doi: 10.1016/j.spasta.2018.11.002.

Dodge, Y., 1996. A natural random number generator. *International Statistical Review*, 64(3), 329-344.

Efstratiadis, A., Mamassis, N., and Tsoukalas, I., 2019. Synoptic report on the evaluation of the flood risk for areas affected by the ongoing spilling of the Hylike-Paralimni system. Modernization of the management of the water supply system of Athens - Update, Department of Water Resources and Environmental Engineering – National Technical University of Athens, 19 pp., https://www.itia.ntua.gr/1988/.

European Commission, 2007. Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks, *Official Journal of the European Communities*, L 288, 6.11.2007, p. 27.

Ewing, J., 1920. LXII. The specific heat of saturated vapour and the entropy-temperature diagrams of certain fluids. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 39 (234), 633–646.

Fagan, B., 2008. The great warming: Climate change and the rise and fall of civilizations. Bloomsbury Publishing, USA.

Fischer, H., 2010. *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer, New York, USA.

Fisher, R.A., and Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Math. Proc. Camb. Phil. Soc.*, 24(2), 180–190, doi: 10.1017/S0305004100015681.

Fréchet, M., 1927. Sur la loi de probabilité de l'écart maximum, *Ann. Soc. Polon. Math.*, 6, 93.

Galton, F., 1890. Dice for statistical experiments. *Nature*, 42, 13–14.

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklöv, J., Forsberg, B., and Leone, M., 2015. Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The Lancet*, 386 (9991), 369-375.

Gauch, H.G., Jr., 2003. *Scientific Method in Practice*. Cambridge University Press, Cambridge, UK.

Glynis, K., 2019. Stochastic investigation of the behavior of land surface air temperature on global scale. Diploma thesis, 159 pp., Athens, www.itia.ntua.gr/2008/.

Gnedenko, B., 1943. Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.*, 44(3), 423–453, doi: 10.2307/1968974.

Gnedenko, B.V., and Kolmogorov, A.N., 1949. Predelnye raspredelniya dlya summ nezavisimykh sluchainykh velichin (Limit distributions for sums of independent random variables), Gostekhizdat, Moscow-Leningrad, 1949 (in Russian).

Griffin, D., and Anchukaitis, K.J., 2014. How unusual is the 2012–2014 California drought?. *Geophys. Res. Lett.*, 41, 9017–9023, doi: 10.1002/2014GL062433.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C., and Wallis, J.R., 1979. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resour. Res.*, 15 (5), 1049–1054, doi: 10.1029/WR015i005p01049.

Gringorten, I.I., 1963. A plotting rule for extreme probability paper. *J. Geophys. Res.*, 68(3): 813--814.

Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples, *Technometrics*, 11 (1), 1-21, doi: 10.1080/00401706.1969.10490657.

Gumbel, E.J., 1958. *Statistics of Extremes*, Columbia Univ. Press, New York, USA.

Hanel M., Rakovec, O., Markonis, Y., Máca, P. Samaniego, L., Kyselý, J., and Kumar, R., 2018. Revisiting the recent European droughts from a long-term perspective. *Scientific Reports*, 8, 9499, doi: 10.1038/s41598-018-27464-4.

Hazen, A., 1914. The storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, 77, 1539-1669.

Hemelrijk, J., 1966. Underlining random variables. *Statistica Neerlandica*, 20 (1), 1-7.

Herbertson, A.J., 1907. *Outlines of Physiography, an Introduction to the Study of the Earth*, Arnold, London, UK.

Hershfield, D.M., 1961. Estimating the probable maximum precipitation, *Proc. Am. Soc. Civil Eng., J. Hydraul. Div.*, 87(HY5), 99-106.

Hershfield, D.M. and Wilson, W.T., 1957. Generalizing of rainfall-intensity-frequency data. AIHS. Gen. Ass. Toronto 1957, 1, pp.499-506, http://hydrologie.org/redbooks/a043/043051.pdf.

Hirsch, R.M., and Ryberg, K.R., 2012. Has the magnitude of floods across the USA changed with global CO2 levels?. *Hydrological Sciences Journal*, 57 (1), 1-9, doi: 10.1080/02626667.2011.621895.

Hoffman, J.I., 2015. *Biostatistics for medical and biomedical practitioners*. Academic Press, London, UK.

Hosking, J.R.M.,1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B*, 52, 105–124.

Hosking, J.R.M., 1994. The four-parameter kappa distribution. *IBM Journal of Research and Development*, 38(3), 251–258, doi: 10.1147/rd.383.0251.

Hosking, J.R., and Wallis, J.R., 1997. *Regional Frequency Analysis—An Approach Based on L-Moments*. Cambridge Univ. Press, New York.

Hosking, J.R., Wallis, J.R. and Wood, E.F., 1985a. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3), 251-261.

Hosking, J.R., Wallis, J.R. and Wood, E.F., 1985b. An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. *Hydrological Sciences Journal*, 30(1), 85-109.

Hurst, H.E., 1951. Long term storage capacities of reservoirs. *Trans. Am. Soc. Civil Eng.*, 116, 776–808.

Iliopoulou, T., and Koutsoyiannis, D., 2019. Revealing hidden persistence in maximum rainfall records. *Hydrological Sciences Journal*, 64 (14), 1673–1689, doi:10.1080/02626667.2019.1657578.

Iliopoulou, T., and Koutsoyiannis, D., 2020. Projecting the future of rainfall extremes: better classic than trendy. *Journal of Hydrology*, 588, 125005, doi:10.1016/j.jhydrol.2020.125005.

Iliopoulou, T., Aguilar, C., Arheimer, B., Bermúdez, M., Bezak, N., Ficchi, A., Koutsoyiannis, D., Parajka, J., Polo, M.J., Thirel, G., and Montanari, A., 2019. A large sample analysis of European rivers on seasonal river flow correlation and its physical drivers, *Hydrology and Earth System Sciences*, 23, 73–91, doi: 10.5194/hess-23-73-2019.

IPCC, 2013a. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, NY, 1535 pp. http://www.climatechange2013.org/report/.

IPCC, 2013b: Annex III: Glossary [Planton, S. (ed.)]. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [ed. by Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley]. Cambridge University Press, Cambridge, UK, and New York, NY, USA.

Jaynes, E.T., 1957. Information theory and statistical mechanics. *Physical Review*, 106 (4), 620-630.

Jaynes, E.T., 2003. *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge, UK, 728 pp.

Kendall, M.G., and Stuart, A., 1963. *The Advanced Theory of Statistics*, Vol. 1, Distribution Theory, 2nd edition, Charles Griffin & Co., London, 434 pp.

Keshner, M.S., 1982. 1/f noise. *Proc. IEEE*, 70, 212–218.

Khintchine, A., 1933. Zu Birkhoffs Lösung des Ergodenproblems. *Math. Ann.*, 107, 485-488.

Khintchine, A., 1934. Korrelationstheorie der stationären stochastischen Prozesse. *Mathematische Annalen*, 109 (1), 604–615.

Klein Tank, A.M.G., Wijngaard, J.B., Können, G.P., Böhm, R., Demarée, G., Gocheva, A., et al., 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Intern. J. Climatol.*, 22, 1441-1453, doi:10.1002/joc.773.

Klemeš, V., 1986. Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, 31(1), 13-24.

Klemeš, V., 1987. One hundred years of applied storage reservoir theory. *Water Resources Management*, 1(3), 159-175.

Kolmogorov, A.N., 1931. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math. Ann.*, 104, 415-458. (English translation: On analytical methods in probability theory, In: Kolmogorov, A.N., *Selected Works of A. N. Kolmogorov - Volume 2, Probability Theory and Mathematical Statistics*, ed. by A.N. Shiryayev, Kluwer, Dordrecht, The Netherlands, 62-108, 1992).

Kolmogorov, A.N., 1933. Grundbegriffe der Wahrscheinlichkeitsrechnung, *Ergebnisse der Math.* (2), Berlin (2nd English Edition: Foundations of the Theory of Probability, 84 pp. Chelsea Publishing Company, New York, 1956).

Kolmogorov, A.N., 1938. A simplified proof of the Birkhoff-Khinchin ergodic theorem. *Uspekhi Mat. Nauk*, 5, 52-56. (English edition: Kolmogorov, A.N., 1991, *Selected Works of A. N. Kolmogorov - Volume 1, Mathematics and Mechanics*, ed. by Tikhomirov, V.M., Kluwer, Dordrecht, The Netherlands, 271-276).

Kolmogorov, A.N., 1940. Wiener spirals and some other interesting curves in a Hilbert space. *Dokl. Akad. Nauk SSSR*, 26, 115-118. (English edition: Kolmogorov, A.N., 1991, *Selected Works of A. N. Kolmogorov - Volume 1, Mathematics and Mechanics*, ed. by Tikhomirov, V.M., Kluwer, Dordrecht, The Netherlands,, 324-326).

Kolmogorov, A.N., 1947. Statistical theory of oscillations with continuous spectrum. *Collected papers on the 30th anniversary of the Great October Socialist Revolution*, Vol. 1, *Akad. Nauk SSSR*, Moscow-Leningrad, 242-252. (English edition: Kolmogorov, A.N., 1992. *Selected Works of A. N. Kolmogorov - Volume 2, Probability Theory and Mathematical Statistics*, ed. by Shiryayev, A.N., Kluwer, Dordrecht, The Netherlands, 321-330).

Köppen, W., 1918. Klassifikation der Klimate nach Temperatur, Niederschlag und Jahreslauf, *Petermanns Geogr. Mitteilungen*, 64, 103-203.

Koutsoyiannis, D., 1994. A stochastic disaggregation method for design storm and flood synthesis. *Journal of Hydrology*, 156, 193–225, doi:10.1016/0022-1694(94)90078-7.

Koutsoyiannis, D. 1997. *Statistical Hydrology*. Edition 4 (in Greek), National Technical University of Athens, Athens, 312 pp., doi: 10.13140/RG.2.1.5118.2325.

Koutsoyiannis, D., 1999. A probabilistic view of Hershfield's method for estimating probable maximum precipitation. *Water Resources Research*, 35 (4), 1313–1322.

Koutsoyiannis, D., 2000. A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. *Water Resources Research*, 36 (6), 1519–1533.
Koutsoyiannis, D., 2001. Coupling stochastic models of different time scales. *Water Resources Research*, 37 (2), 379–391, doi: 10.1029/2000WR900200.
Koutsoyiannis, D., 2002. The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences Journal*, 47 (4), 573–595, doi: 10.1080/02626660209492961.
Koutsoyiannis, D., 2003. Climate change, the Hurst phenomenon, and hydrological statistics. *Hydrological Sciences Journal*, 48 (1), 3–24, doi: 10.1623/hysj.48.1.3.43481.
Koutsoyiannis, D., 2004a. Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation. *Hydrological Sciences Journal*, 49 (4), 575–590.
Koutsoyiannis, D., 2004b. Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records. *Hydrological Sciences Journal*, 49 (4), 591–610.
Koutsoyiannis, D., 2005a. Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling. *Hydrological Sciences Journal*, 50 (3), 381–404, doi: 10.1623/hysj.50.3.381.65031.
Koutsoyiannis, D., 2005b. Uncertainty, entropy, scaling and hydrological stochastics, 2, Time dependence of hydrological processes and time scaling. *Hydrological Sciences Journal*, 50 (3), 405–426, doi: 10.1623/hysj.50.3.405.65028.
Koutsoyiannis, D., 2005c. Internal report, Supplement to "Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling", http://www.itia.ntua.gr/641/.
Koutsoyiannis, D., 2006a. An entropic-stochastic representation of rainfall intermittency: The origin of clustering and persistence. *Water Resources Research*, 42 (1), W01401, doi: 10.1029/2005WR004175.
Koutsoyiannis, D., 2006b. Nonstationarity versus scaling in hydrology, *Journal of Hydrology*, 324, 239–254.
Koutsoyiannis, D., 2007. A critical review of probability of extreme rainfall: principles and models, *Advances in Urban Flood Management*, ed. by Ashley, R., Garvin, S., Pasche, E., Vassilopoulos, A., and Zevenbergen, C., 139–166, doi: 10.1201/9780203945988.ch7, Taylor and Francis, London.
Koutsoyiannis, D., 2008. *Probability and Statistics for Geophysical Processes*, National Technical University of Athens, Athens, doi: 10.13140/RG.2.1.2300.1849/1.
Koutsoyiannis, D., 2010. A random walk on water. *Hydrology and Earth System Sciences*, 14, 585–601.
Koutsoyiannis, D., 2011a. Hurst-Kolmogorov dynamics and uncertainty, *Journal of the American Water Resources Association*, 47 (3), 481–495.
Koutsoyiannis, D., 2011b. Hurst-Kolmogorov dynamics as a result of extremal entropy production. *Physica A*, 390 (8), 1424–1432.
Koutsoyiannis, D., 2012. Clausius-Clapeyron equation and saturation vapour pressure: simple theory reconciled with practice. *European Journal of Physics*, 33 (2), 295–305, doi:10.1088/0143-0807/33/2/295.
Koutsoyiannis, D., 2013a. Physics of uncertainty, the Gibbs paradox and indistinguishable particles. *Studies in History and Philosophy of Modern Physics*, 44, 480–489.
Koutsoyiannis, D., 2013b. Hydrology and Change. *Hydrological Sciences Journal*, 58 (6), 1177–1197, doi: 10.1080/02626667.2013.804626.
Koutsoyiannis, D., 2014a. Entropy: from thermodynamics to hydrology. *Entropy*, 16 (3), 1287–1314.
Koutsoyiannis, D., 2014b. Random musings on stochastics (Lorenz Lecture), *AGU 2014 Fall Meeting*, San Francisco, USA, American Geophysical Union, doi: 10.13140/RG.2.1.2852.8804.
Koutsoyiannis, D., 2016. Generic and parsimonious stochastic modelling for hydrology and beyond. *Hydrological Sciences Journal*, 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., 2019a. Knowable moments for high-order stochastic characterization and modelling of hydrological processes. *Hydrological Sciences Journal*, 64 (1), 19–33, doi: 10.1080/02626667.2018.1556794.
Koutsoyiannis, D., 2019b. Time's arrow in stochastic characterization and simulation of atmospheric and hydrological processes. *Hydrological Sciences Journal*, 64 (9), 1013–1037, doi: 10.1080/02626667.2019.1600700.
Koutsoyiannis, D., 2020a. Simple stochastic simulation of time irreversible and reversible processes. *Hydrological Sciences Journal*, 65 (4), 536–551, doi: 10.1080/02626667.2019.1705302 (also: http://www.itia.ntua.gr/1975/).
Koutsoyiannis, D., 2020b. Revisiting the global hydrological cycle: is it intensifying?. *Hydrology and Earth System Sciences*. 24, 3899–3932, doi:10.5194/hess-24-3899-2020.

Koutsoyiannis, D., 2020c: Rebuttal to review comments on "Revisiting global hydrological cycle: Is it intensifying?" (Interactive comment on "Revisiting global hydrological cycle: Is it intensifying?" by Demetris Koutsoyiannis). *Hydrol. Earth Syst. Sci. Discuss.*, doi: 10.5194/hess-2020-120-AC1.

Koutsoyiannis, D., 2020d. A voyage in climate, hydrology and life on a 4.5-billion-years old planet, Self-organized lecture, School of Civil Engineering – National Technical University of Athens, Athens, doi: 10.13140/RG.2.2.27000.26883.

Koutsoyiannis, D., Dimitriadis, P., Lombardo, F., and Stevens, S., 2018. From fractals to stochastics: Seeking theoretical consistency in analysis of geophysical data. *Advances in Nonlinear Geosciences*; ed. by Tsonis, A., Springer, New York, NY, USA.

Koutsoyiannis, D., and Economou, A., 2003. Evaluation of the parameterization-simulation-optimization approach for the control of reservoir systems. *Water Resources Research*, 39 (6), 1170, doi: 10.1029/2003WR002148.

Koutsoyiannis, D., Efstratiadis, A., and Georgakakos, K., 2007. Uncertainty assessment of future hydroclimatic predictions: A comparison of probabilistic and scenario-based approaches. *Journal of Hydrometeorology*, 8 (3), 261–281, doi: 10.1175/JHM576.1.

Koutsoyiannis, D. Efstratiadis, A., Mamassis, N., and Christofides, A., 2008. On the credibility of climate predictions. *Hydrological Sciences Journal*, 53 (4), 671–684, doi: 10.1623/hysj.53.4.671.

Koutsoyiannis, D., and Georgakakos, A., 2006. Lessons from the long flow records of the Nile: determinism vs indeterminism and maximum entropy. *20 Years of Nonlinear Dynamics in Geosciences*, Rhodes, Greece, doi: 10.13140/RG.2.2.10996.14727.

Koutsoyiannis, D., Kozonis, D., and Manetas, A. 1998. A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206 (1-2), 118–135.

Koutsoyiannis, D., and Kundzewicz, Z.W., 2020. Atmospheric temperature and CO_2: Hen-or-egg causality?, *Sci*, 2 (3), 72, doi: 10.3390/sci2040077.

Koutsoyiannis, D., and Langousis, A., 2011. Precipitation, *Treatise on Water Science*, edited by P. Wilderer and S. Uhlenbrook, 2, 27–78, Academic Press, Oxford.

Koutsoyiannis, D., and Manetas, A., 1996. Simple disaggregation by accurate adjusting procedures. *Water Resources Research*, 32 (7), 2105–2117, doi: 10.1029/96WR00488.

Koutsoyiannis, D., Markonis, Y., Koukouvinos, A., Papalexiou, S.M., Mamassis, N., and Dimitriadis, P., 2010. Hydrological study of severe rainfall in the Kephisos basin, Greece. *Study of the management of Kephisos*, Commissioner: General Secretariat of Public Works – Ministry of Environment, Planning and Public Works, Contractors: Exarhou Nikolopoulos Bensasson, Denco, G. Karavokiris, et al., 154 pages, Athens (http://www.itia.ntua.gr/970/).

Koutsoyiannis, D., and Montanari, A., 2007. Statistical analysis of hydroclimatic time series: Uncertainty and insights. *Water Resources Research*, 43 (5), W05429, doi: 10.1029/2006WR005592.

Koutsoyiannis, D., and Montanari, A., 2015. Negligent killing of scientific concepts: the stationarity case. *Hydrological Sciences Journal*, 60 (7-8), 1174–1183, doi: 10.1080/02626667.2014.959959.

Koutsoyiannis, D., and Papalexiou, S.M., 2017. Extreme rainfall: Global perspective. *Handbook of Applied Hydrology, Second Edition*, ed. by Singh, V.P., 74.1–74.16, McGraw-Hill, New York.

Koutsoyiannis, D., Paschalis, A., and Theodoratos, N., 2011. Two-dimensional Hurst-Kolmogorov process and its application to rainfall fields. *Journal of Hydrology*, 398 (1-2), 91–100.

Koutsoyiannis, D. and Xanthopoulos, T., 1999. *Engineering Hydrology*, Edition 3 (in Greek), 418 pp., National Technical University of Athens, Athens, doi: 10.13140/RG.2.1.4856.0888.

Kron, W., Eichner, J. and Kundzewicz, Z.W., 2019. Reduction of flood risk in Europe–Reflections from a reinsurance perspective. *Journal of Hydrology*, 576, 197-209, doi: 10.1016/j.jhydrol.2019.06.050.

Krumbein, W.C., 1968. Statistical models in sedimentology. *Sedimentology*, 10 (1), 7-23.

Kundzewicz, Z.W., Pińskwar, I., and Koutsoyiannis, D., 2020. Variability of global mean annual temperature is significantly influenced by the rhythm of ocean-atmosphere oscillations, *Science of the Total Environment*, 747, 141256, doi: 10.1016/j.scitotenv.2020.141256.

Kyun, K. and Kim, K., 2006. *Equilibrium Business Cycle Theory in Historical Perspective*. Cambridge University Press, Cambridge, UK.

Lamb, H.H., 1972. *Climate: Past, Present, and Future, Vol. 1: Fundamentals and Climate Now*. Methuen, London, UK.

Lampret, V., 2015. Accurate double inequalities for generalized harmonic numbers. *Applied Mathematics and Computation*, 265, 557-567.

Landwehr, J.M., Matalas, N.C., and Wallis, J.R., 1979. Probability-weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resour. Res.*, 15(5), 1055-1064.

Lasota, A., and Mackey, M.C., 1994. *Chaos, Fractals and Noise*, Springer-Verlag, New York, USA.

Leadbetter, M.R., 1983. Extremes and local dependence in stationary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65, 291–306.

Linsley, R.K., Kohler, M.A. and Paulhus, J.L.H., 1975. *Hydrology for Engineers*. 2nd Edition, McGrawHill, New York.

Livadiotis, G., and McComas, D.J., 2013. Understanding kappa distributions: A toolbox for space science and astrophysics. *Space Science Reviews*, 175 (1-4), 183-214.

Lombardo, F., Napolitano, F., and Russo, F., 2006. On the use of radar reflectivity for estimation of the areal reduction factor, *Natural Hazards and Earth System Sciences*, 6, 377-386.

Lombardo, F., Napolitano, F., Russo, F., and Koutsoyiannis, D., 2019. On the exact distribution of correlated extremes in hydrology. *Water Resources Research*, 55 (12), 10405–10423, doi: 10.1029/2019WR025547.

Lombardo, F., Volpi, E., Koutsoyiannis, D., and Papalexiou, S.M., 2014. Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology. *Hydrology and Earth System Sciences*, 18, 243–255, doi: 10.5194/hess-18-243-2014.

Mackey, M.C., 2003. *Time's Arrow: The Origins of Thermodynamic Behavior*, Dover, Mineola, NY, USA, 175 pp.

Makropoulos, C., Efstratiadis, A., Karakatsanis, G., Nikolopoulos, D., and Koukouvinos, A., 2018. Estimation of financial cost of raw water - Final report, Update of financial cost of raw water for Athens water supply (in Greek), Department of Water Resources and Environmental Engineering, National Technical University of Athens, Athens, Greece.

Mandelbrot, B.B., and Van Ness, J.W., 1968. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4), 422-437.

Mandelbrot, B.B. and Wallis, J.R., 1968. Noah, Joseph, and operational hydrology. *Water Resources Research*, 4(5), 909-918.

Markonis, Y., and Koutsoyiannis, D., 2013. Climatic variability over time scales spanning nine orders of magnitude: Connecting Milankovitch cycles with Hurst–Kolmogorov dynamics. *Surveys in Geophysics*, 34 (2), 181–207, doi: 10.1007/s10712-012-9208-9.

Markonis, Y., and Koutsoyiannis, D., 2016. Scale-dependence of persistence in precipitation records, *Nature Climate Change*, 6, 399–401, doi: 10.1038/nclimate2894.

Matalas, N.C., 1967. Mathematical assessment of synthetic hydrology. *Water Resour. Res.,* 3(4), 937-945.

Matalas, N.C., and Langbein, W.B. 1962. Information content of the mean. *J. Geophys. Res.*, 67(9), 3441–3448.

Mathieu, M., 1988. On the origin of the notion 'Ergodic Theory'. *Expositiones Mathematicae*, 6, 373-377.

Mazliak, L., 2018. The beginnings of the Soviet Encyclopedia. The utopia and misery of mathematics in the political turmoil of the 1920s. *Centaurus*, 60(1-2), 25-51.

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., and Houston, T.G., 2012. An overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Oceanic Technol.*, 29, 897-910, doi: 10.1175/JTECH-D-11-00103.1.

Metropolis, N., Reitwiesner, G., and von Neumann, J., 1950. Statistical treatment of values of first 2,000 decimal digits of e and of π calculated on the ENIAC. *Mathematical Tables and Other Aids to Computation*, 4, 109-112.

Metropolis, N., and Ulam, S., 1949. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247), 335-341.

Mielke, P.W. Jr., 1973. Another family of distributions for describing and analyzing precipitation data. *Journal* of *Applied Meteorology,* 12 (2), 275-280.

Mielke P.W. Jr., and Johnson, E.S., 1973. Three-parameter kappa distribution maximum likelihood estimates and likelihood ratio tests. *Monthly Weather Review, 101(9)*, 701-707.

Milly, P.C.D., Betancourt, J. Falkenmark, M. Hirsch, R.M. Kundzewicz, Z.W. Lettenmaier, D.P. and Stouffer, R.J., 2008. Stationarity Is Dead: Whither Water Management?. *Science*, 319, 573-574.

Mimiyianni, I., 2010. Stochastic investigation of the utility of multiple hydrological records for improving the reliability of estimation, Postgraduate Thesis (in Greek), Department of Water Resources and Environmental Engineering – National Technical University of Athens, http://www.itia.ntua.gr/1058/.

Montanari, A., 2012. Hydrology of the Po River: looking for changing patterns in river discharge. *Hydrol. Earth Syst. Sci.*, 16, 3739–3747, doi: 10.5194/hess-16-3739-2012.

Montanari, A., and Koutsoyiannis, D., 2012. A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, 48, W09555, doi: 10.1029/2011WR011412.

Montanari, A., and Koutsoyiannis, D., 2014. Modeling and mitigating natural hazards: Stationarity is immortal!. *Water Resources Research*, 50 (12), 9748–9756, doi: 10.1002/2014WR016092.

Mora, R.D., Bouvier, C., Neppel L., and Niel, H., 2005. Approche régionale pour l'estimation des distributions ponctuelles des pluies journalières dans le Languedoc-Roussillon (France). *Hydrol. Sci. J.,* 50 (1), 17-29.

Moxon, J., 1700. *Mathematicks Made Easie: Or, a Mathematical Dictionary: Explaining the Terms of Art, and Difficult Phrases Used in Arithmetick, Geometry, Astronomy, Astrology, and Other Mathematical Sciences* … J. Moxon, at the sign of Atlas in Warwick Lane.

Nadarajah, S. and Kotz, S., 2008. Exact distribution of the max/min of two Gaussian random variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16 (2), 210-212, doi: 10.1109/TVLSI.2007.912191.

NERC (Natural Environment Research Council), 1975. *Flood Studies Report*. Institute of Hydrology, Wallingford, UK.

Niederreiter, H., 1992. *Random Number Generation and Quasi-Monte Carlo Methods*, Society for Industrial and Applied Mathematics, Philadelphia, USA.

Neumann, J. and Flohn, H., 1987. Great historical events that were significantly affected by the weather: Part 8, Germany's war on the Soviet Union, 1941–45. I. Long-range weather forecasts for 1941–42 and climatological studies. *Bulletin of the American Meteorological Society*, 68(6), 620-630.

O'Connell, P.E., Koutsoyiannis, D., Lins, H.F., Markonis, Y., Montanari, A., and Cohn, T.A., 2016. The scientific legacy of Harold Edwin Hurst (1880 – 1978). *Hydrological Sciences Journal*, 61 (9), 1571–1590, doi: 10.1080/02626667.2015.1125998.

Olbert, S., 1968. Summary of Experimental Results from M.I.T. Detector on IMP-1. *Physics of the Magnetosphere, Astrophysics and Space Science Library*, ed. by Carovillano R.L., McClay J.F., Radoski H.R., vol 10. Springer, Dordrecht, The Netherlands.

Overeem, A., Buishand, T.A., Holleman, I., and Uijlenhoet, R., 2010. Extreme value modeling of areal rainfall from weather radar. *Water Resour. Res.*, 46, W09514, doi:10.1029/2009WR008517.

Papacharalampous, G., Koutsoyiannis, D., and Montanari, A., 2020a. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: Methodology development and investigation using toy models, *Advances in Water Resources*, 136, 103471, doi: 10.1016/j.advwatres.2019.103471.

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., and Montanari, A., 2020b. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. *Advances in Water Resources*, 136, 103470, doi: 10.1016/j.advwatres.2019.103470.

Papalexiou, S.M., and Koutsoyiannis, D. 2012. Entropy based derivation of probability distributions: A case study to daily rainfall. *Advances in Water Resources*, 45, 51–57, doi: 10.1016/j.advwatres.2011.11.007.

Papalexiou, S.M., and Koutsoyiannis, D. 2013, Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, 49 (1), 187–201, doi: 10.1029/2012WR012557.

Papoulis, A., 1990. *Probability and Statistics*, Prentice-Hall, New Jersey, USA.

Papoulis, A., 1991. *Probability, Random Variables and Stochastic Processes*, 3rd edn., McGraw-Hill, New York, USA.

Pappas, C., Papalexiou, S.M., and Koutsoyiannis, D., 2014. A quick gap-filling of missing hydrometeorological data. *Journal of Geophysical Research-Atmospheres*, 119 (15), 9290–9300, doi: 10.1002/2014JD021633.

Perry, J., 1903. The Thermodynamics of Heat Engines. *Nature*, 67, 602–605, doi: 10.1038/067602a0.

Persons, W.M. 1919. *Measuring and Forecasting General Business Conditions*. Harvard University & American Institute of Finance, https://babel.hathitrust.org/cgi/pt?id=hvd.32044018749697.

Reitwiesner, G., 1950. An ENIAC determination of π and e to more than 2000 decimal places. *Mathematical Tables and Other Aids to Computation*, 4, 11-15.

Ripley, B.D., 1987. *Stochastic Simulation*, Wiley, New York, USA.

Robertson, H.S., 1993. *Statistical Thermophysics*, Prentice Hall, Englewood Cliffs, NJ, USA, 582 pp.

Salas, J.D., 1993. Analysis and modeling of hydrologic time series. *Handbook of Hydrology*, ed. by Maidment, D., Ch. 19, 19.1-19.72, McGraw-Hill, New York, USA.

Sargentis, G.-F., Iliopoulou, T., Sigourou, S., Dimitriadis, P., and Koutsoyiannis, D., 2020. Evolution of clustering quantified by a stochastic method — Case studies on natural and human social structures. *Sustainability*, 12 (19), 7972, doi: 10.3390/su12197972.

Schwarz, G.E., 1978. Estimating the dimension of a model. *Annals of Statistics*, 6 (2), 461–464, doi: 10.1214/aos/1176344136.

Serinaldi, F., 2015. Dismissing return periods!. *Stochastic Environmental Research and Risk Assessment*, 29 (4), 1179-1189.

Serinaldi, F., and Kilsby, C.G., 2015. Stationarity is undead: Uncertainty dominates the distribution of extremes. *Advances in Water Resources*, 77, 17-36.

Serinaldi, R., Kilsby, C.G., and Lombardo, F., 2018. Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology. *Advances in Water Resources*, 111,132-155.

Schaefer, M.G., Barker, B.L., Taylor G.H., and Wallis, J.R., 2006. Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Wahington State. *Geophysical Research Abstracts*, Vol. 8, 10899.

Shaffer, G., Olsen, S.M., and Pedersen, J.O.P., 2009. Long-term ocean oxygen depletion in response to carbon dioxide emissions from fossil fuels. *Nature Geoscience*, doi: 10.1038/NGEO420.

Shannon, C.E., 1948. The mathematical theory of communication. *Bell System Technical Journal*, 27 (3), 379-423.

Shcheglov, D., 2007. Hipparchus' table of climata and Ptolemy's geography. *Orbis Terrarum*, 9.

Sikorska, A., Montanari, A., and Koutsoyiannis, D., 2015. Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. *Journal of Hydrologic Engineering (ASCE)*, 20 (1), doi: 10.1061/(ASCE)HE.1943-5584.0000926.

Singh, V.P., and Rajagopal, A.K., 1986. A new method of parameter estimation for hydrologic frequency analysis. *Hydrological Science and Technology*, 2 (3) 33-44.

Sisson, S.A., Pericchi, L.R., and Coles, S.G., 2006. A case for a reassessment of the risks of extreme hydrological hazards in the Caribbean. *Stoch. Environ. Res. Risk Assess.*, 20, 296–306.

Slutzky, E.,1925, Über stochastische Asymptoten und Grenzwerte, *Metron*, 5(3), 3-89.

Slutsky, E., 1927. Slozhenie sluchainykh prichin, kak istochnik tsiklicheskikh protsessov. *Voprosy kon''yunktury*, 3, 34 – 64. 1927 (English edition: Slutzky, E., 1937. The summation of random causes as the source of cyclic processes. *Econometrica: Journal of the Econometric Society*, 105-146).

Slutsky, E., 1928a. Sur ùn critérium de la convergence stochastique des ensembles des valeurs éventuelles). *Comptes rendus Acad. Sci.*, 187, 370.

Slutsky, E., 1928b. Sur les fonctions éventuelles continues, intégrables et dérivables dans le sens stochastiques. *Comptes rendus des séances de l'Académie des Sciences,* 187, 878.

Slutsky, E.,1929. Quelques propositions sur les limites stochastiques éventuelles. *Comptes rendus des séances de l'Académie des Sciences*, 189, 384.

Smith, P. J., 1995. A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *The American Statistician*, 49 (2), 217–218.

Stamp, L.D., 1957. Major natural regions: Herbertson after fifty years. *Geography*, 42(4), 201-216.

Stedinger, J.R., 1983. Estimating a regional flood frequency distribution, *Water Resour. Res.*, 19, 503– 510.

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events, *Handbook of Hydrology*, ed. by Maidment, D.R., Ch. 19, McGraw-Hill, New York, USA.

Stigler, S.M., 2002. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, Mass., USA.

Sudler, C., 1927. Storage required for the regulation of streamflow. *Trans. Am. Soc. Civ. Eng.*, 91, 622-660.

Sun, F., Roderick, M.L., and Farquhar, G.D., 2012. Changes in the variability of global land precipitation. *Geophys. Res. Lett.*, 39, L19402, doi: 10.1029/2012GL053369.

Svensson, C., and Jones, D.A., 2010. Review of methods for deriving areal reduction factors. *Journal of Flood Risk Management*, 3 (3), 232-245.

Swendsen, R.H., 2006. Statistical mechanics of colloids and Boltzmann's definition of the entropy. *American Journal of Physics*, 74 (3), 187-190.

Swendsen, R.H., 2011. How physicists disagree on the meaning of entropy. *Am. J. Phys.*, 79 (4), 342-348.

Swinburne, J., 1904. Entropy. *Nature*, 70, 54–55, doi: 10.1038/070054b0.

Thomas, H.A., and Fiering, M.B. 1962. Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation. *Design of Water Resource Systems*, ed. by Maass, A., Hufschmidt, M.M., Dorfman, R., Thomas, H.A., Jr., Marglin, S.A., and Fair, G.M., Harvard University Press, Cambridge, Mass.

Tippett, L.H.C., 1927. *Random Sampling Numbers*. Cambridge University Press, Cambridge, UK.

Todorovic, P., and Zelenhasic, E., 1970. A stochastic model for flood analysis. *Water Resources Research*, 6(6), 1641–1648.

Toussoun, O., 1925. *Mémoire sur l'histoire du Nil. Mémoire de l'Institut d'Egypte*. Institut français d'archeologie orientale, Cairo, Egypt.

Tsaknias, D., Bouziotas, D., and Koutsoyiannis, D., 2016. Statistical comparison of observed temperature and rainfall extremes with climate model outputs in the Mediterranean region, *ResearchGate*, doi: 10.13140/RG.2.2.11993.93281.

Tsekouras, G., and Koutsoyiannis, D., 2014. Stochastic analysis and simulation of hydrometeorological processes associated with wind and solar energy, *Renewable Energy*, 63, 624–633, doi: 10.1016/j.renene.2013.10.018.

Tsoukalas, I., Makropoulos, C., and Koutsoyiannis, D., 2018. Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions. *Water Resources Research*, 54 (11), 9484–9513, doi: 10.1029/2017WR022462.

Tukey, J.W., 1962. The future of data analysis. *Ann. Math. Stat.*, 33 (1), 21-24.

Tyralis, H., and, Koutsoyiannis, D., 2011. Simultaneous estimation of the parameters of the Hurst-Kolmogorov stochastic process. *Stochastic Environmental Research & Risk Assessment*, 25 (1), 21–33.

Tyralis, H., and, Koutsoyiannis, D., 2014. A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. *Climate Dynamics*, 42 (11-12), 2867–2883, doi: 10.1007/s00382-013-1804-y.

Tyralis, H., and, Koutsoyiannis, D., 2017. On the prediction of persistent processes using the output of deterministic models. *Hydrological Sciences Journal*, 62 (13), 2083–2102, doi: 10.1080/02626667.2017.1361535.

Tyralis, H., Koutsoyiannis, D., and Kozanis, S., 2013. An algorithm to construct Monte Carlo confidence intervals for an arbitrary function of probability distribution parameters. *Computational Statistics*, 28 (4), 1501–1527, doi: 10.1007/s00180-012-0364-7.

Uffink, J., 1995. Can the maximum entropy principle be explained as a consistency requirement?. *Studies In History and Philosophy of Modern Physics*, 26 (3), 223-261.

US Department of the Interior (Bureau of Reclamation) 1987. *Design of Small Dams*, 3rd edn. US Government Printing Office, Denver, Colorado, USA.

US Weather Bureau, Generalized estimates of probable maximum precipitation west of the 105th meridian, Technical Paper no 38, U.S. Department of Commerce, Washington, DC, 1960, https://www.nws.noaa.gov/oh/hdsc/Technical_papers/TP38.pdf.

Veneziano, D., Langousis, A. , and Lepore, C., 2009. New asymptotic and pre-asymptotic results on rainfall maxima from multifractal theory. *Wat. Resour. Res.*, 45, W11421, doi: 10.1029/2009WR008257.

Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., and Koutsoyiannis, D., 2015. One hundred years of return period: Strengths and limitations. *Water Resources Research*, doi: 10.1002/2015WR017820.

Volpi, E., Fiori, A., Grimaldi, S., Lombardo, F., and Koutsoyiannis, D., 2019. Save hydrological observations! Return period estimation without data decimation. *Journal of Hydrology*, doi: 10.1016/j.jhydrol.2019.02.017.

von Mises, R., 1936. La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, 1, 141-160.

Wallis, J. R., and O'Connell, P.E., 1972. Small sample estimation of ρ_1. *Water Resour. Res.*, 8(3), 707-712.

Walker, G., 1931. On periodicity in series of related terms. *Proceedings of the Royal Society of London, Ser. A*, 131, 518–532.

Weibull, W., 1939. A statistical theory of strength of materials. *Ing. Vetensk. Akad. Handl.*, 151, 1–45.

Whittle, P., 1951. *Hypothesis Testing in Times Series Analysis*. PhD thesis, Almqvist & Wiksells, Uppsala, Sweden.

Whittle, P., 1952. Tests of fit in time series. *Biometrika*, 39(3/4), 309-318.

Whittle, P., 1953. The analysis of multiple stationary time series. *Journal of the Royal Statistical Society B*, 15 (1), 125-139.

WMO (World Meteorological Organization), 1992. *International Meteorological Vocabulary*. WMO, No. 182, Geneva, Switzerland, https://library.wmo.int/doc_num.php?explnum_id=4712.

Wold, H.O., 1938. *A Study in the Analysis of Stationary Time-Series*. PhD thesis, Almquist and Wicksell, Uppsala, Sweden.

Wold, H.O., 1948. On prediction in stationary time series. *The Annals of Mathematical Statistics*, 19(4), 558-567.

Wornell, G.W., 1993. Wavelet-based representations for the 1/f family of fractal processes. *Proc. IEEE*, 81, 1428–1450.

Yule, G.U. 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London A*, 226, 267–298.

Yule, G. U., 1945. A method of studying time series based on their internal correlations. *J. R. Stat. Soc.*, 108, 208–225.

Afterword comments

[to be placed in the back cover in Edition 1]

- by Harry Lins (World Meteorological Organization, Commission for Hydrology)

- by Alberto Montanari (University of Bologna and European Geosciences Union)