Forecasting of geophysical processes using stochastic and machine learning algorithms

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

5-9 July 2017
Athens, Greece

EWRA 2017

10th World Congress on Water Resources and Environment "Panta Rhei"
Conference II: Climate and Water Resources

Available online at: **itia.ntua.gr/1717**

1. Abstract

We perform an extensive comparison between four stochastic and two machine learning (ML) forecasting algorithms by conducting a multiple-case study. The latter is composed by 50 single-case studies, which use time series of total monthly precipitation and mean monthly temperature observed in Greece. We apply a fixed methodology to each individual case and, subsequently, we perform a cross-case synthesis to facilitate the detection of systematic patterns. The stochastic algorithms include the Autoregressive order one model, an algorithm from the family of Autoregressive Fractionally Integrated Moving Average models, an Exponential Smoothing State Space algorithm and the Theta algorithm, while the ML algorithms are Neural Networks and Support Vector Machines. We also use the last observation as a Naive benchmark in the comparisons. We apply the forecasting methods to the deseasonalized time series. We compare the one-step ahead as also the multi-step ahead forecasting properties of the algorithms. Regarding the one-step ahead forecasting properties, the assessment is based on the absolute error of the forecast of the last observation. For the comparison of the multi-step ahead forecasting properties we use five metrics applied to the test set (last twelve observations), i.e. the root mean square error, the Nash-Sutcliffe efficiency, the ratio of standard deviations, the index of agreement and the coefficient of correlation. Concerning the ML algorithms, we also perform a sensitivity analysis for time lag selection. Additionally, we compare more sophisticated ML methods as regards to the hyperparameter optimization to simple ones.

2. Introduction

- Machine learning (ML) algorithms are widely used for the forecasting of geophysical processes as an alternative to stochastic algorithms.
- Popular ML algorithms are the Neural Networks (NN) and the Support Vector Machines (SVM). The large number of the relevant applications is imprinted in Maier and Dandy (2000) and Raghavendra and Deka (2014).
- The research in geophysical sciences often focuses on comparing stochastic to ML forecasting algorithms.
- The comparisons performed are usually based on single-case studies (e.g. Koutsoyiannis et al. 2008; Valipour et al. 2013).
- Single-case studies offer the benefit of studying the phenomena in detail as also in their context. On the other hand, they do not allow generalizations in any extent (Achen and Snidal 1989).
- Generalizations could be derived by examining a sufficient number of different cases, as implemented in Papacharalampous (2016) and Papacharalampous et al. (2017).
- Here we conduct a multiple-case study composed by 50 individual cases, each of them based on geophysical time series data from Greece.
- In more detail:
  - ✓ We apply a fixed methodology to each individual case for the comparison between several stochastic and ML methods regarding their one-step ahead and multi-step ahead forecasting properties.
  - ✓ Concerning the ML methods, we also perform a sensitivity analysis for time lag selection. Additionally, we compare more sophisticated ML methods as regards to the hyperparameter optimization to simple ones.
  - ✓ Finally, we perform a cross-case synthesis to facilitate the detection of systematic patterns.
- The multiple-case study method can provide a form of generalization named "contingent empirical generalization", while retaining the immediacy of the single-case study method (Achen and Snidal 1989).

3. Methodology outline

- We use 50 time series of total monthly precipitation (data source: Peterson and Vose 1997) and mean monthly temperature (data source: Lawrimore et al. 2011) observed in Greece (see 4).
- We select only those with few missing values (blocks with length equal or less than one). Subsequently, we use the Kalman filter algorithm from the zoo R package (Zeileis and Grothendieck 2005) for filling in the missing values.
- We use the deseasonalized time series for the application of the forecasting methods (see 5), as suggested in Taieb et al. (2012).
- To describe the long-term persistence of the deseasonalized time series, we estimate the Hurst parameter H for each of them using the maximum likelihood method (Tyralis and Koutsoyiannis 2011) implemented with the HKprocess R package (Tyralis 2016).
- We apply the following methodology to each time series:
  - ✓ First, we split the time series into a fitting and a test set. The latter is the last observation for the one-step ahead forecasting experiments and the last 12 observations for the multi-step ahead forecasting experiments.
  - ✓ Second, we fit the models to the deseasonalized fitting set and make predictions corresponding to the test set.
  - ✓ Third, we add the seasonality to the predicted values and compare them to their corresponding observed using several metrics (see bellow).
  - ✓ Regarding the one-step ahead forecasting properties, the assessment is based on the absolute error (AE) of the forecast of the last observation.
  - ✓ For the comparison of the multi-step ahead forecasting properties we use the Root Mean Square Error (RMSE), the Nash-Sutcliffe efficiency (NSE), the ratio of standard deviations (rSD), the index of agreement (d) and the coefficient of correlation (Pr) applied to the test set. The definitions of the metrics NSE, d and Pr are available in Krause et al. (2005), while the definition of the rSD in Zambrano-Bigiarini (2014).
  - ✓ Finally, we conduct the cross-case synthesis to demonstrate similarities and differences between the single-case studies conducted.

4. Time series

	s/n	Code	Location	Length (months)	H estimate*
Precipitation time series	1	prec_1	Agrinion	384	0.48
	2	prec_2	Alexandroupoli	480	0.59
	3	prec_3	Aliartos	1008	0.53
	4	prec_4	Anogeia	252	0.52
	5	prec_5	Anogeia	360	0.53
	6	prec_6	Araxos	624	0.51
	7	prec_7	Athens	264	0.48
	8	prec_8	Athens	1428	0.53
	9	prec_9	Athens	204	0.52
	10	prec_10	Fragma	780	0.54
	11	prec_11	Heraklion	540	0.50
	12	prec_12	Igoumenitsa	480	0.49
	13	prec_13	Ioannina	480	0.58
	14	prec_14	Kalamata	180	0.51
	15	prec_15	Kalo Chorio	420	0.50
	16	prec_16	Kastelli	336	0.55
	17	prec_17	Kerkyra	540	0.51
	18	prec_18	Kythira	276	0.48
	19	prec_19	Kos	396	0.49
	20	prec_20	Kozani	396	0.57
	21	prec_21	Larissa	564	0.55
	22	prec_22	Lemnos	600	0.52
	23	prec_23	Methoni	492	0.49
	24	prec_24	Milos	480	0.57
	25	prec_25	Mytilene	468	0.55

	s/n	Code	Location	Length (months)	H estimate*
Precipitation time series	26	prec_26	Naxos	204	0.46
	27	prec_27	Patra	1008	0.52
	28	prec_28	Sitia	288	0.56
	29	prec_29	Skyros	396	0.50
	30	prec_30	Thessaloniki	804	0.58
	31	prec_31	Thessaloniki	120	0.56
	32	prec_32	Trikala	480	0.56
	33	prec_33	Tripoli	420	0.53
Temperature time series	34	temp_1	Araxos	360	0.66
	35	temp_2	Athens	1416	0.67
	36	temp_3	Athens	156	0.68
	37	temp_4	Athens	744	0.65
	38	temp_5	Heraklion	792	0.69
	39	temp_6	Kalamata	720	0.74
	40	temp_7	Kerkyra	792	0.67
	41	temp_8	Larissa	1416	0.64
	42	temp_9	Lemnos	576	0.75
	43	temp_10	Methoni	264	0.59
	44	temp_11	Methoni	312	0.61
	45	temp_12	Patra	468	0.69
	46	temp_13	Samos	180	0.64
	47	temp_14	Samos	360	0.64
	48	temp_15	Souda	660	0.71
	49	temp_16	Thessaloniki	1500	0.71
	50	temp_17	Thessaloniki	120	0.67

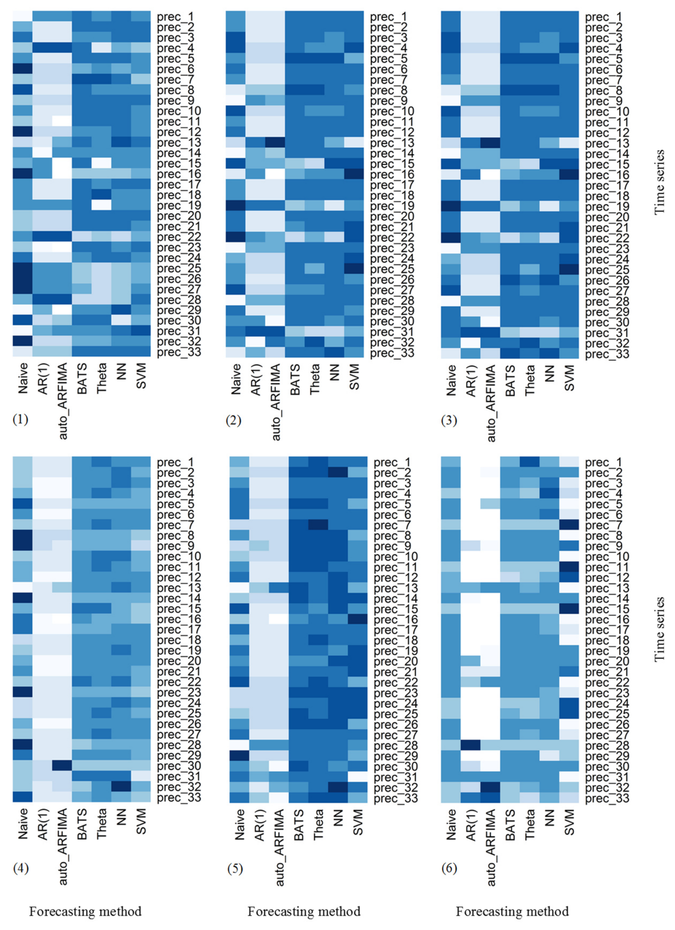
* The Hurst parameter H is estimated for the deseasonalized time series.

5. Forecasting methods

Benchmark	Stochastic	ML
Naive *(last observation)*	AR(1)	NN
	auto_ARFIMA	SVM
	BATS	
	Theta	

- We apply the benchmark and stochastic algorithms using the forecast R package (Hyndman 2016; Hyndman and Khandakar 2008) and the ML using the rminer R package (Cortez 2010, 2015).
- The Naive, AR(1), auto_ARFIMA and BATS algorithms apply Box-Cox transformation to the input data before fitting a model to them.
- While the stochastic forecasting methods are simply defined by the stochastic algorithm, the ML methods are defined by the set {ML algorithm, hyperparameter selection procedure, time lags}.
- We compare two procedures for hyperparameter selection, i.e. predefined hyperparameters or defined after optimization, and 21 regression matrices, each using the first n time lags, n = 1, 2, …, 21. The hyperparameter optimization is performed with the hold-out method.
- Hereafter, we consider that the ML models are used with predefined hyperparameters and that the regression matrix is built only by the first time lag, unless mentioned differently.
- We use two ML forecasting methods (one for each algorithm) in the comparisons conducted between stochastic and machine learning.
- We also use 42 forecasting methods (21 for each algorithm) to perform a sensitivity analysis for time lag selection and four ML forecasting methods (two for each algorithm) for the investigation of the effect of the hyperparameter optimization.

6. Comparison on precipitation time series



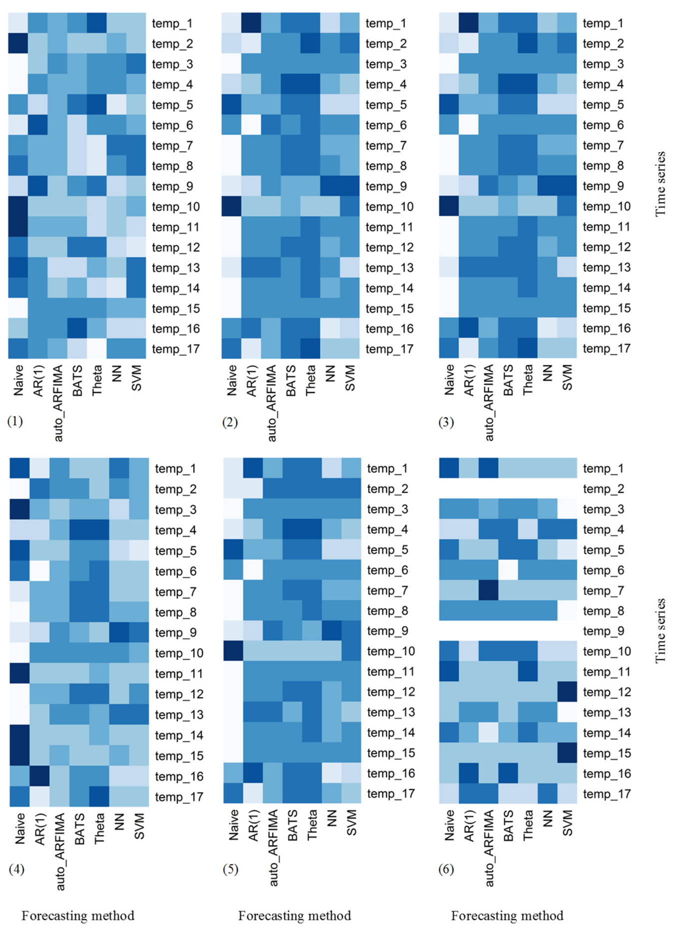
Heatmaps for the comparison between the stochastic and ML methods on the precipitation time series. The darker the colour the better the forecasts.

Notes

- The results of the single-case studies vary significantly.
- There is no best or worst forecasting method regarding all the criteria set simultaneously.
- The former observations apply equally to the stochastic and the ML forecasting methods.
- The Naive benchmark is as competent as the forecasting methods regarding all the criteria set.
- The forecasting methods AR(1) and auto_ARFIMA are the least proper to use on our precipitation data.

(1) one-step ahead forecasting - AE
(2) multi-step ahead forecasting - RMSE
(3) multi-step ahead forecasting - NSE
(4) multi-step ahead forecasting - rSD
(5) multi-step ahead forecasting - d
(6) multi-step ahead forecasting - Pr

7. Comparison on temperature time series



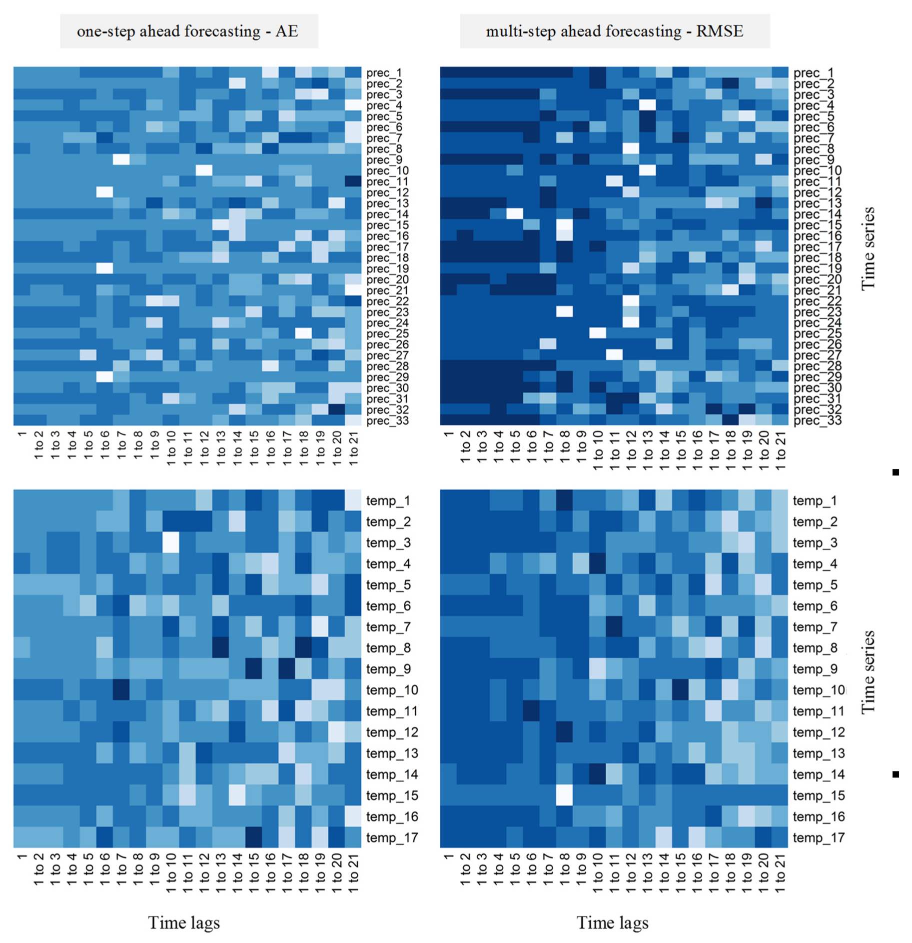
Heatmaps for the comparison between the stochastic and ML methods on the temperature time series. The darker the colour the better the forecasts.

Notes

- The first four notes of 6 also apply to 7.
- The forecasting methods AR(1) and auto_ARFIMA are competent on the temperature time series data.
- By studying the numerical results we note that the forecasts for temperature are remarkably better than the forecasts for precipitation.

(1) one-step ahead forecasting - AE
(2) multi-step ahead forecasting - RMSE
(3) multi-step ahead forecasting - NSE
(4) multi-step ahead forecasting - rSD
(5) multi-step ahead forecasting - d
(6) multi-step ahead forecasting - Pr

8. Time lag selection: NN algorithm

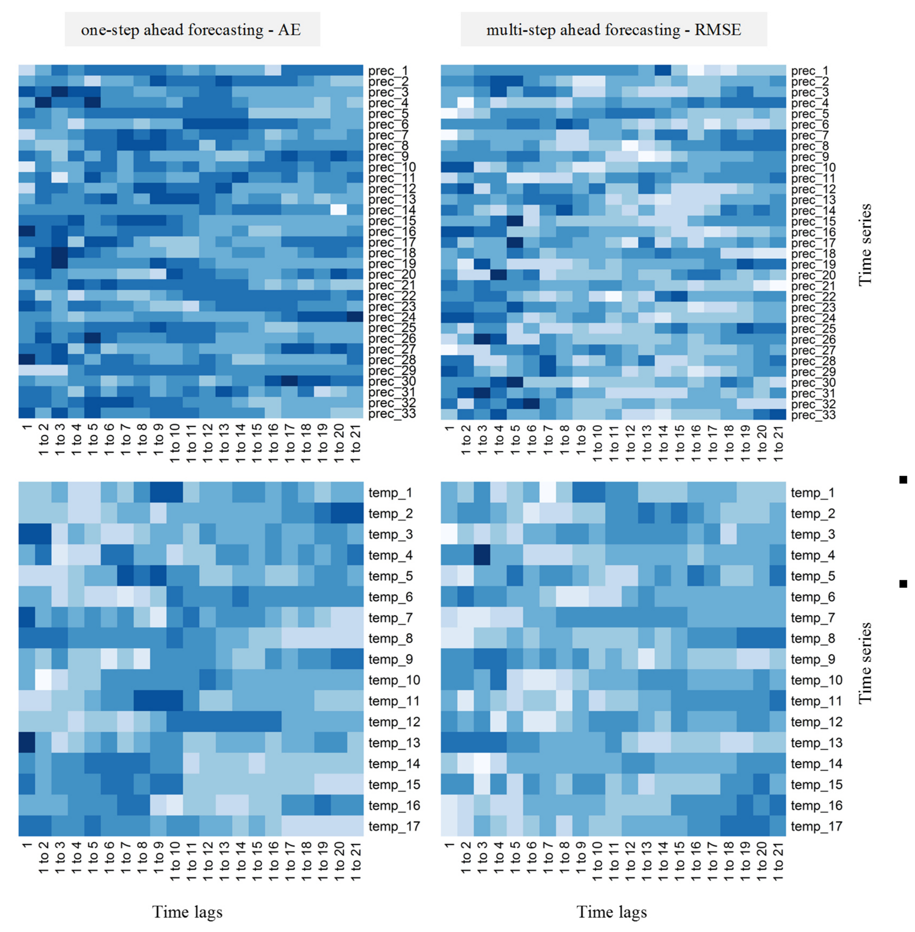


Heatmaps for the sensitivity analysis on the time lags in time series forecasting using the NN algorithm. The darker the colour the better the forecasts.

Notes

- We observe significant variations in the results across the individual cases, to an extent that it is impossible to decide on a best or worst ML forecasting method among the single-case studies.
- The left parts of the heatmaps are smoother with no white cells.

9. Time lag selection: SVM algorithm

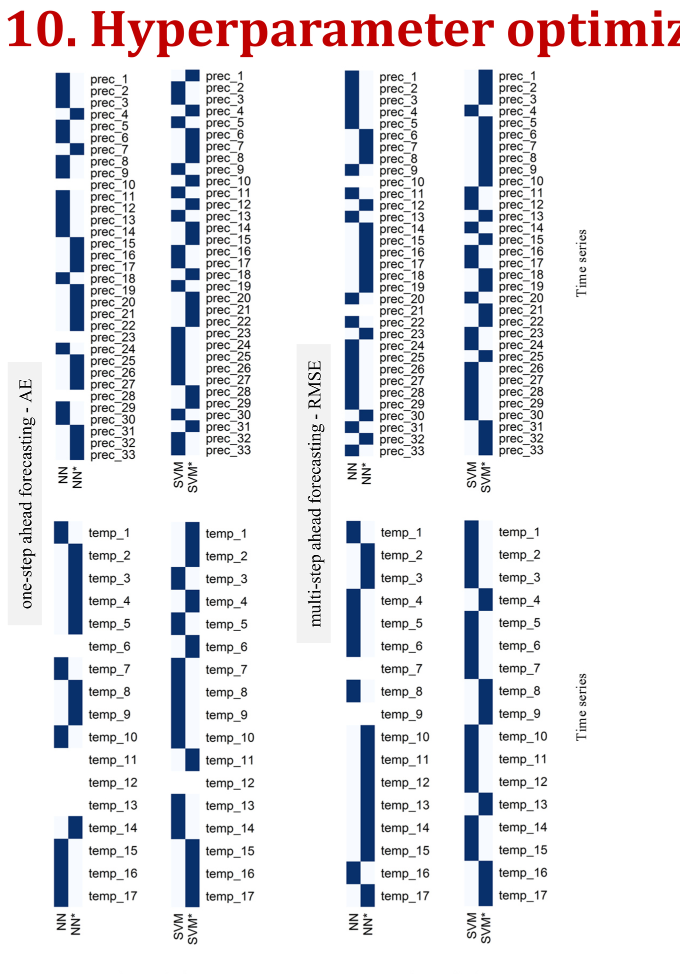


Heatmaps for the sensitivity analysis on the time lags in time series forecasting using the SVM algorithm. The darker the colour the better the forecasts.

Notes

- The first note of 8 also applies to 9.
- We observe no systematic patterns and the variations are rather random.

10. Hyperparameter optimization



*Heatmaps for the investigation of the effect of hyperparameter optimization on the forecast quality. The darker the colour the better the forecasts. The symbol * in the name of a forecasting method denotes that the model's hyperparameters have been optimized.*

Notes

- The results vary across the single-case studies in a rather random manner.
- The hyperparameter optimization does not necessary lead to better forecasts for the NN and SVM algorithms.

11. Summary and conclusions

- We compare four stochastic and two ML forecasting algorithms by conducting a multiple-case study, which is composed by 50 single-case studies.
- The latter use time series of total monthly precipitation and mean monthly temperature observed in Greece.
- We compare the one- and multi-step ahead forecasting properties of the algorithms.
- Regarding the ML algorithms, we also perform a sensitivity analysis for time lag selection.
- Furthermore, we compare more sophisticated ML methods as regards to the hyperparameter optimization to simple ones.
- The present study must be encountered as a contingent empirical evidence on several issues that have drawn the attention in the field of time series forecasting.
- The findings suggest that the stochastic and ML methods can perform equally well, but always under limitations.
- The best forecasting method depends on the case examined and the criterion of interest, while it can be either stochastic or ML. However, the ML methods are computationally intensive.
- Regarding the time lag selection, the best choice seems to depend mainly on the case, while the ML algorithm might has also some effect.
- Finally, for the algorithms used in the present study hyperparameter optimization does not necessarily lead to better forecasts.

References

Achen, C.H., and Snidal, D., 1989. Rational deterrence theory and comparative case studies. *World Politics*, 41 (2), 143-169. doi:10.2307/2010405

Cortez, P., 2010. Data Mining with Neural Networks and Support Vector Machines Using the R/rminer Tool. In: P. Perner, eds. *Advances in Data Mining*. Applications and Theoretical Aspects. Springer Berlin Heidelberg, pp 572-583. doi:10.1007/978-3-642-14400-4_44

Cortez, P., 2015. rminer: Data Mining Classification and Regression Methods. R package version 1.4.1.

Hyndman, R.J., O'Hara-Wild, M., Bergmeir, C., Razbash, S., and Wang, E., 2017. forecast: Forecasting functions for time series and linear models. R package version 7.1.

Hyndman, R.J., and Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 27 (3), 1-22. doi:10.18637/jss.v027.i03

Koutsoyiannis, D., Yao, H., and Georgakakos, A., 2008. Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods. *Hydrological Sciences Journal*, 53 (1), 142-164. doi:10.1623/hysj.53.1.142

Krause, P., Boyle, D.P., and Bäse F., 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5, 89-97.

Lawrimore, J.H., Menne, M.J., Gleason, B.E., Williams, C.N., Wuertz, D.B., Vose, R.S., and Rennie, J., 2011. An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, 116 (D19121). doi:10.1029/2011JD016187

Maier, H.R., and Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*, 15 (1), 101-124. doi:10.1016/S1364-8152(99)00007-9

Papacharalampous, G.A., 2016. Theoretical and empirical comparison of stochastic and machine learning methods for hydrological processes forecasting. MSc thesis.

Papacharalampous, G.A., Tyralis, H., and Koutsoyiannis, D., 2017. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. In preparation.

Peterson, T.C., and Vose, R.S., 1997. An overview of the Global Historical Climatology Network temperature database. *Bulletin of the American Meteorological Society*, 78 (12), 2837-2849. doi:10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2

Raghavendra, N.S., and Deka, P.C., 2014. Support vector machine applications in the field of hydrology: a review. *Applied Soft Computing*, 19, 372-386. doi:10.1016/j.asoc.2014.02.002

Taieb, S.B., Bontempi, G., Atiya, A.F., and Sorjamaa, A., 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications*, 39 (8), 7067-7083. doi:10.1016/j.eswa.2012.01.039

Tyralis, H., 2016. HKprocess: Hurst-Kolmogorov Process. R package version 0.0-2.

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. doi:10.1007/s00477-010-0408-x

Valipour, M., Banihabib, M.E., and Behbahani, S.M.R., 2013. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology*, 476 (7), 433-441. doi:10.1016/j.jhydrol.2012.11.017

Zambrano-Bigiarini, M., 2014. hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3-8.

Zeileis, A., and Grothendieck, G., 2005. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14 (6), 1-27. doi:10.18637/jss.v014.i06