Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes

Georgia A. Papacharalampous*, Hristos Tyralis and Demetris Koutsoyiannis

Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Iroon Polytechniou 5, 157 80 Zografou, Greece

* Corresponding author, papacharalampous.georgia@gmail.com

Abstract: We perform an extensive comparison between 11 stochastic to 9 machine learning methods regarding their multi-step ahead forecasting properties by conducting 12 large-scale computational experiments. Each of these experiments uses 2 000 time series generated by linear stationary stochastic processes. We conduct each simulation experiment twice; the first time using time series of 110 values and the second time using time series of 310 values. Additionally, we conduct 92 real-world case studies using mean monthly time series of streamflow and particularly focus on one of them to reinforce the findings and highlight important facts. We quantify the performance of the methods using 18 metrics. The results indicate that the machine learning methods do not differ dramatically from the stochastic, while none of the methods under comparison is uniformly better or worse than the rest. However, there are methods that are regularly better or worse than others according to specific metrics.

Key Words: multi-step ahead forecasting; neural networks; random forests; stochastic vs machine learning models; support vector machines; time series

© 2017 by the author(s). Distributed under a Creative Commons CC BY license.

1. Introduction

1.1 Time series forecasting in hydrology and beyond

Point forecasting (hereafter, "forecasting", unless specified differently) is of great importance in operational hydrology (Wang et al. 2009). Moreover, the scientific interest accompanying this practical interest is largely reflected in the literature and mainly motivated by the fact that all forecasting tasks share a particular nature upheld by the pursuance of properly modelling the future behaviour of the process under investigation. Admittedly, nobody could ever dispute the challenging character of forecasting (*"prediction is very difficult, especially about the future"*), while it would be rather impractical to despise a forecast that was proven useful in practice.

A classification of the available methodologies for time series forecasting regarding the forecasting horizon is one- and multi-step ahead forecasting. The accumulation of errors in multi-step ahead forecasting renders the latter far more demanding than one-step ahead forecasting. There are five strategies for multi-step ahead forecasting, i.e. the recursive, direct, DirRec, MIMO and DIRMO (Taieb et al., 2012; Bontempi et al., 2013). One- and multi-step ahead forecasting are both frequent practices in hydrology. Examples of one-step ahead forecasting of hydrological processes include Lambrakis et al. (2000), Ballini et al. (2001), Yu et al. (2004), Yu and Liong (2007), Hong (2008), Koutsoyiannis et al. (2008) and Tran et al. (2015). Several studies performing multi-step ahead forecasting are Ballini et al. (2001), Kim and Valdés (2003), Asefa et al. (2005), Khan and Coulibaly (2006), Lin et al. (2006), Cheng et al. (2008), Guo et al. (2011) and Valipour et al. (2013). In these studies, time series exhibiting seasonality are analysed. Furthermore, Hu et al. (2001) and Tongal and Berndtsson (2016) perform multi-step ahead forecasting of time series without seasonal behaviour.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Regardless of the methodology adopted and due to the stochastic nature of forecasting, it is well known in advance that all forecasts will be finally proven wrong. Despite this fact, researchers have long been chasing the most accurate forecast for their data, a universally best technique. On the other hand, there is an argument that it is the data and the application of interest that determine the proper methodology for each case, rather than vice versa (Hong and Fan 2016). Another argument is that perhaps research should invest more on probabilistic forecasting (e.g. using Bayesian statistics as in Tyralis and Koutsoyiannis 2014) and less on point forecasting (Krzysztofowicz 2001).

In fact, the opinions on forecast evaluation are often diverging, as they tend to depend on the perspective from which the forecasts are examined. An interesting study on this subject can be found in Murphy (1993). The latter identifies three criteria for this specific evaluation, which are adopted as a foundation for further discussion in later studies, e.g. Ramos et al. (2010) and Weijs et al. (2010). These criteria are (1) the consistency during the forecasting process, (2) the quality or the correspondence between the forecasts and the target values and (3) the value or the profit that the forecast provide to the decision makers. Weijs et al. (2010) note that criterion (2) concerns more the pure science, while criterion (3) is closer related to the decisions made within the engineering applications (of science), rather than science itself. Thus, only a few studies are dedicated to criterion (3), such as Ramos et al. (2010) and Ramos et al. (2013), while the greatest part of the literature focuses on criterion (2). The latter likewise applies to the present study and to all of its references aiming to deal with the modelling issue (*which model should I use?*) within specific hydrological concepts.

Right after the introduction of the currently classical Autoregressive Integrated Moving Average (ARIMA) models by Box and Jenkins (1968), Carlson et al. (1970) used several stationary models of this specific family to forecast the evolution of four annual time series of streamflow processes. Today the available models for time series forecasting are numerous and can be classified according to De Gooijer and Hyndman (2006) into eight categories, i.e. (a) exponential smoothing, (b) ARIMA, (c) seasonal models, (d) state space and structural models and the Kalman filter, (e) nonlinear models, (f) long-range dependence models, e.g. the family of Autoregressive Fractionally Integrated Moving Average (ARFIMA) models, (g) Autoregressive Conditional Heteroscedastic/Generalized Autoregressive Conditional Heteroscedastic (ARCH/GARCH) models and (h) count data forecasting. The models from the categories (a)-(g) are of potential interest in hydrology.

The theoretical properties of the models of categories (a)-(d), (f), (g) (hereafter, referred to as "stochastic") more or less have been investigated, in contrast to those of the nonlinear models and in particular the Machine Learning (ML) algorithms, also referred to in the literature as black-box models. These two main categories of models are known to represent two different cultures in statistical modelling, the data modelling culture and the algorithmic modelling culture (Breiman 2001b). The former assumes that an analytically formulated stochastic model is behind the generation of the data, while the latter that behind this process is something complex and unknown, which does not have to be analytically formulated, as long as a purely algorithmic model can offer high forecast accuracy. In other words, profoundly understanding and properly modelling the (future) behaviour of a process are strongly connected within the data modelling culture, but completely irrelevant within the algorithmic modelling culture. The distinction between causal explanation, prediction and description is acknowledged and clarified in terms of modelling in Shmueli (2010). Still, one could question whether the (rather artificial) separation of models with respect to the "stochastic-ML dipole" actually corresponds to a striking difference in their forecasting performance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

What cannot be questioned, on the other hand, is the popularity that the various ML forecasting methods have gained in many scientific fields, including hydrology. Amongst the most popular ML algorithms are the Neural Networks (NN), the Random Forests (RF) and the Support Vector Machines (SVM). The SVM are presented in their current form by Cortes and Vapnik (1995) (see also Vapnik 1995, 1999), while the RF by Breiman (2001a). For the implementation of the NN for time series forecasting the reader is referred to Zhang et al. (1998) and Zhang (2001). Regarding the use of the SVM for this specific purpose, a review can be found in Sapankevych and Sankar (2009). The large number of the relevant applications of the NN and SVM algorithms in the field of hydrology is imprinted in Maier and Dandy (2000) and Raghavendra and Deka (2014) respectively, while the RF algorithms are barely used for the forecasting of hydrological processes.

The long list of studies using NN algorithms for hydrological forecasting tasks within a specific hydrological context includes Atiya et al. (1999), Lambrakis et al. (2000), Kişi (2007), Cheng et al. (2008) and Yaseen et al. (2016), while the reader can find relevant studies using SVM algorithms in Sivapragasam et al. (2001), Shi and Han (2007), Kişi and Cimen (2011) and Lu and Wang (2011); also some critical comments for such studies have been raised by Koutsoyiannis (2007). Furthermore, the literature contains a large number of studies proposing hybrid forecasting methods, e.g. Hu et al. (2001), Kim and Valdés (2003), Yu and Liong (2007) and Hong (2008), which can be for example combinations of ARMA (Autoregressive Moving Average) and ML algorithms. Moreover, a common practice in hydrology and beyond is to compare several ML methods on a few given observed time series and infer about the optimal model. Comparisons between NN and SVM can be found in Liong and Sivapragasam (2002), Khalil et al. (2006), Pai and Hong (2007), Hong (2008), Guo et al. (2011), Kişi and Cimen (2012), Kalteh (2013) and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

He et al. (2014). Also frequent are the comparisons between ML and stochastic forecasting methods (see Section 1.2.1), in which the latter are almost always used as benchmarks.

Most of the studies mentioned in the previous paragraph present forecasts about the behaviour of streamflow processes, i.e. Atiya et al. (1999), Lambrakis et al. (2000), Kişi (2007), Shi and Han (2007), Yu and Liong (2007), Cheng et al. (2008), Guo et al. (2011), Kişi and Cimen (2011), Kalteh (2013), He et al. (2014) and Yaseen et al. (2016). Another type of process attracting scientific and practical forecasting interest is precipitation (e.g. Sivapragasam et al. (2001), Hong (2008), Pai and Hong (2007), Lu and Wang (2011), Sivapragasam et al. (2001) and Kişi and Cimen (2012)). Nevertheless, as emphasized in Zaini et al. (2015) precipitation and streamflow are amongst the most difficult geophysical processes to forecast.

1.2 Comparison of stochastic and machine learning forecasting methods

1.2.1 A literature survey with a focus on hydrology

Research within the field of hydrology often focuses on comparing ML forecasting methods to stochastic, while the comparisons performed are all based on case studies. For instance, Jain et al. (1999) use a time series of average monthly streamflow to compare a NN model and an ARIMA model regarding their one-step ahead forecasting properties and Kişi (2004) use another time series of the same type to compare several NN and AR (Autoregressive) models regarding their multi-step ahead forecasting properties. Admittedly, most of the relevant studies compare NN to ARIMA models, such as Ballini et al. (2001), Mishra et al. (2007), Abudu et al. (2010), Kişi et al. (2012) and Valipour et al. (2013). The same applies to Khan and Coulibaly (2006), Lin et al. (2006), Wang et al. (2009), Shabri and Suhartono (2012), Belayneh et al. (2014) and Patel and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Ramachandran (2014), albeit the latter studies also include SVM methods in the comparisons.

Apart from Khan and Coulibaly (2006), Mishra et al. (2007), Kişi et al. (2012) and Belayneh et al. (2014) all studies mentioned right above use monthly time series of streamflow processes. Similarly, Yu et al. (2004) compare several forecasting methods, including an ARIMA model and SVM, on two daily time series of runoff and Tongal and Berndtsson (2016) compare several stochastic and ML forecasting methods on three time series of streamflow processes. Additionally, in Chen et al. (2012) the reader can find one of the few studies using RF for hydrological forecasting tasks. The latter algorithm is compared to ARIMA models on four time series regarding its multi-step ahead forecasting properties.

Furthermore, Koutsoyiannis et al. (2008) conduct a sufficient number of one-month ahead forecasting experiments on a monthly time series of Nile flows, which is 131 years long, to compare the performance of several stochastic and ML methods. It is concluded that the forecasting methods from both main categories seem to be competent in time series forecasting, while it is emphasized that stochastic models that depart from the "typical ARMA formalism" might outperform in explaining the behaviour of natural processes and, thus, they might be more appropriate for simulation.

Finally, in Papacharalampous et al. (2017) a multiple-case study using 50 monthly time series of precipitation or temperature processes is conducted for the comparison of the forecasting performance of four stochastic and two ML algorithms. As regards the ML algorithms the effect on the forecasting quality of the time lag selection and the hyperparameter optimization is also investigated. One- and multi-step ahead forecasting experiments are conducted and six performance criteria are examined. Cross-case synthesis is performed for the detection of similarities across the individual cases. All

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

forecasting methods, either stochastic or ML, seem to perform equally well in this study, each being better or worse than the rest depending on the individual case and the criterion of interest.

1.2.2 The broader perspective

Each forecasting method has several theoretical properties, which render it better or worse than other forecasting methods within a specific context. The latter is determined by various factors, such as the process under examination, the available time series data and its observation frequency, the forecasting methods and the criteria used for the evaluation of forecast quality. Although this specific fact becomes clearly evident even through a small-scale literature survey, there are still very essential theoretical questions concerning hydrological time series forecasting that remain unanswered in the literature. The answers to those questions are, by nature, context-independent and, thus, require the use of a different research method from the case study, which serves ideally purposes of context-dependent research.

Generally, the single-case study method is suitable for studying a context-dependent phenomenon in detail and, therefore, its implementation can provide useful insights (Yin 1994). However, single-case studies do not allow generalizations to any extend (Achen and Snidal 1989) and their results cannot even stand alone as evidence for theory discovery. On the contrary, a multiple-case study can offer contingent empirical evidence on theoretical issues (Achen and Snidal 1989), mainly because of its comparative character. Prescribed by evidence or not, generalizations can be derived either analytically (most preferably) or empirically by designing and conducting experiments, which use real-world and/or simulated data in conjunction with statistical methods (context-independent research). Moreover, the multi-method of research

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

resulting from the integration of context-independent and context-dependent research methods can deliver highly robust results (Gable 1994). Within this specific research method the case study functions complementary to the context-independent research method for serving purposes of theory validation as also for emphasizing on specific aspects that would stay hidden within the statistical representation of the phenomena under investigation.

Regarding the so far conducted comparisons between forecasting methods, their majority in all scientific fields is based on case studies. Nevertheless, in some few cases beyond the field of hydrology the number of the examined time series is quite large. These time series are realizations of several phenomena, which however are fundamentally different from being hydrological, and their examination includes concepts that are rather inappropriate in hydrological terms (e.g. paying attention to small quantitative differences in the forecasting performance of the methods). Examples of such studies can be found in Makridakis and Hibon (1987), Makridakis and Hibon (2000) and Ahmed et al. (2010), which examine 1 001, 3 003 and 1 045 time series respectively. Within these studies a statistical analysis is performed and the results are presented correspondingly. Furthermore, the literature includes two studies (Zhang 2001; Thissen et al. 2003) in which the performance of the methods is assessed on simulated time series from linear stochastic processes. The scale of the simulation experiment is small in both cases. Thissen et al. (2003) examine one long time series from the ARMA family, while Zhang (2001) examine 8 stochastic processes from the ARMA family and 30 simulated time series for each stochastic process. The forecasting methods are ARMA models, NN and SVM in the former study and ARMA models and NN in the latter study, while Makridakis and Hibon (1987), Makridakis and Hibon (2000) and Ahmed et al. (2010) do not focus their comparisons on the stochastic-ML dipole.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Admittedly, the studies mentioned in the previous paragraph as well as Papacharalampous et al. (2017) pursue generalized results to greater extent than most of the available studies. However, the gap still remains. What specifically needs to be addressed, amongst other context-independent research questions, is whether the stochastic-ML dipole actually corresponds to a clear difference in the forecasting performance of the methods, especially in the light of published studies, which claim that they found a technique, which is better than others are. Given the fact that each forecasting case is indisputably unique, this task would necessarily require the examination of a sufficiently large and representative sample of forecasting cases within the same (properly designed) methodological framework.

Extensive simulations combined with statistical analysis and benchmarking can constitute, nevertheless, a highly effective approach to solving the problem under discussion on a theoretical basis. In more detail, for the generalized comparison of stochastic and ML forecasting methods, a sufficient number of different and representative of the underlying phenomena time series could be used for the estimation of the expected performance of several forecasting methods regarding several criteria of interest. The need of using simulated time series to assess the performance of forecasting methods is emphasized by forecasting experts (Bontempi 2013). The analytical approach in assessing the performance of ML algorithms is usually not possible, therefore the only alternative approach is using simulations. Concerning the testing procedure, while the available metrics for the assessment of the forecast quality are a lot, most of the studies use only a few (Krause et al. 2005), understating the importance of the testing process despite relevant suggestions (e.g. Humphrey et al. 2017). Similarly, the number of the compared forecasting methods is usually small, although benchmarks are commonly included in the relevant comparisons

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

(Pappenberger et al. 2015). Apparently, the larger the scale of the simulation experiments, the more general would be the results.

1.2.3 The present study

In the context described in the above sections, we perform an extensive comparison between several stochastic and ML methods for the forecasting of hydrological processes by conducting large-scale computational experiments based on simulations. The comparison refers to the multi-step ahead forecasting properties of the methods, although one-step ahead forecasting is also of practical and scientific interest. The reason for this specific decision is that multi-step ahead forecasting constitutes a greater challenge than one-step ahead forecasting. The time series are generated by linear stationary stochastic processes, which are commonly used for modelling hydrological processes. In fact, stationary models, in contrast to the non-stationary, are established as the appropriate modelling choice when dealing with natural processes, unless tangible and quantitative information that can fully support a deterministic description (not based on data but on physical laws) of change in time are available (Koutsoyiannis 2011; Koutsoyiannis and Montanari 2015). Additionally to the simulation experiments, we individually examine several real-world time series and specifically focus on one of them to reinforce the findings and highlight important facts. Our aim is to fill the gap detected in the literature by providing generalized results on several questions that have attracted the attention in the field of hydrological time series forecasting, with an emphasis on the comparison of stochastic and ML forecasting methods.

To increase our aim's feasibility, we have put extra effort into the design of our simulation experiments. Most significantly, we simulate a large number of time series to ensure that the latter compose a wide range of different cases. Furthermore, we use a sufficient, but still manageable, number of forecasting methods from both the stochastic and the ML categories. Another important methodological point is the selection of popular algorithms. In more detail, the stochastic algorithms originate from the families of ARIMA, ARFIMA, state space and exponential smoothing. Likewise, the ML algorithms are NN, RF and SVM. We note that the ARIMA, ARFIMA, NN and SVM models are widely used within the field of hydrology, while the use of the remaining upper listed algorithms in the present study is rather innovative. Finally, our methodological approach to the problem also includes an adequate number of metrics, corresponding to several criteria, for the comparative assessment of the forecasting methods. Some of these metrics have been introduced for the evaluation of hydrological forecasts.

The main methodological elements (time series, forecasting methods and metrics) are combined into a simple yet promising and unique methodological framework. The latter emphasizes on answering several theoretical research questions, but it also provides quantification of the forecasting methods' performance on linear processes as also material for the evaluation of the utility of the metrics. For the about 13 000 figures, conducted in the context of an exploratory visualization, as well as the numerical summaries of the results the reader is referred to the fully reproducible reports, which are available together with their codes in the Supplementary material. In fact, reproducibility is a primary consideration of the present study, as it can promote scientific progress in a reliable manner by allowing validation tests to be made (LeVeque et al. 2012). Therefore, we encourage the reader to use the codes to reproduce our analyses as also to perform further experimentation conducting case studies of his/her interest. The preliminary research for this paper was conducted for the Postgraduate Thesis of the first author (Papacharalampous 2016).

2. Methodology

In Section 2 we present the basic methodological elements of this study, while the reader is referred to the Appendices for a brief theoretical background, as also to the scientific literature for a more complete coverage of the relevant theory.

2.1 Simulated processes

We simulate time series according to several stochastic models from the frequently used families of ARMA and ARFIMA. Although this specific modelling is accompanied by certain problems (Koutsoyiannis 2016), it is considered rather satisfying for the generalization pursued here and has been widely applied in hydrology (e.g. Montanari et al. 1997, 2000). The simulated stochastic processes are presented in Table 1, while for the related definitions the reader is referred to Appendix A. We use the arima.sim built in R algorithm (R Core Team 2017) to simulate the ARMA(p,q) processes and the fracdiff.sim algorithm of the fracdiff R package (Fraley et al. 2012) to simulate the ARFIMA(p,d,q) processes.

Table 1. Simulated stochastic processes of the present study. Their definitions are given in Appendix A.

s/n	Stochastic model	Parameters of the stochastic model
1	AR(1)	$\varphi_1 = 0.7$
2	AR(1)	$\varphi_1 = -0.7$
3	AR(2)	$\varphi_1 = 0.7$, $\varphi_2 = 0.2$
4	MA(1)	$\theta_1 = 0.7$
5	MA(1)	$\theta_1 = -0.7$
6	ARMA(1,1)	$\varphi_1 = 0.7$, $\theta_1 = 0.7$
7	ARMA(1,1)	$\varphi_1 = -0.7$, $\theta_1 = -0.7$
8	ARFIMA(0,0.45,0)	
9	ARFIMA(1,0.45,0)	$\varphi_1 = 0.7$
10	ARFIMA(0,0.45,1)	$\theta_1 = -0.7$
11	ARFIMA(1,0.45,1)	$\varphi_1 = 0.7$, $\theta_1 = -0.7$
12	ARFIMA(2,0.45,2)	$\varphi_1 = 0.7$, $\varphi_2 = 0.2$, $\theta_1 = -0.7$, $\theta_2 = -0.2$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

2.2 Real-world time series

We individually examine 92 mean monthly time series of streamflow processes (Str_1, Str_2, …, Str_92), which originate from catchments in Australia and are a subset of a larger dataset (Peel et al. 2000), containing at least 10 and up to 93 years of continuous observations with a median value of 25 years. We use the deseasonalized time series for the application of the forecasting methods, as suggested for the improvement of forecast quality (Taieb et al. 2012; Hyndman and Athanasopoulos 2013).

To describe the long-term persistence of each deseasonalized time series we estimate the Hurst parameter H for it using the mleHK algorithm of the HKprocess R package (Tyralis 2016), which implements the maximum likelihood method (Tyralis and Koutsoyiannis 2011). The parameter H takes values in the interval (0,1). The larger it is the larger the long-range dependence of the Hurst - Kolmogorov stochastic process, which is widely used for the modelling of geophysical processes instead of the ARFIMA(0,d,0) model. The estimated values range between 0.56 and 0.99 with a mean value of 0.78. For detailed information about the real-world time series of the present study the reader is referred to the Supplementary material.

Among these 92 time series, in the present manuscript we focus on Str_77, which is presented in Figure 1. The length of this specific time series is 42 years (January 1955 to December 1996) or 504 observations. For its exploration we calculate the sample Autocorrelation Function (ACF) and the sample Partial Autocorrelation Function (PACF), which are also presented in Figure 1. The respective H estimate equals 0.90. The use of this specific time series is not dictated from any significant reason. In fact, we could have selected any time series of our interest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

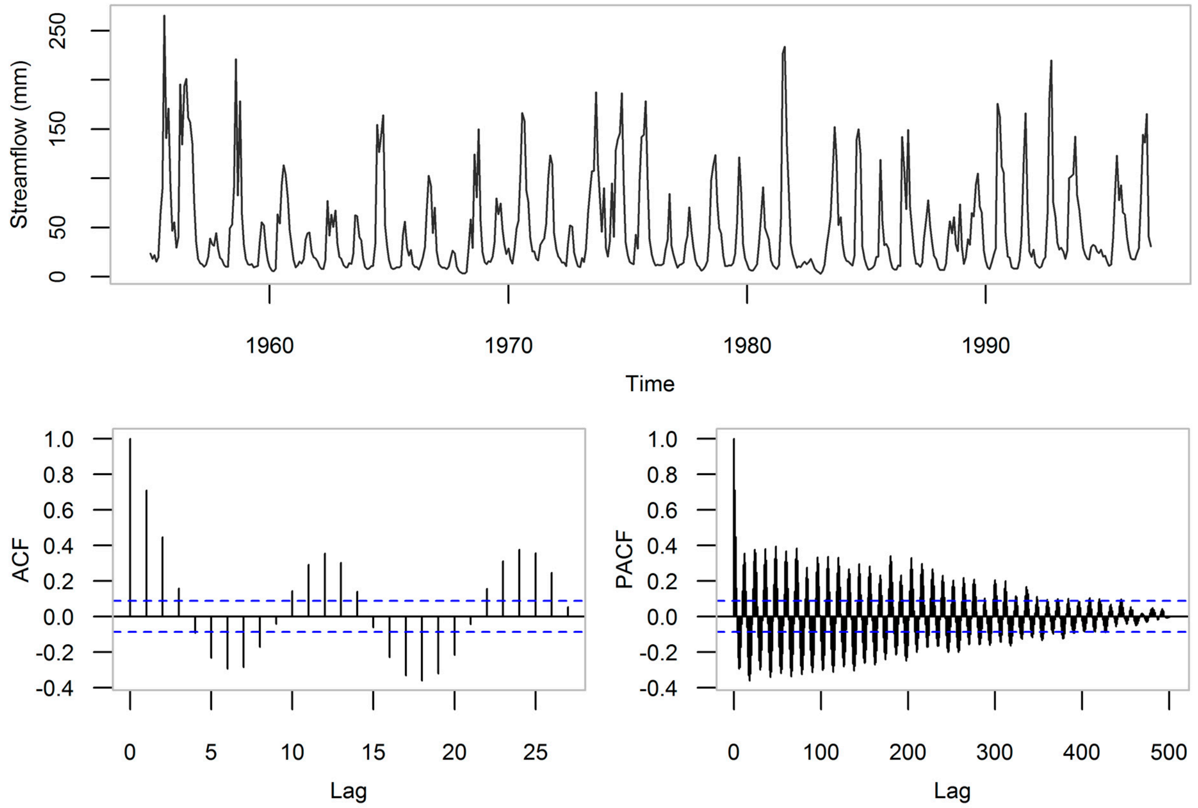

Figure 1. The Str_77 time series. Data source: Peel et al. (2000).

2.3 Forecasting methods

We compare 11 stochastic to 9 ML forecasting methods. The stochastic methods are classified into five main categories as presented in Table 2. Similarly, the ML methods are classified into three main categories as presented in Table 3 and Table 4. For the implementation of the forecasting methods the reader is referred to the Supplementary material.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Table 2. Stochastic forecasting methods. The forecasting methods are available in code form in the Supplementary material.

s/n	Abbreviated name	Category
1	Naïve	Simple
2	RW	
3	ARIMA_f	ARIMA
4	ARIMA_s	
5	auto_ARIMA_f	
6	auto_ARIMA_s	
7	auto_ARFIMA	ARFIMA
8	BATS	State space
9	ETS_s	
10	SES	Exponential smoothing
11	Theta	

Table 3. ML forecasting methods. The time lag selection procedures adopted are defined in Table 4. The forecasting methods are available in code form in the Supplementary material.

s/n	Abbreviated name	Category	Model structure information	Hyperparameter optimized (possible values)	Time lag selection procedure
1	NN_1	NN	Single hidden layer Multilayer Perceptron (MLP)	Number of hidden nodes (0, 1, …, 15)	1
2	NN_2				2
3	NN_3				3
4	RF_1	RF	Breiman' s random forests algorithm with 500 grown trees	Number of variables randomly sampled as candidates at each split (1, …, 5)	1
5	RF_2				2
6	RF_3				3
7	SVM_1	SVM	Radial Basis kernel "Gaussian" function, C = 1, epsilon = 0.1	Sigma inverse kernel width (2^n, n = -8, -7, …, 6)	1
8	SVM_2				2
9	SVM_3				3

Table 4. Time lag selection procedures adopted for the ML methods. The forecasting methods are available in code form in the Supplementary material.

s/n	Time lags
1	The corresponding to an estimated value for the ACF using the acf R algorithm (built in R algorithm), i.e. the time lags 1, …, 20 for a time series of 100 values and the time lags 1, …, 24 for a time series of 300 values
2	The corresponding to a statistical important estimated value for the ACF using the acf R algorithm (built in R algorithm). If there is no statistical important estimated value for the ACF, the corresponding to the largest estimated value
3	According to the nnetar R function (package forecast), i.e. the time lags 1, …, n, where n is the number of AR parameters that are fitted to the time series data using the ar R algorithm (built in R algorithm)

We use two simple forecasting methods in the comparisons. The Naïve forecasting method, one of the most commonly used benchmarks (Hyndman and Athanasopoulos 2013; Pappenberger et al. 2015), simply sets all forecasts equal to the last value. The RW forecasting method, a variation of the Naïve forecasting method, is equivalent to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

drawing a line between the first and the last value and extrapolating it into the future (Hyndman and Athanasopoulos 2013). The stochastic methods also include the ARIMA and ARFIMA methods. These five methods apply the maximum likelihood method to estimate the values of the parameters of the AR and MA parts of the models. For the ARIMA_f and ARIMA_s forecasting methods the numbers of the AR (*p*) and MA (*q*) parameters are set to be the same to those used in the simulated processes, while the number of differencing (*d*) is set to be zero. The auto_ARIMA_f and auto_ARIMA_s methods estimate the values of *p*, *d*, *q* of the ARIMA model using the Akaike Information Criterion with a correction for finite sample sizes (AICc), as described in Hyndman and Athanasopoulos (2013). The same applies to the auto_ARFIMA method for the estimation of the values of *p*, *d*, *q* of the ARFIMA models.

The BATS and ETS_s forecasting methods use the point forecasts from an exponential smoothing state space model with several key features, i.e. capability of performing Box-Cox transformation and/or including ARMA errors correction, Trend and Seasonal components (BATS), also allowing an optimal model selection using the Akaike Information Criterion (AIC), and an exponential smoothing state space model with automatic selection of the Error, Trend and Seasonal components (ETS) respectively. We additionally include the SES (Simple Exponential Smoothing) and Theta forecasting methods in the comparisons. The latter method was presented by Assimakopoulos and Nikolopoulos (2000) and had the best performance in the M3-Competition, during which it was applied to 3 003 historical time series from various categories of data (Makridakis and Hibon 2000). The reader is referred to Hyndman et al. (2008) and Hyndman and Athanasopoulos (2013) for the theoretical background of the exponential smoothing and space state models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Regarding the NN, the RF and the SVM forecasting methods, there are some additional concerns to the selection of the algorithms, originating from the nature of the ML methods. The choices to be considered for the selection of the time lags used to build the regression matrix (input data matrix), as well as the choices for the values of the hyperparameters of the models (e.g. the hidden nodes in a NN model), are many. Usually, hyperparameters are not directly decided by the ML algorithm during the fitting process. A fact is that the ML models are by design rather more flexible than needed in most cases and, thus, hyperparameter optimization is often used to detect and prevent overfitting as much as possible. In Tables 3 and 4 we summarize the basic information about the model structures, the hyperparameter optimization and the time lag selection procedures adopted.

We apply the stochastic methods using mainly the R package forecast (Hyndman and Khandakar 2008, Hyndman et al. 2017) and the ML methods using the R package rminer (Cortez 2010, 2016) and the nnetar algorithm from the R package forecast, as also several built in R algorithms. The R package rminer uses the nnet algorithm of the nnet R package (Venables and Ripley 2002), the randomForest algorithm of the randomForest R package (Liaw and Wiener 2002) and the ksvm algorithm of the kernlab R package (Karatzoglou et al. 2004) for the application of the NN, the RF and the SVM methods respectively.

2.4 Metrics

The metrics used for the comparative assessment of the forecasting methods are classified into five main categories according to the criterion presented in Table 5. They provide assessment regarding two types of accuracy, the capture of the variance and the correlation. By Type 1 accuracy we mean the closeness of the forecasted time series to the actual, while by Type 2 accuracy we mean the closeness of the mean of the forecasted values of each time series to the mean of the actual ones. The definitions of the metrics are listed in Appendix B, while the reader is also referred to Nash and Sutcliffe (1970), Kitanidis and Bras (1980), Yapo et al. (1996), Krause et al. (2005), Criss and Winston (2008), Gupta et al. (2009), Zambrano-Bigiarini (2014) for further information.

Table 5. Metrics used in the present study; their definitions are given in Appendix B.

s/n	Abbreviated Name	Full name	Criterion
1	MAE	Mean Absolute Error	Type 1 accuracy
2	MAPE	Mean Absolute Percentage Error	
3	RMSE	Root Mean Square Error	
4	NSE	Nash-Suttcliffe Efficiency	
5	mNSE	Modified Nash-Suttcliffe Efficiency	
6	rNSE	Relative Nash-Suttcliffe Efficiency	
7	cp	Persistence Index	
8	ME	Mean Error	Type 2 accuracy
9	MPE	Mean Percentage Error	
10	PBIAS	Percent Bias	
11	VE	Volumetric Efficiency	
12	rSD	Ratio of Standard Deviations	Capture of the variance
13	Pr	Pearson’ s Correlation Coefficient	Correlation
14	r2	Coefficient of Determination	
15	d	Index of Agreement	Type 1 accuracy, capture of the variance
16	md	Modified Index of Agreement	
17	rd	Relative Index of Agreement	
18	KGE	Kling-Gupta Efficiency	Type 2 accuracy, capture of the variance, correlation

2.5 Methodology outline

For the comparison of the forecasting methods (see Section 2.3) we conduct 12 large-scale computational experiments based on simulations. Within each of the latter we simulate 2 000 time series according to a model of a stochastic process (see Section 2.1). We conduct each computational experiment twice; the first time using simulated time series of 110 values and the second time using simulated time series of 310 values. The simulation experiments conducted are named as presented in Table 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Table 6. Simulation experiments of the present study. The simulated processes are presented in Table 1.

s/n	Code	Simulated process	Length of the time series
1	SE_1a	1	110 values
2	SE_2a	2	
3	SE_3a	3	
4	SE_4a	4	
5	SE_5a	5	
6	SE_6a	6	
7	SE_7a	7	
8	SE_8a	8	
9	SE_9a	9	
10	SE_10a	10	
11	SE_11a	11	
12	SE_12a	12	
13	SE_1b	1	310 values
14	SE_2b	2	
15	SE_3b	3	
16	SE_4b	4	
17	SE_5b	5	
18	SE_6b	6	
19	SE_7b	7	
20	SE_8b	8	
21	SE_9b	9	
22	SE_10b	10	
23	SE_11b	11	
24	SE_12b	12	

Additionally, we conduct 92 real-world case studies using the time series presented in Section 2.2. The case studies are respectively named as CS_1, CS_2, …, CS_92 after the time series Str_1, Str_2, …, Str_92 that they examine. As regards CS_77, we also examine its 4 variations defined in Table 7. We apply the forecasting methods to the simulated and the real-world time series according to Table 8.

Table 7. Variations of the CS_77 case study using parts of the Str_77 time series, which is presented in Figure 1. The *H* parameter is estimated for the deseasonalized time series.

s/n	Variation code	Original time series	Start	End	Length (months)	*H*
1	CS_77_v1	Str_77	Jan 1955	Dec 1995	492	0.90
2	CS_77_v2		Jan 1955	Dec 1994	480	0.89
3	CS_77_v3		Jan 1955	Dec 1993	468	0.89
4	CS_77_v4		Jan 1955	Dec 1992	456	0.89

Table 8. Use of the forecasting methods on the time series.

Forecasting method	ARMA simulated processes	ARFIMA simulated process	Real-world time series
Naive	✓	✓	✓
RW	✓	✓	✓
ARIMA_f	✓	×	×
ARIMA_s	✓	×	×
auto_ARIMA_f	✓	×	×
auto_ARIMA_s	✓	×	×
auto_ARFIMA	×	✓	✓
BATS	✓	✓	✓
ETS_s	✓	✓	✓
SES	✓	✓	✓
Theta	✓	✓	✓
NN_1	✓	✓	✓
NN_2	✓	✓	✓
NN_3	✓	✓	✓
RF_1	✓	✓	✓
RF_2	✓	✓	✓
RF_3	✓	✓	✓
SVM_1	✓	✓	✓
SVM_2	✓	✓	✓
SVM_3	✓	✓	✓

For the application of the stochastic methods we divide each time series into two segments, i.e. the fitting segment and the test segment, which contain n_1 and n_2 values respectively, as indicated in Figure 2. We fit the stochastic models to the former and make predictions corresponding to the latter using the recursive multi-step ahead forecasting method. For the simulation experiments n_2 equals 10, while for the real-world case studies 12. For the application of the ML forecasting methods, we additionally divide the segment of n_1 values into two segments, i.e. the fitting segment (first $[2n_1/3]$ values of the time series) and the validation segment, as indicated in Figure 3. Regarding the real-world time series, the fitting and validation segments are used after mean-value deseasonalization, which is performed using a multiplicative model of time series decomposition. For a brief coverage of the subject of the time series decomposition methods the reader is referred to Hyndman and Athanasopoulos (2013). Moreover, Box-Cox transformation of the real-world data aiming at their normalization also precede the application of most of the stochastic methods. Nevertheless, the ML methods are non-parametric and, thus, they are not affected by the non-normality.

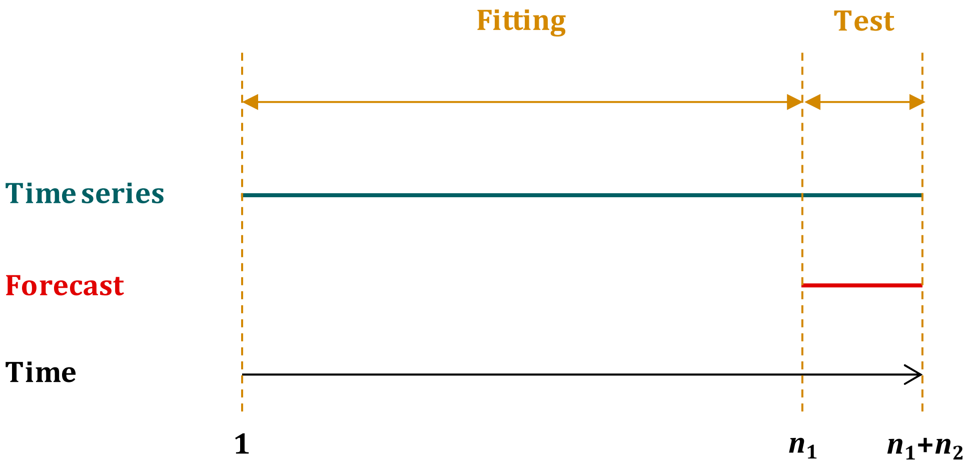

Figure 2. Division of a time series into two segments for the application of the stochastic methods.

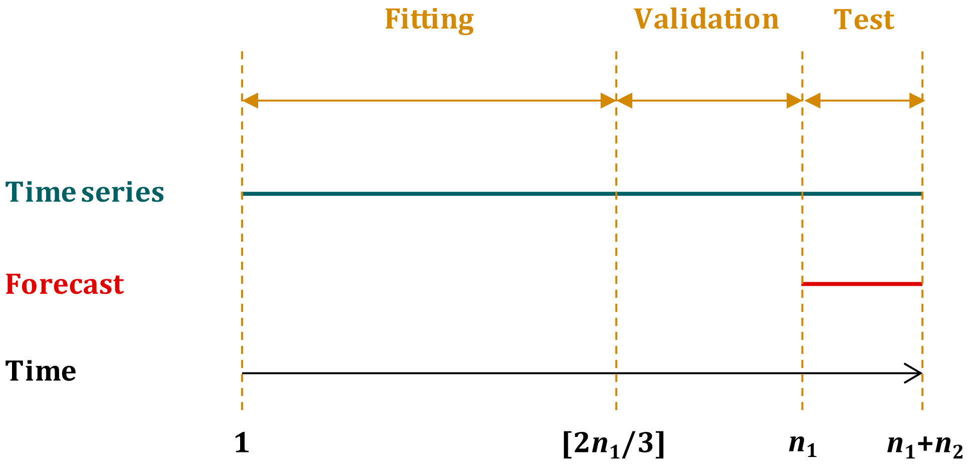

Figure 3. Division of a time series into three segments for the application of the ML methods.

The validation segment serves the hyperparameter optimization procedure, as explained subsequently. We use the fitting segment to fit several ML models that differ only as it comes to the values of a specific hyperparameter. We use each of those models to make predictions corresponding to the validation segment and measure the RMSE of those predictions. Finally, we decide on the value of the hyperparameter, i.e. the corresponding to the model with the smallest RMSE on the validation segment

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

(optimum model). We fit a model with the selected hyperparameter value to data of both the fitting and validation segments and make predictions corresponding to the test segment. Regarding the real-world case studies, we add the seasonality to the predicted time series.

To form a qualitative image of the forecasting methods we first apply several diverse graphical methods on the data sets shaped within each simulation experiment. Here, we only present forecasting examples on two simulated time series of 110 values in Figures 4 and 5, as well as on the Str_77 time series. Figure 6 presents the forecasted time series within the CS_77 case study with seasonality both included and excluded. The concomitant to the addition of seasonality improvement of the forecasts is certainly a noteworthy fact. Furthermore, Figure 7 presents forecasting examples within the examined variations of the CS_77 case study highlighting the uniqueness of each forecasting case. For an extensive graphical exploration of the simulation experiments and the real-world case studies the reader is referred to the Supplementary material.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

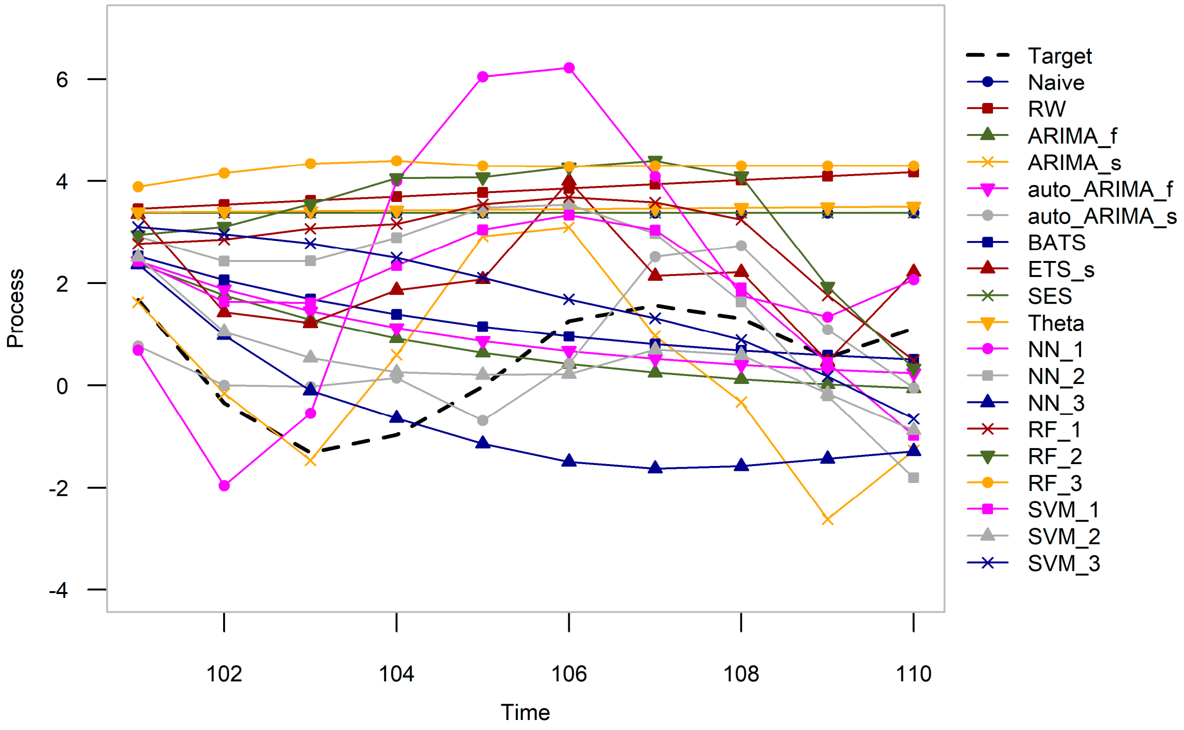

Figure 4. Forecasting examples on a time series resulting from the simulation of an ARMA process within the SE_6a simulation experiment.

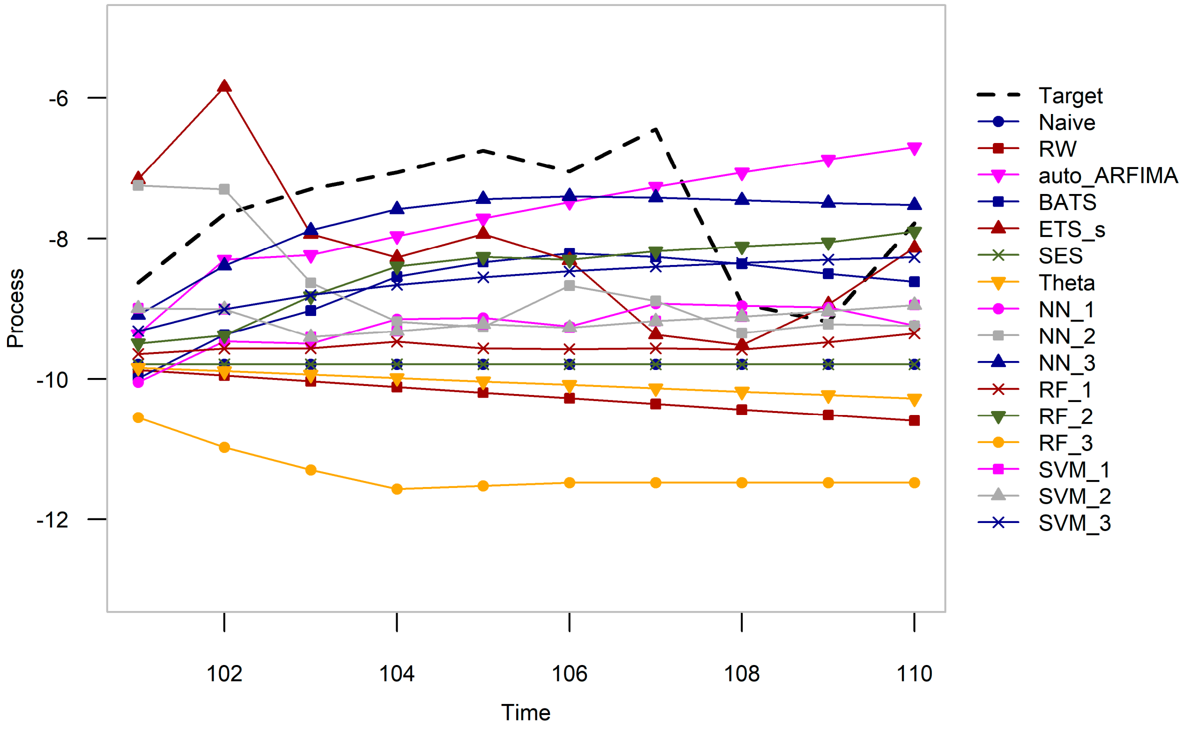

Figure 5. Forecasting examples on a time series resulting from the simulation of an ARFIMA process within the SE_9a simulation experiment.

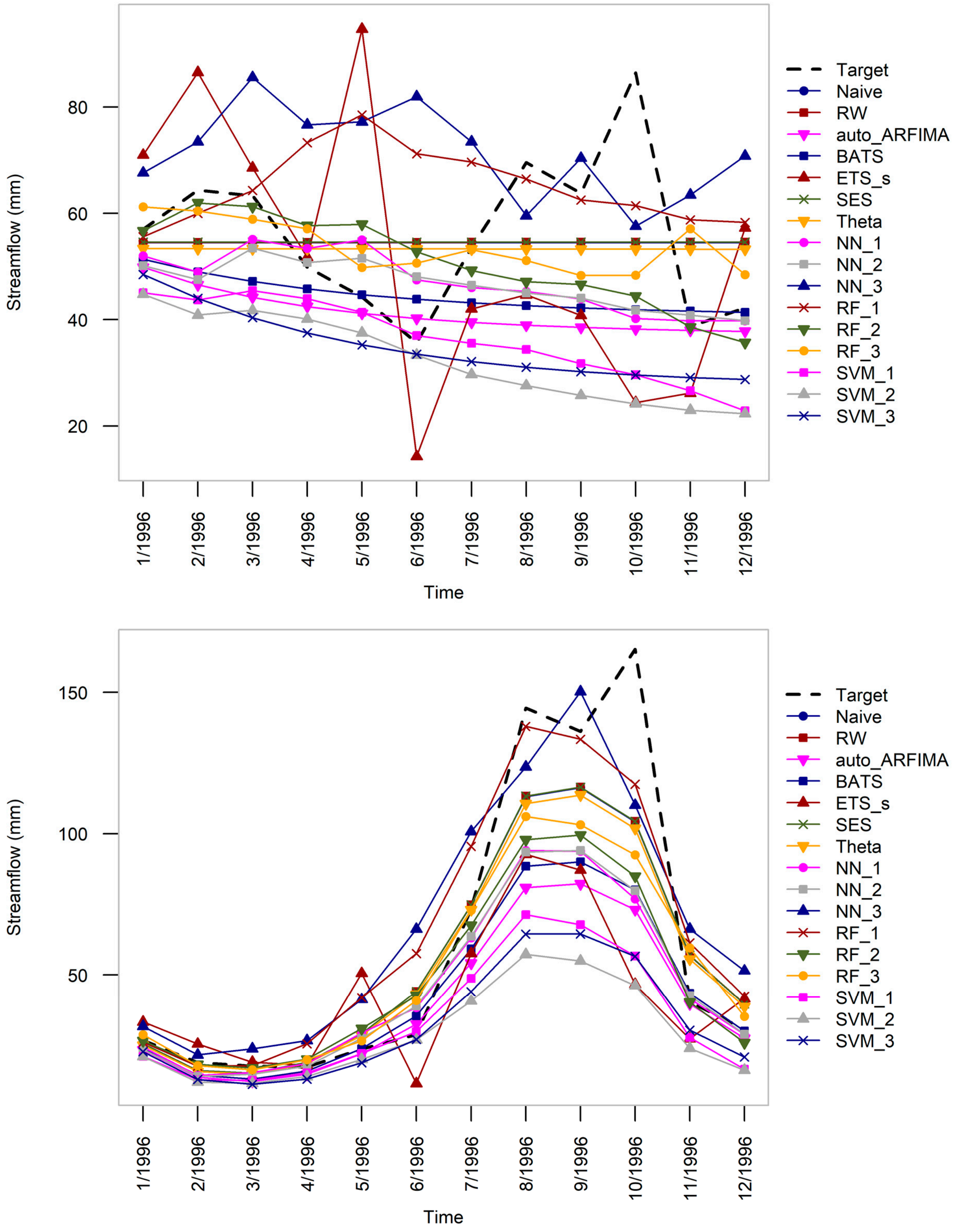

Figure 6. Forecasting examples within the CS_77 case study. Seasonality is excluded from the time series of the upper graph and included in the time series of the lower graph.

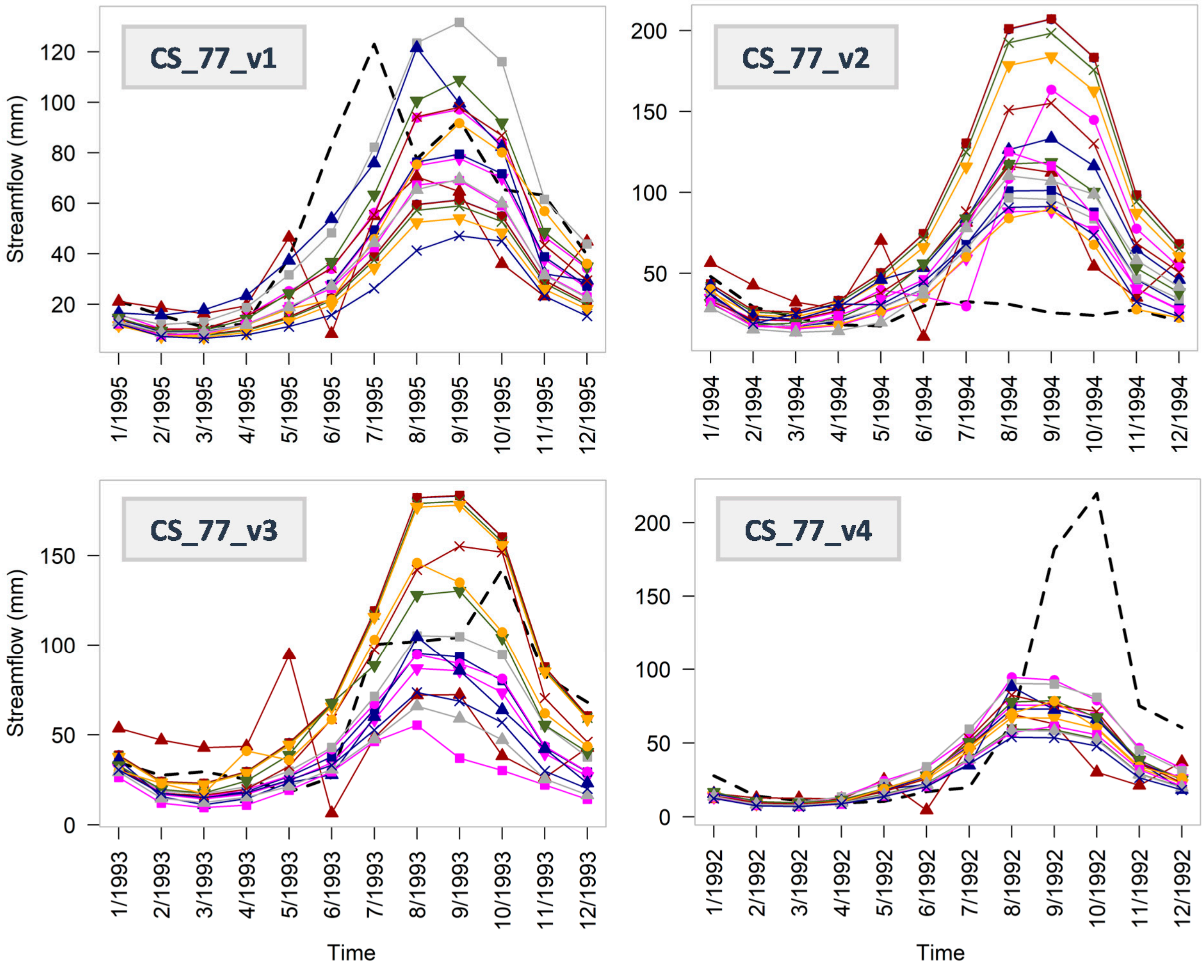

Figure 7. Forecasting examples within the examined variations of the CS_77 case study.

Despite its large interest, this specific visualization cannot support a massive and objective evaluation, which is absolutely essential when pursuing generalized results. Therefore, we compute the values of the metrics presented in Section 2.4 for each forecasting test. The computation takes place on the test segment, which functions as a reference for the comparative assessment of the forecasting methods' performance. Finally, we calculate several descriptive statistics, namely the minimum (min), maximum (max), mean, median, standard deviation (sd), interquartile range (iqr), kurtosis and skewness, of the metric values distributions for all the forecasting tests related to each forecasting method within each forecasting experiment. We use those values for the comparative assessment of the forecasting methods, mainly the medians and iqr values. We compare the medians, as described in Table 9. The same applies to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

the comparison of the values of the each metric computed within the case study, while the smallest the iqr the better the forecasts.

Table 9. Use of the metrics for the comparative assessment of the forecasting methods. The metrics are defined in Appendix B.

Metric	Values	Optimum Value	Condition (the desired)
MAE	[0, +∞)	0	smaller MAE
MAPE	[0, +∞)	0	smaller MAPE
RMSE	[0, +∞)	0	smaller RMSE
NSE	(-∞, 1]	1	larger NSE
mNSE	(-∞, 1]	1	larger mNSE
rNSE	(-∞, 1]	1	larger rNSE
cp	(-∞, 1]	1	larger cp
ME	(-∞,+∞)	0	smaller |ME|
MPE	(-∞,+∞)	0	smaller |MPE|
PBIAS	(-∞,+∞)	0	smaller |PBIAS|
VE	(-∞,+∞)	1	smaller |VE - 1|
rSD	[0, +∞)	1	larger min{rSD, 1/rSD}
Pr	[-1, 1]	1	larger Pr
r2	[0, 1]	1	larger r2
d	[0, 1]	1	larger d
md	[0, 1]	1	larger md
rd	(-∞, 1]	1	larger rd
KGE	(-∞, 1]	1	larger KGE

Although our computational experiments are designed to produce new knowledge in the field of time series forecasting, there are several outcomes rather well known at the forefront of our methodological framework (benchmarking). In more detail, the ARIMA_f and also the auto_ARIMA_f forecasting methods are expected to have the best performance regarding the Type 1 accuracy, mainly in terms of RMSE, on the time series resulting from the simulation of ARMA processes because of their theoretical background (for details see Wei 2006, pp. 88-93). Likewise, this applies to the performance of ARIMA_s and auto_ARIMA_s regarding the capture of the variance exhibited by the time series within the same simulation experiments. Furthermore, the ARIMA_f and ARIMA_s forecasting methods share an additional advantage, since they use by design the *p*, *d*, *q* numbers used in the simulation process. Similarly to the ARIMA_f and auto_ARIMA_f forecasting methods, auto_ARFIMA is expected to be the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

best in terms of RMSE on the time series resulting from the simulation of ARFIMA processes. The five forecasting methods mentioned in the present paragraph, together with the two simple methods, play a key role within our methodological approach.

The results are presented in the context of an exploratory visualization and are available in both quantitative and qualitative forms. All numerical results are provided in table form, while the graphs implemented include histograms, barplots, side-by-side boxplots and heatmaps. This specific context facilitates the comparisons in an efficient way, as it ensures the multifaceted representation of the information provided. Furthermore, the heatmap visualization within and across the various simulation experiments eases the detection of systematic patterns that are closely related to theoretical elements. Therefore, we emphasize on the qualitative form of the results. For the simulation experiments, we mainly focus on two general categories of heatmaps, i.e. the "Type 1 heatmaps" and "Type 2 heatmaps", which use the median values of the metrics. The former type integrates the information provided by all metrics for the average-case performance of the forecasting methods within a specific simulation experiment, while the latter imprints the average-case performance of the forecasting methods across the various simulation experiments according to a specific metric.

3. Results

3.1 Simulation experiments

Section 3.1 aims at providing a synopsis of the results of the simulation experiments. The latter constitute the basis for the comparison of the forecasting methods on a theoretical level. To support our key findings, which are derived through a comprehensive examination of the results, here we present a small, though representative, sample of the entire information, while the reader is referred to the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Supplementary material for a thorough overview of the simulation experiments. We especially encourage the reading of Section 3.1 alongside with the report entitled "Selected figures for the qualitative comparison of the forecasting methods" of the Supplementary material.

We choose to start our brief exploration into the simulation experiments by presenting the descriptive statistics calculated for the distributions of the NSE metric values within the SE_1a simulation experiment (see Table 10). This choice is prompted by the fact that the NSE is a metric particularly important for the field of hydrology. In fact, if we had to base our comparisons to one metric, this would probably be this specific one. While examining the numerical results under discussion, we most importantly note that the medians are all negative, which means that at least the 50% of the forecasts given by all the forecasting methods are less close to their corresponding values than the mean of the latter. Secondly, we proceed to ranking the various forecasting methods according to the median (or the mean) values and the condition stated on Table 9. We note that ARIMA_f produces the best forecasts in terms of NSE, as assumed in Section 2.5 for several metrics providing assessment regarding the Type 1 accuracy. We also observe that the simple methods have a rather moderate than bad performance in terms of NSE, since they are better than four forecasting methods, namely ARIMA_s, auto_ARIMA_s, ETS_s and NN_1. Additionally, we note that the sd or iqr values also vary significantly from one forecasting method to the other, with the latter mentioned forecasting methods being worse than the rest also regarding these descriptive statistics. Up to this point, we surely have started to wonder how different the forecasting methods' ranking would be, if we had examined the results provided by some other metric for the same simulation experiment. Furthermore, we wonder in which extent this ranking could differ across the various simulation experiments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Table 10. Descriptive statistics calculated for the distributions of the NSE metric values for all forecasting tests within the SE_1a simulation experiment. The simulation experiments are named according to Table 6.

	min	max	mean	median	sd	iqr	kurtosis	skewness
Naïve	-38.176	0.000	-2.109	-0.896	3.426	2.306	19.912	-3.706
RW	-41.012	0.112	-2.337	-1.004	3.787	2.571	20.082	-3.740
ARIMA_f	-24.000	0.651	-0.901	-0.346	1.746	1.156	42.554	-5.159
ARIMA_s	-52.813	0.823	-3.112	-1.899	4.192	3.212	29.047	-4.123
auto_ARIMA_f	-30.074	0.639	-1.185	-0.421	2.342	1.273	38.271	-5.002
auto_ARIMA_s	-186.344	0.734	-4.343	-2.277	8.392	3.977	156.028	-9.706
BATS	-32.517	0.598	-1.303	-0.476	2.404	1.388	34.249	-4.643
ETS_s	-137.892	0.804	-7.678	-3.642	12.071	7.594	28.102	-4.332
SES	-37.180	0.000	-1.921	-0.791	3.107	2.131	22.453	-3.785
Theta	-37.478	0.055	-1.944	-0.809	3.143	2.160	22.116	-3.773
NN_1	-85.677	0.697	-3.132	-1.660	5.417	3.268	67.804	-6.464
NN_2	-822.800	0.655	-1.919	-0.612	18.602	1.521	1893.157	-42.990
NN_3	-578.538	0.773	-1.422	-0.431	13.091	1.271	1885.438	-42.856
RF_1	-39.846	0.790	-1.344	-0.532	2.728	1.419	64.811	-6.464
RF_2	-39.379	0.769	-1.401	-0.594	2.699	1.483	51.352	-5.742
RF_3	-38.670	0.713	-1.602	-0.772	2.518	1.815	37.973	-4.495
SVM_1	-39.495	0.810	-1.396	-0.614	2.763	1.470	65.887	-6.537
SVM_2	-37.764	0.715	-1.327	-0.571	2.370	1.480	44.327	-4.967
SVM_3	-31.314	0.765	-1.094	-0.409	1.970	1.248	40.159	-4.648

Before answering the above worded questions, we consider important to report some preliminary observations extracted from Table 10, but also verified for the rest of the numerical results. Regarding the kurtosis and skewness values, these indicate that the measured NSE metric values are not normally distributed. This specific observation introduces several issues in handling the metric values for the purposes of the present study, which we overcome by basing our comparisons mainly on the medians and the iqr values, as well as by removing the (far) outliers from the figures. Noteworthy is also the fact that the above worded observation does not apply to all of the metrics equally and, thus, we could claim that there are metrics more manageable than others. In fact, although we focus on the comparison of the forecasting methods, the interested reader could discover worth considered information about the metrics in the Supplementary material. Another observation regards the NN forecasting methods, which stand out because of their far outliers. The latter is a clear sign of instability, while the ARIMA_s, auto_ARIMA_s, ETS_s and NN_1 forecasting methods seem to share a different form of instability. Finally, we note the great similarity that the distributions measured for the SES and Theta forecasting methods exhibit, which applies to all the distributions measured for these specific forecasting methods within the present study, confirming the findings of previous studies, e.g. Hyndman and Billah (2003). This similarity is, however, of secondary importance here.

Next, we choose to navigate within the SE_1a simulation experiment using a simplified form of the distributions of the metric values (see Figures 8-10). This specific navigation is indicative of the navigation within any other simulation experiment, and, thus, of great importance in yielding valuable insights into the nature of hydrological time series forecasting, as well as in comparing the forecasting methods or the information provided by each metric regarding the latter.

Some important observations applying to all the simulation experiments are reported subsequently. First, we observe that even the relative metrics, i.e. the corresponding to the same criterion (see Table 5), provide measurements which lead us to different aspects of the same information to an extent bigger or smaller depending on the pair of metrics compared. Second, we note that some of these 18 different aspects are also conflicting to each other and, hence, we realize that we cannot decide on a general ranking of the forecasting methods using the results of a specific simulation experiment, unless we introduce extra constraints. In more detail, we would have to select the criterion and, furthermore, the metric of our interest, for this specific ranking to be possible. Third, we perceive that the collective examination of the quantitative form of the results is quite challenging and, thus, the imposition of a simplification procedure is rather required.

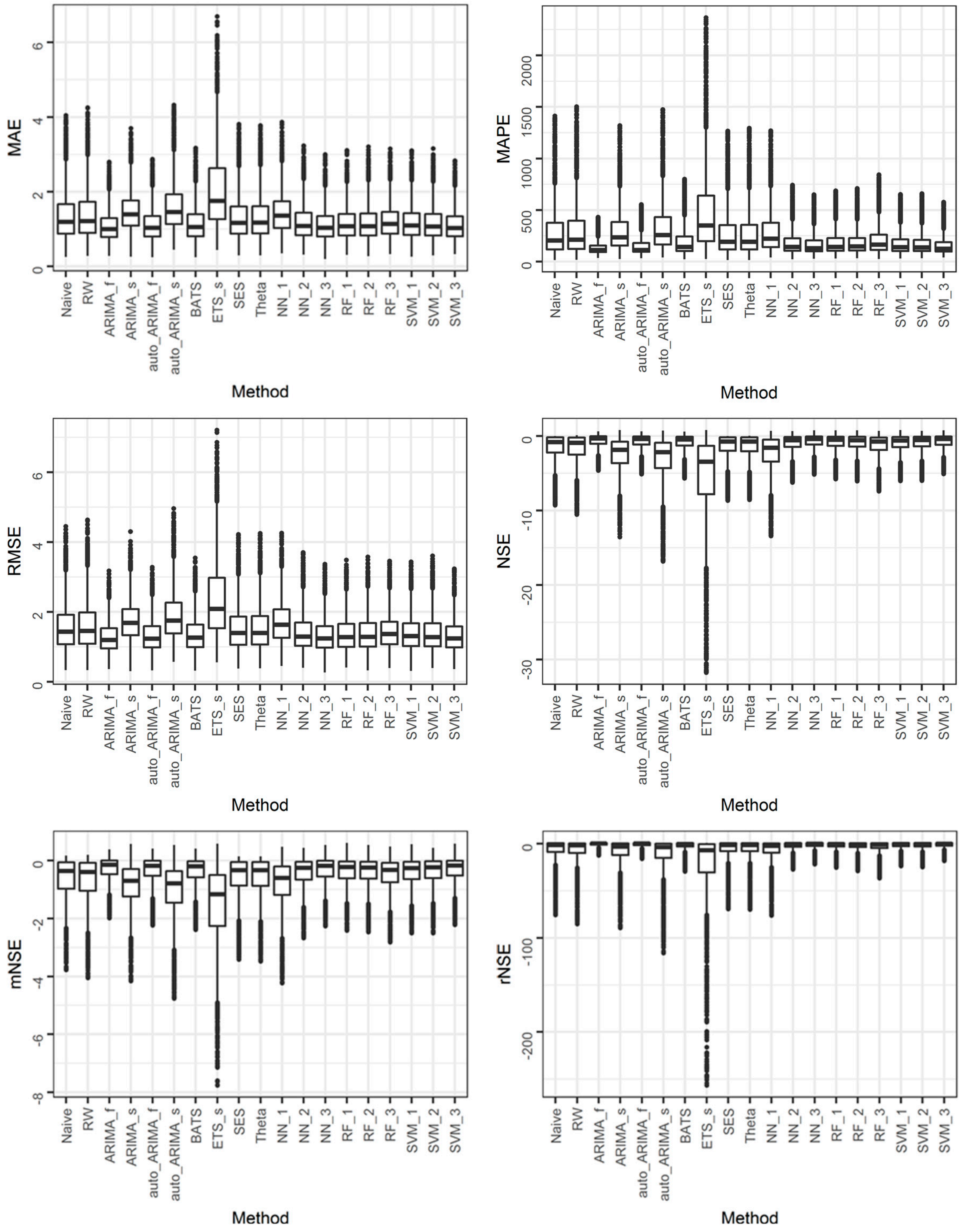

Figure 8. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance within the SE_1a simulation experiment (part 1). The far outliers have been removed.

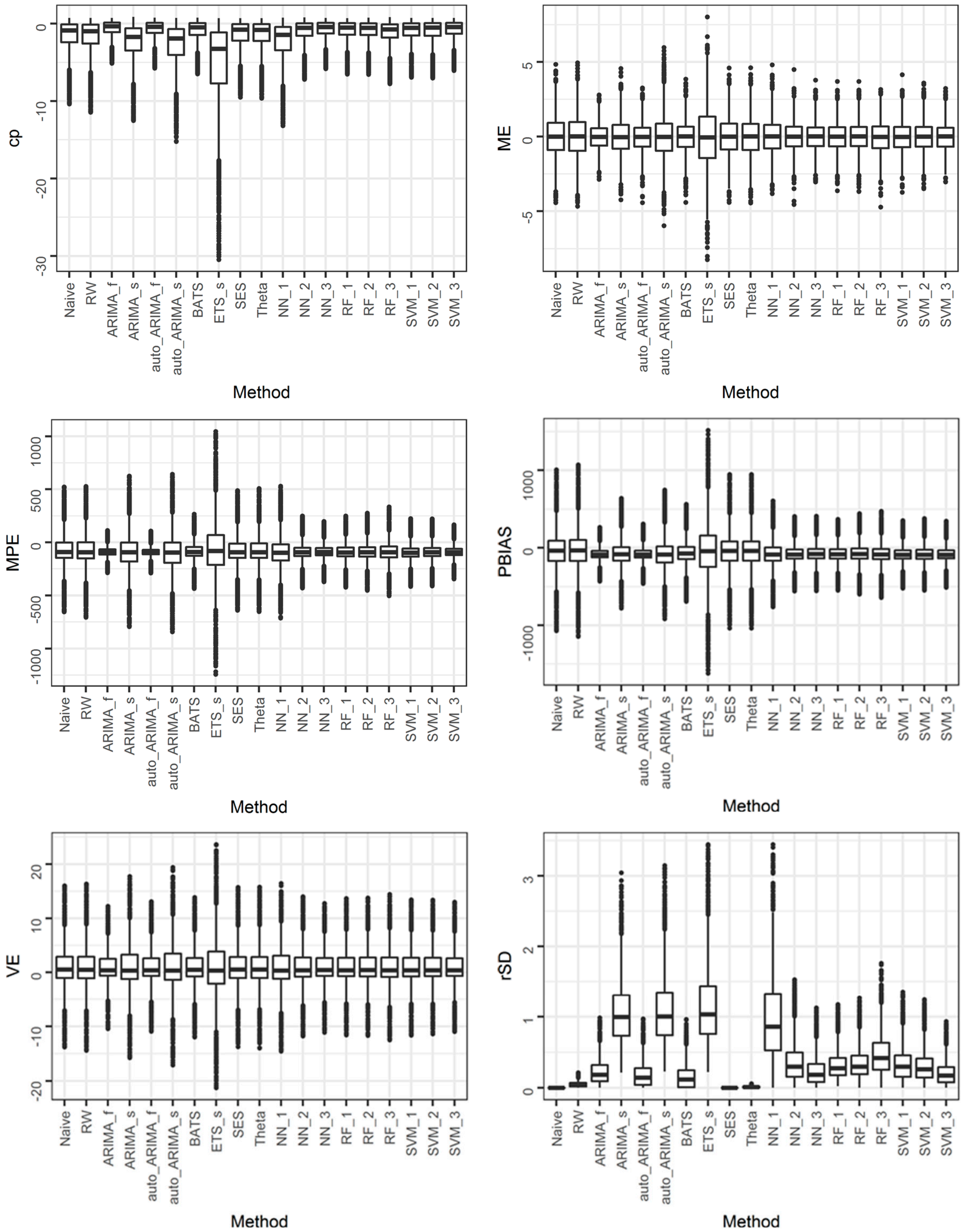

Figure 9. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance within the SE_1a simulation experiment (part 2). The far outliers have been removed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

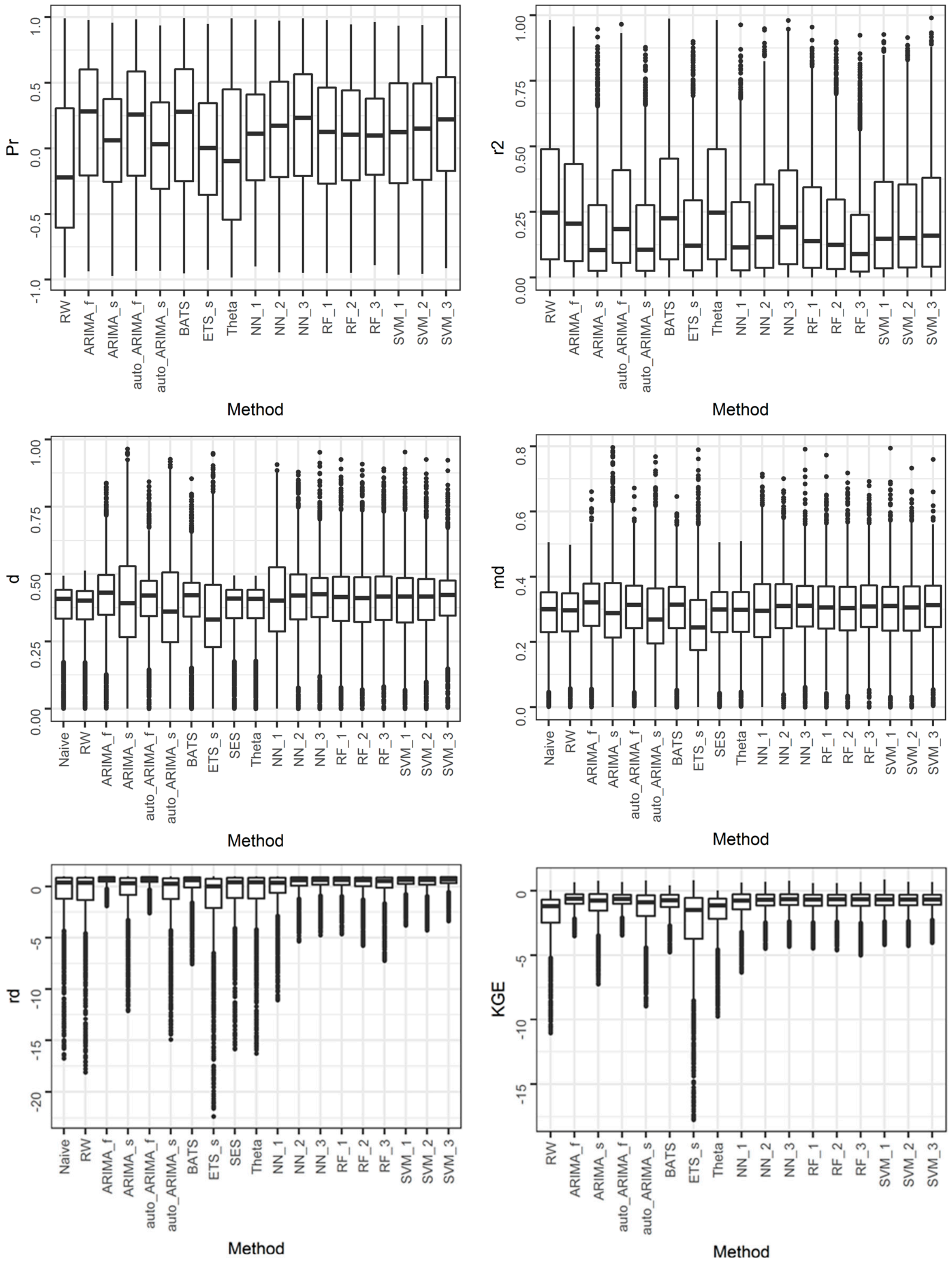

Figure 10. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance within the SE_1a simulation experiment (part 3). Concerning the boxplots of the metrics rd and KGE, the far outliers have been removed.

To this end, in Figure 11 we present the Type 1 heatmaps for the comparison of the forecasting methods within the SE_1a, SE_1b, SE_2a and SE_2b simulation experiments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Additionally, in Figures 12-17, we present the Type 2 heatmaps according to the RMSE, PBIAS, rSD, Pr, d, KGE metrics respectively. In the heatmaps the scaling is performed in the row direction and the darker the colour the better the forecasts. We have also applied clustering analysis on the forecasting methods based on their performance. Having moved step-by-step from the level of a specific metric to the collective examination of the information provided by all the 18 metrics within a specific simulation experiment and subsequently to the collective examination of the entire information using the Type 1 and Type 2 heatmaps, we devote the following paragraphs in summarizing the outcome of the simulation experiments. Figures 11-17 together with Figures 8-10 can facilitate the reading of this synopsis in a rather satisfactory manner. Regarding the decoding of the information provided by the conducted heatmaps, an example is available in Section 3.2.

Admittedly, the most significant outcome of the present study is that none of the forecasting methods is found to be better or worse than the rest regarding all the metrics employed in the evaluation process simultaneously. Alternatively worded, none of the forecasting methods is uniformly best or worst. This observation, first obtained for the SE_1a simulation experiment through the collective examination of Figures 8-10, is easily confirmed for the rest of the simulation experiments using the Type 1 heatmaps, while it reveals that the forecasting quality is subject to limitations. However, there are forecasting methods regularly better or worse than others according to specific metrics, as well as forecasting methods sharing a quite similar performance (e.g. Naïve and RW). The latter easily becomes evident through the examination of the Type 2 heatmaps. For instance, the ARIMA_s, auto_ARIMA_s and ETS_s forecasting methods exhibit far the best average-case performance in terms of rSD for all the simulation experiments of this study (see Figure 14). This kind of information can be useful for the selection of the optimal forecasting method/methods for various engineering applications and also allows the wording of several advantages/disadvantages to some extent, as well as a preliminary clustering of the forecasting methods. Of course, this fact does not apply to all the forecasting methods neither to all the metrics. For example, we observe that the Theta forecasting method can exhibit good, moderate or bad average-case performance in terms of KGE depending on the simulation experiment (see Figure 17), while none of the forecasting methods is found to be regularly better or worse in respect to PBIAS (see Figure 13).

In summary, each forecasting method has some specific theoretical properties and, due to the latter, it performs better or worse than others with respect to specific metrics and/or within specific simulation experiments. We note that the former observations apply equally to the stochastic and the ML forecasting methods. Therefore, forecasting methods originating from both the main categories are found amongst the first, as well as the last positions in each resulting ranking. The latter is possible for a specific metric within a specific simulation experiment. Furthermore, it is noteworthy that Naïve and RW are also competent. These simple methods exhibit rather the best average-case performance in terms of d (see Figure 16) and md. Naïve and RW also perform better or equally well to other forecasting methods regarding several metrics within specific forecasting experiments. We also note that the values of the metrics can vary significantly across the different time series, while in general this variation depends on the metric and the forecasting method and can be extracted from the quantitative form of the results (see for example Figures 8-10). The observations outlined above hold a complete explanation of the results derived by Papacharalampous et al. (2017).

Another important outcome resulting from the examination of the Type 1 heatmaps is that the length of the time series mostly affects the comparative quality of the forecasts and the clustering of the forecasting methods to a smaller extent than the simulated process. In more detail, the length of the time series affects the performance of the NN_1 forecasting method largely, while the performance of the rest forecasting methods is less or even slightly affected depending on the forecasting method as also on the simulated process. Moreover, we observe that forecasting methods resulting from the implementation of the same algorithm can exhibit a far distant or always close performance depending on the algorithm. For instance, the ARIMA forecasting methods can differ with each other to a great extent, a fact also applying to NN, but not to the RF and SVM forecasting methods. Additionally, we note that some of the Type 2 heatmaps are rather smoother than other heatmaps of the same category (see for example Figures 12-17), as the variability across the different simulation experiments can be larger or smaller depending on the metric.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

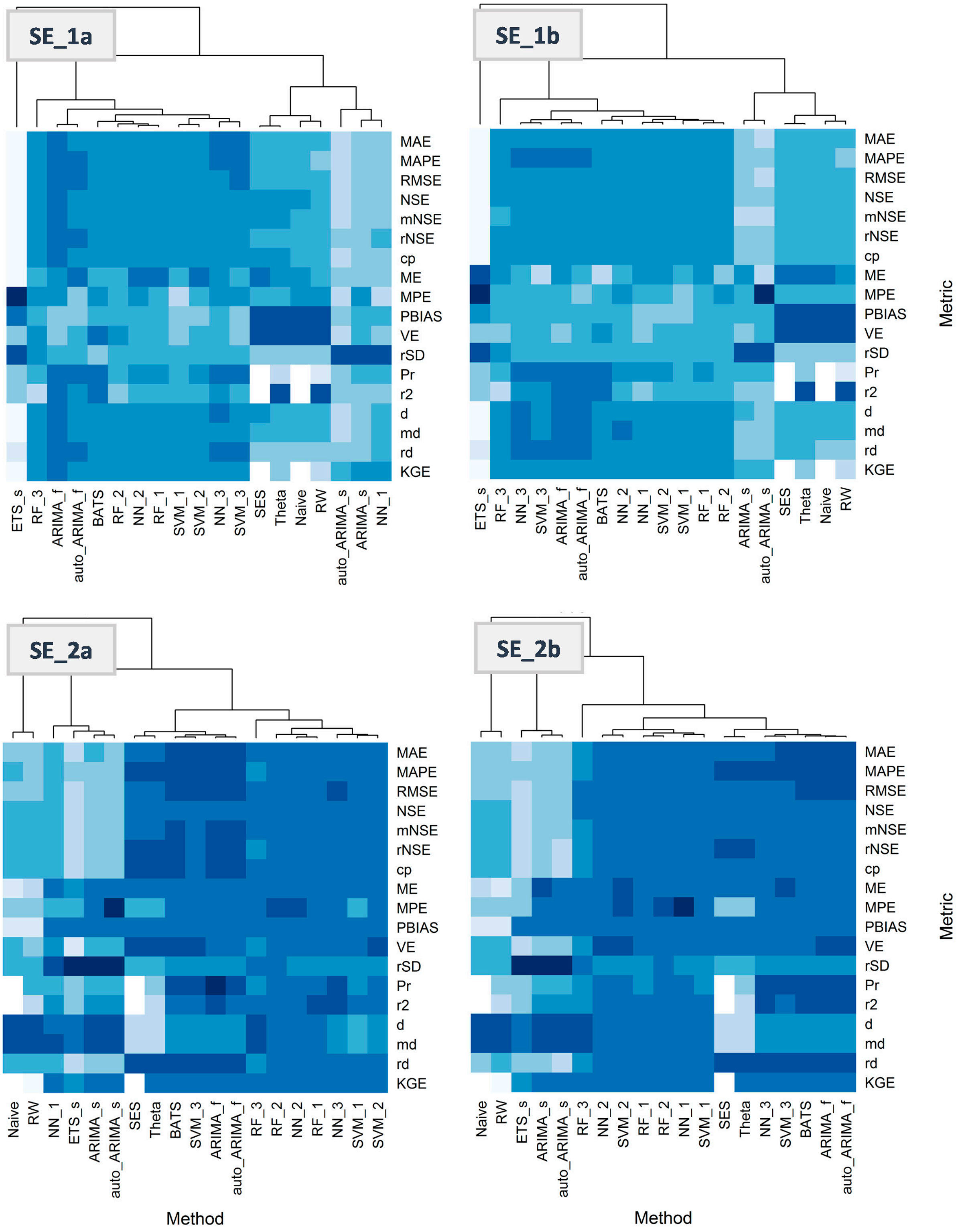

Figure 11. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the metrics and the conditions listed on Table 9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

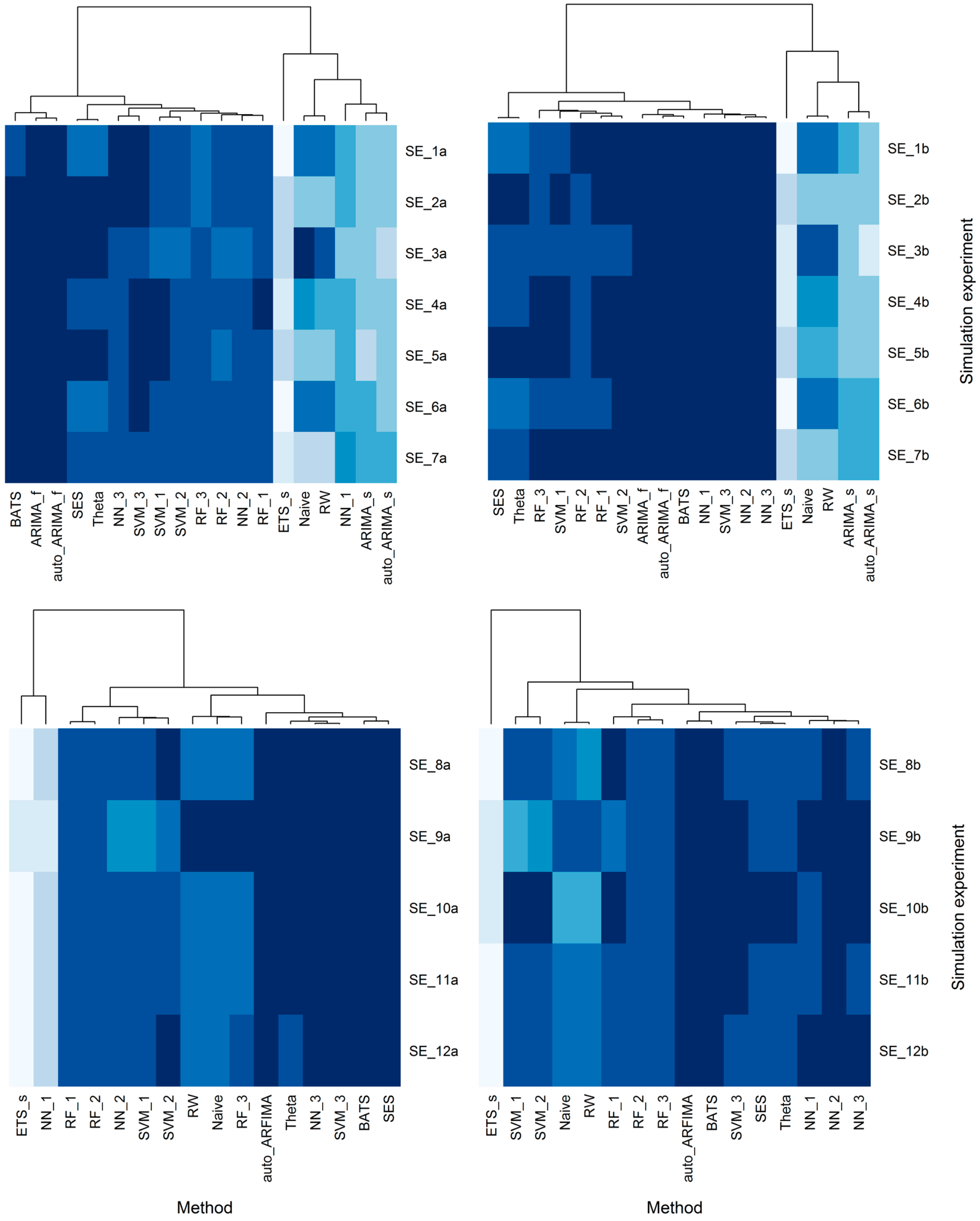

Figure 12. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the RMSE metric and the condition stated on Table 9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

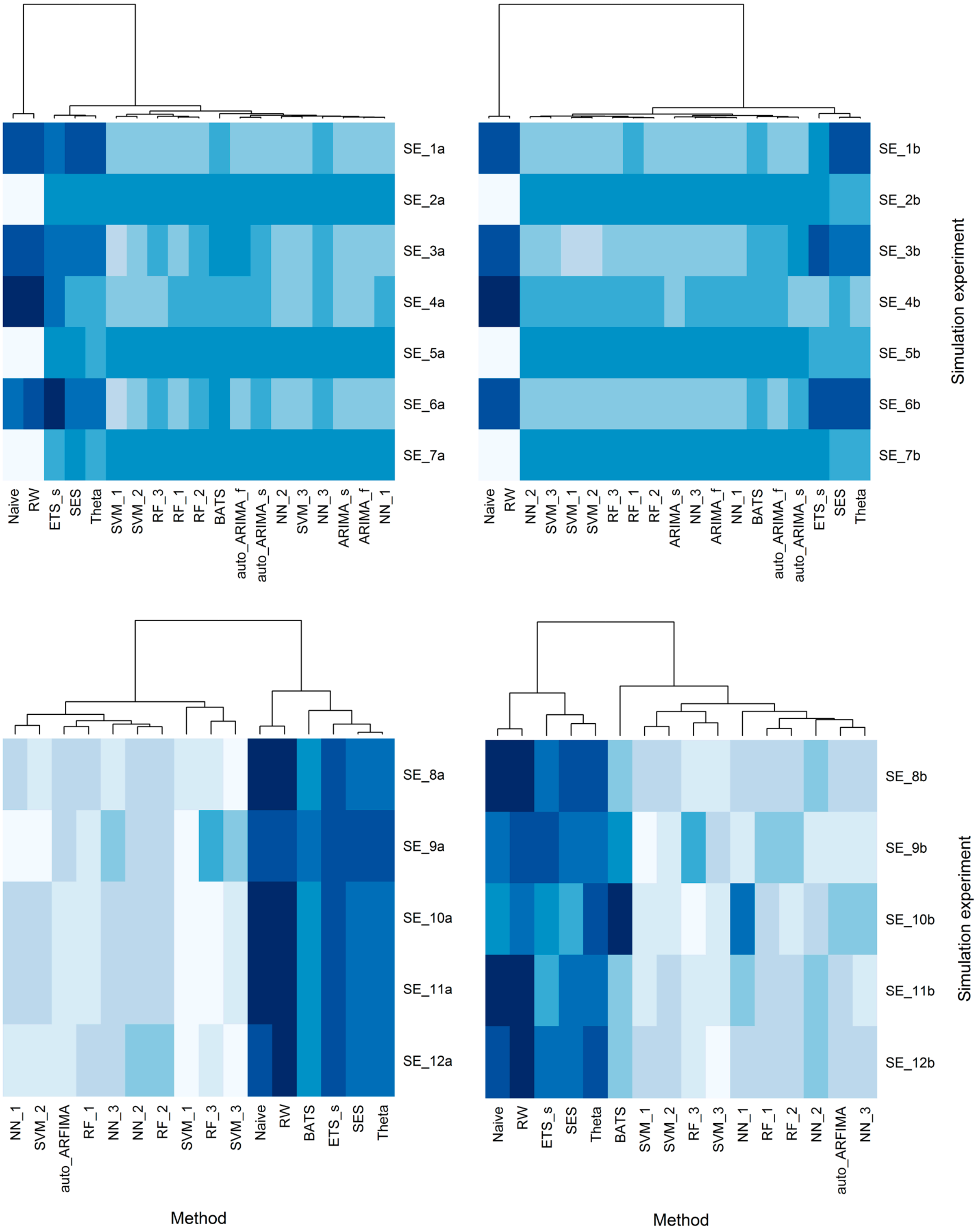

Figure 13. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the PBIAS metric and the condition stated on Table 9.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

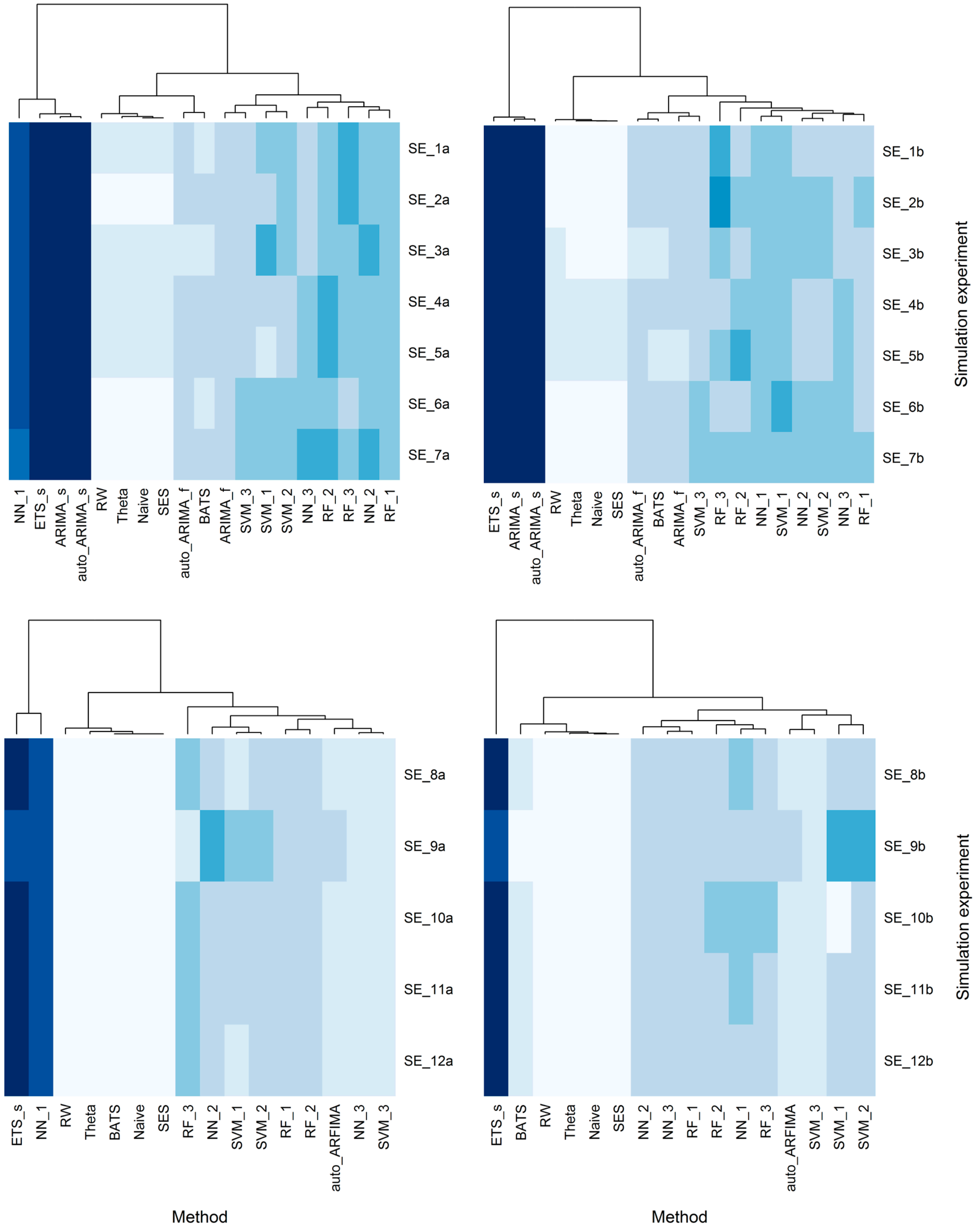

Figure 14. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the rSD metric and the condition stated on Table 9.

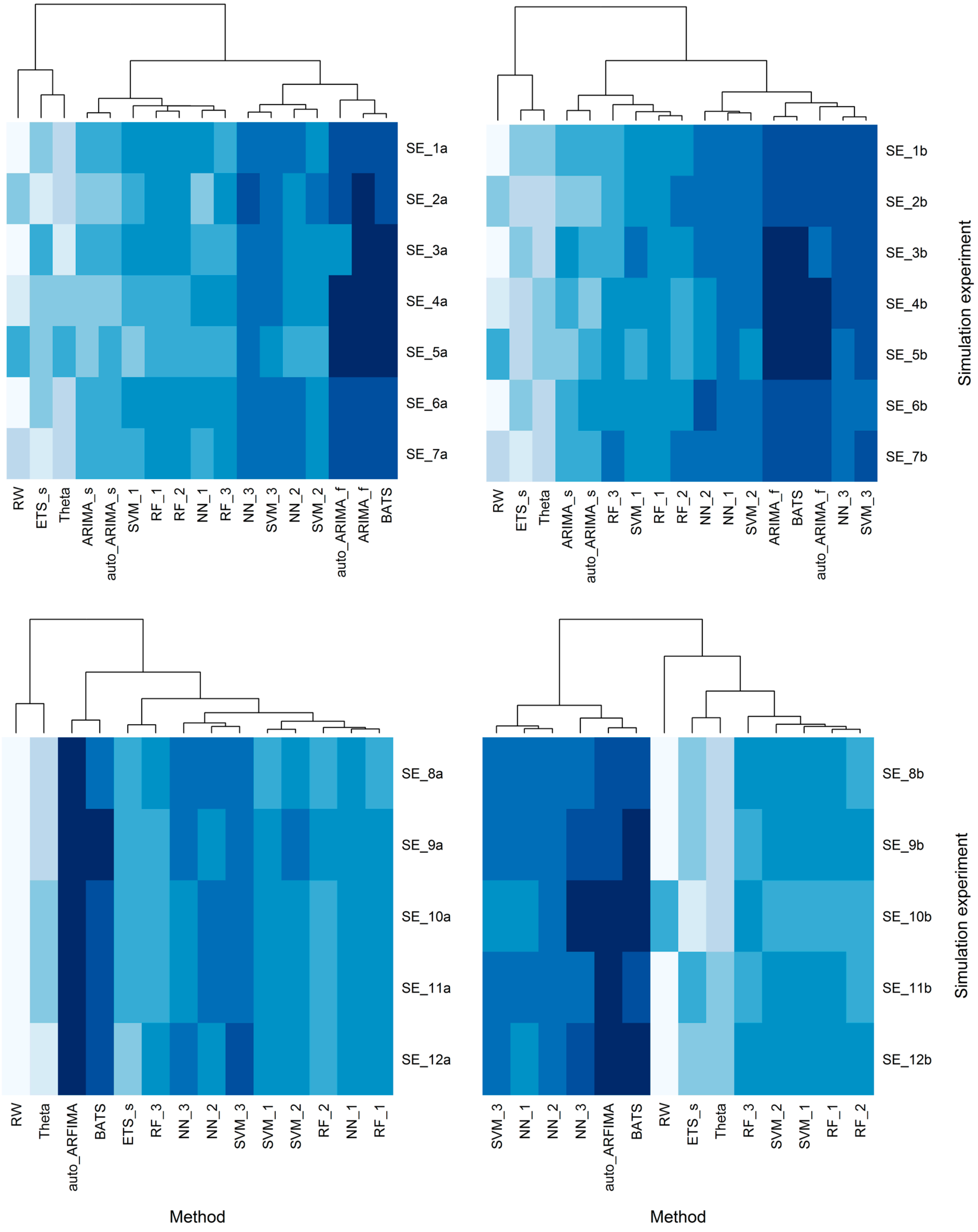

Figure 15. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the Pr metric and the condition stated on Table 9. We note that the Pr metric cannot be calculated for the Naïve and SES forecasting methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

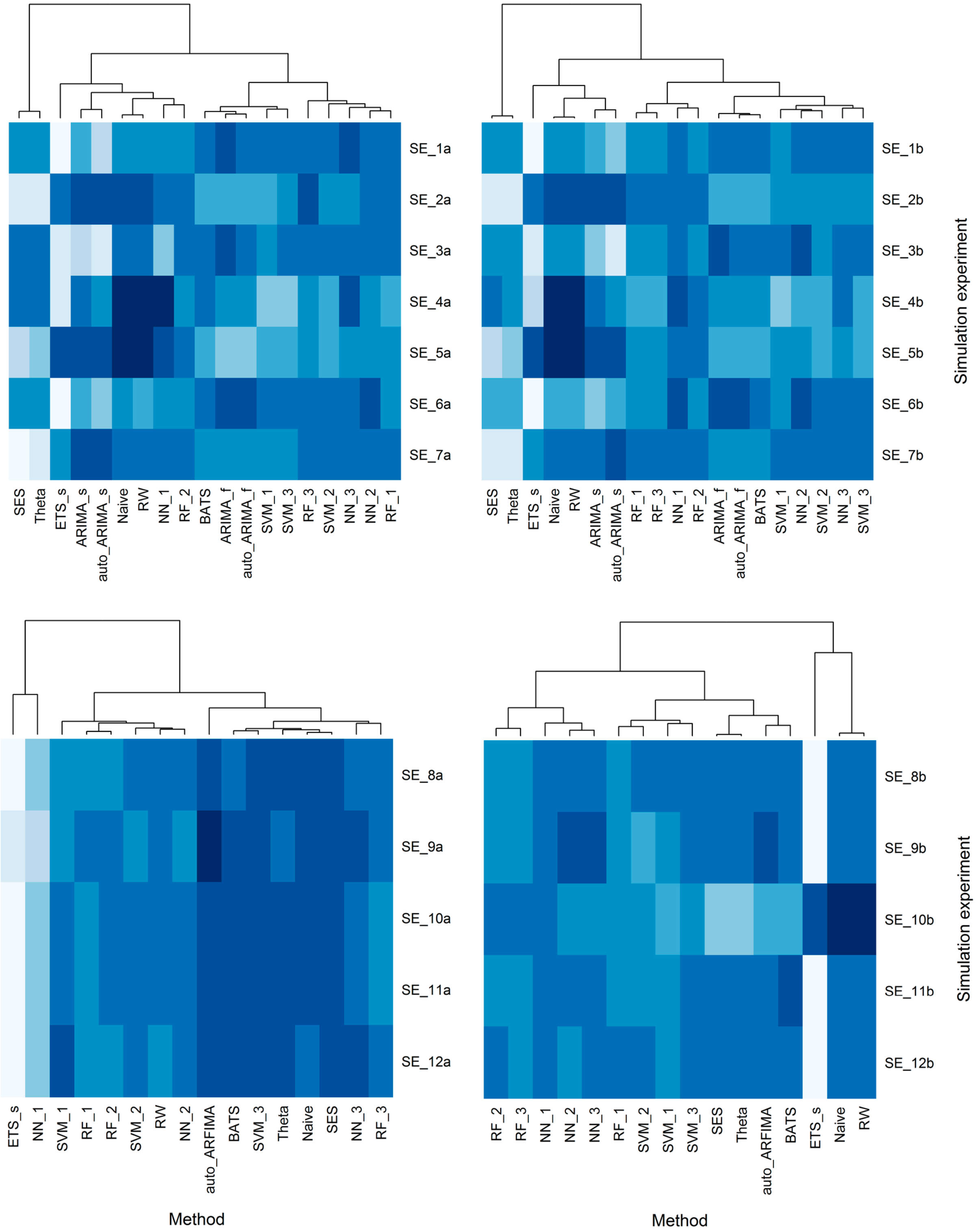

Figure 16. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the d metric and the condition stated on Table 9.

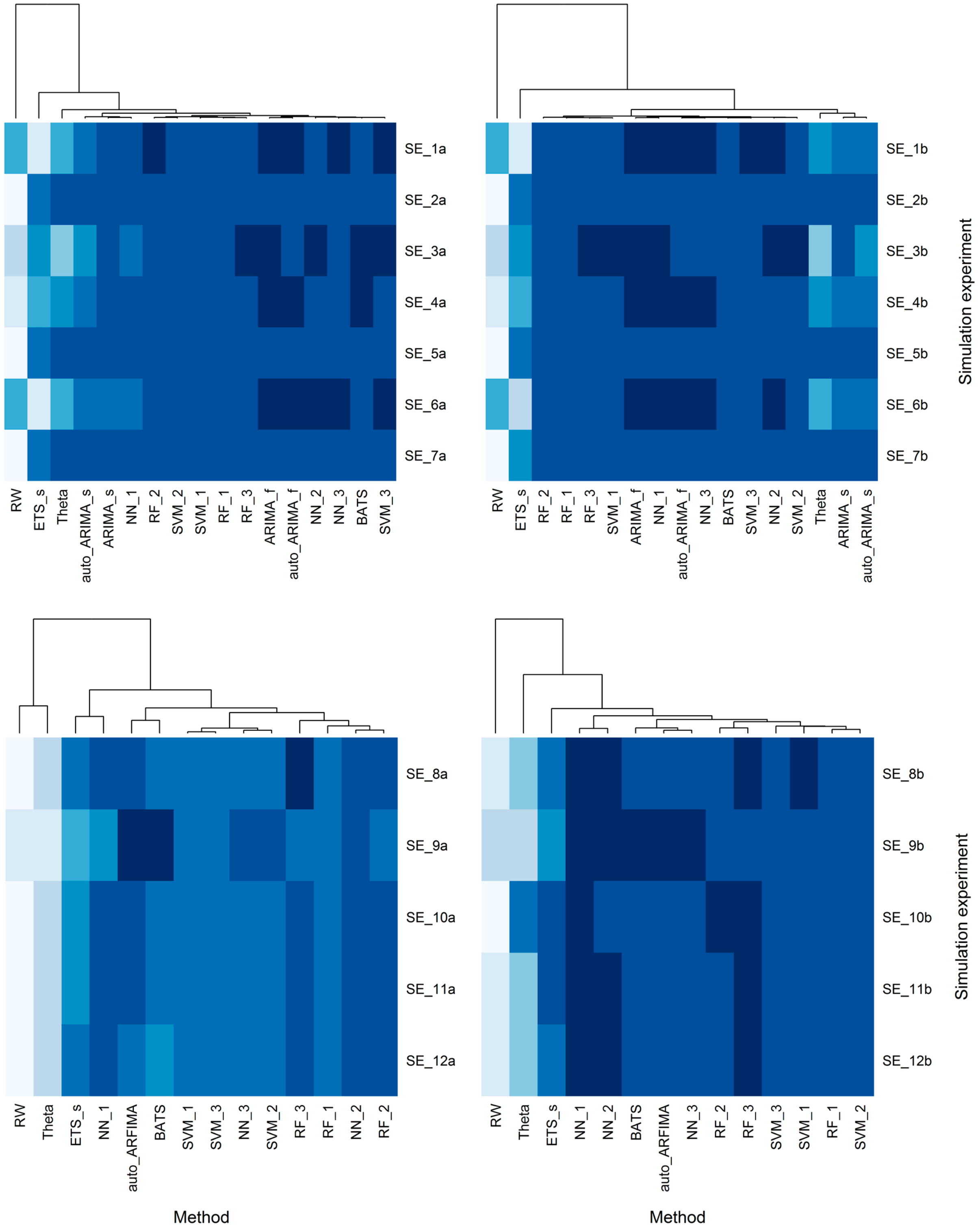

Figure 17. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the KGE metric and the condition stated on Table 9. We note that the KGE metric cannot be calculated for the Naïve and SES forecasting methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

3.2 The CS_77 case study (and variations)

In full correspondence to the simulation experiments, the results of the CS_77 case study are presented in both quantitative and qualitative forms in Figures 18-20 and Figure 21 respectively. Some information extracted from the numerical results regards the improvement of the forecasting outcome, when the forecasts are produced using the deseasonalized time series and subsequently recovering the seasonality, a fact apparent even in Figure 6. The satisfactory values measured for the NSE, mNSE, rNSE, Pr, r2, d, md and rd metrics, in comparison to their corresponding distributions resulted from the simulation experiments (see for example Figures 8-10) are indicative of this specific improvement. For the qualitative comparison of the forecasting methods' performance within the case study under discussion, we can either examine collectively the presented barplots, or alternatively, mine the information provided by the heatmap, for example as presented subsequently. While trying to decode Figure 21, we first note that the performance of SVM_1, SVM_2 and SVM_3 is similar and largely different from the performance of the rest forecasting methods. Such similarities and differences are clearly the reason for the clustering accompanying the qualitative representation of the numerical results. This clustering is illustrated with a proper column rearrangement as also a cluster dendrogram and has proven its usefulness in the decoding processes.

Furthermore, we note that the forecasts resulting from the SVM algorithm are amongst the worst regarding the information provided by several metrics, but rather satisfying in respect to other metrics and the best in respect to VE. The same applies to the forecast resulting from the ETS_s forecasting method, although the quite good performance in its case is measured in terms of ME and VE. This specific forecasting method is clustered together with auto_ARFIMA, RF_2, BATS, NN_1 and NN_2, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

exhibit a rather moderate overall performance. Amongst the latter forecasting methods, RF_2 produces the best forecasts within our case study. Another formed cluster is composed of the NN_3 and RF_1 forecasting methods. Those two ML methods share, in fact, a quite similar performance within this real-world case study, although the forecast resulting from the implementation of RF_1 is clearly better, as well as the best amongst all the forecasting methods in terms of rSD and KGE. Finally, the cluster exhibiting the best overall performance includes Naïve, RW, SES and Theta, as also RF_3. The forecasts resulting from the former four forecasting methods are rather equivalent and better than the forecast produced by the latter.

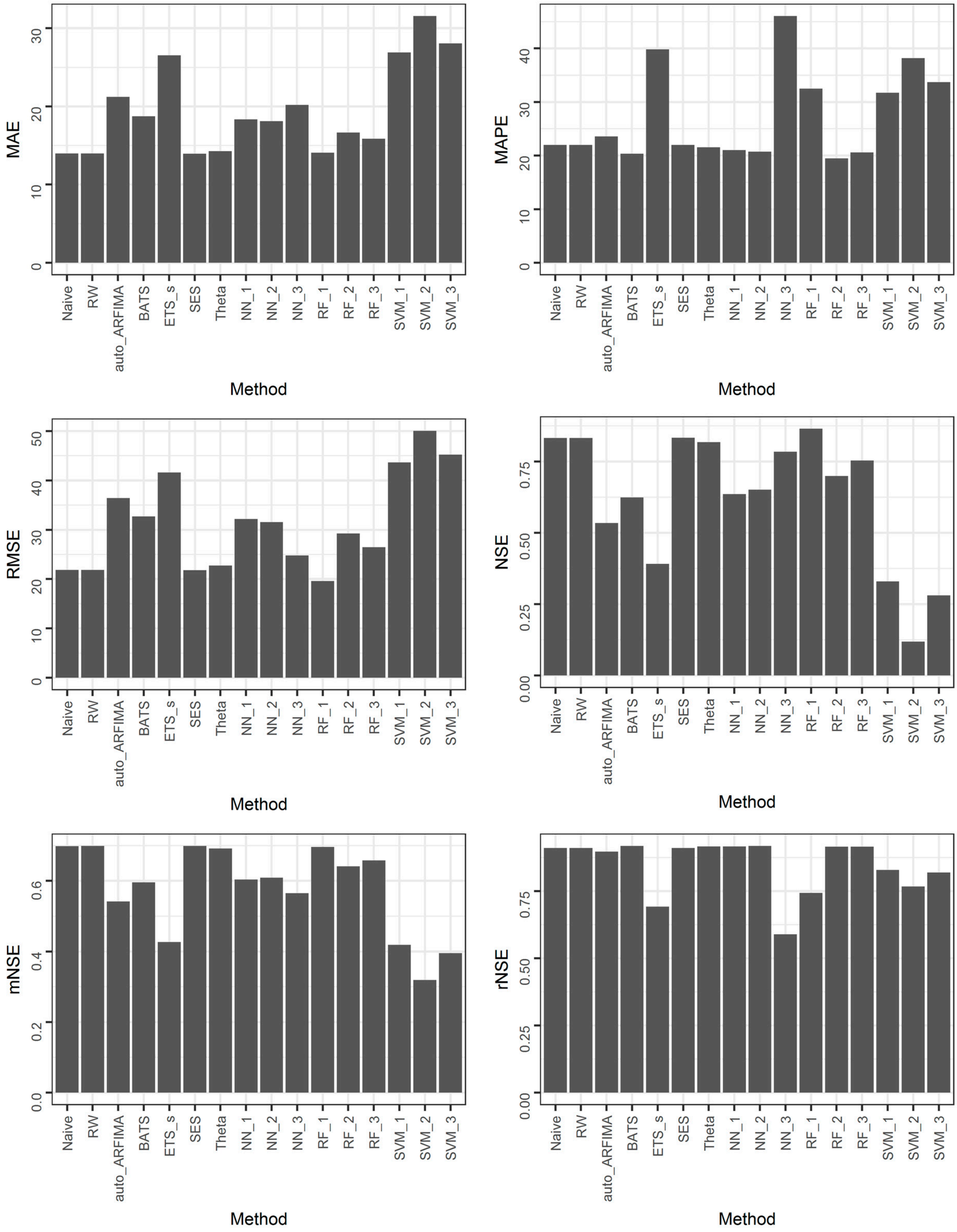

Figure 18. Barplots for the comparative assessment of the forecasting methods regarding their performance within the CS_77 case study (part 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

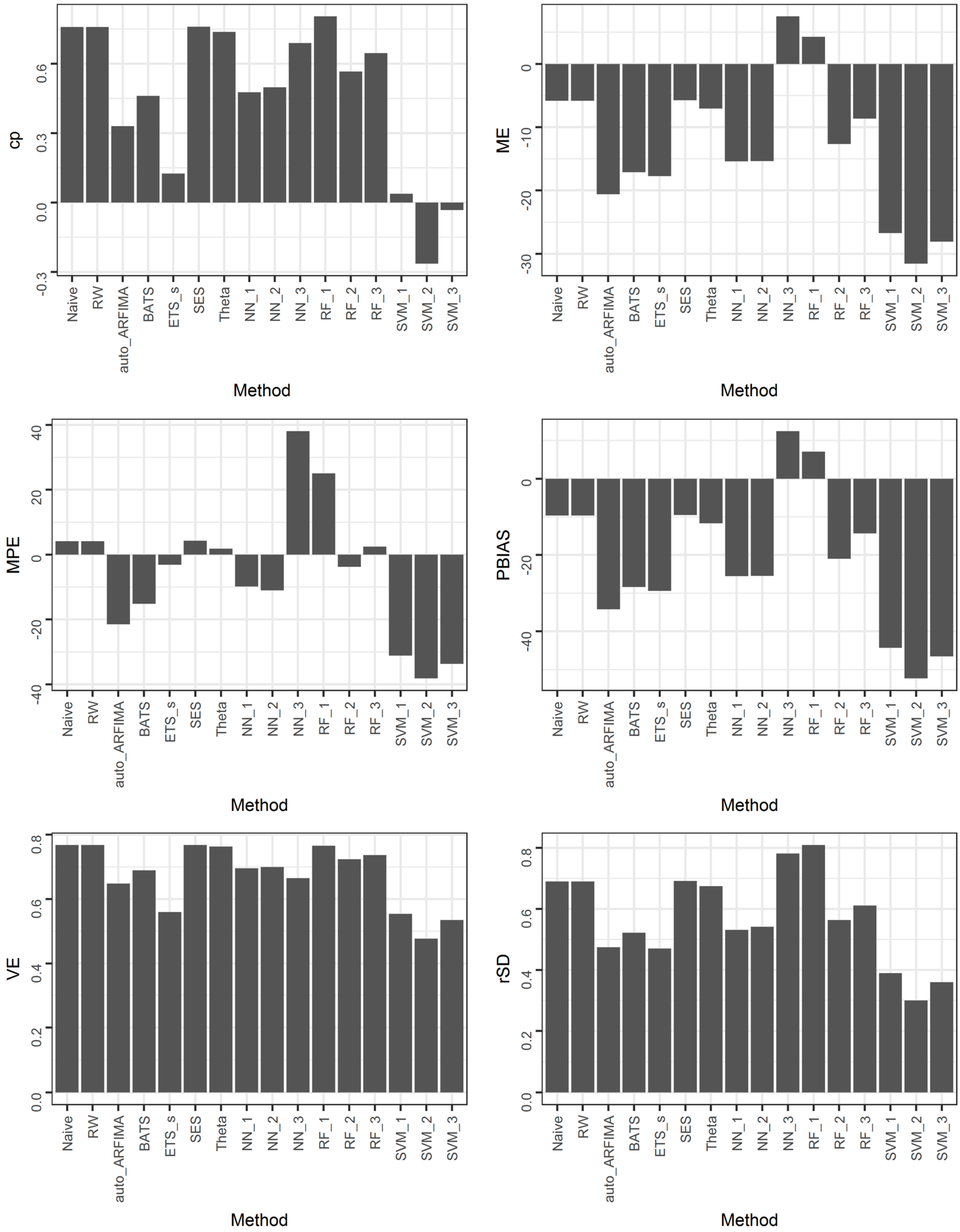

Figure 19. Barplots for the comparative assessment of the forecasting methods regarding their performance within the CS_77 case study (part 2).

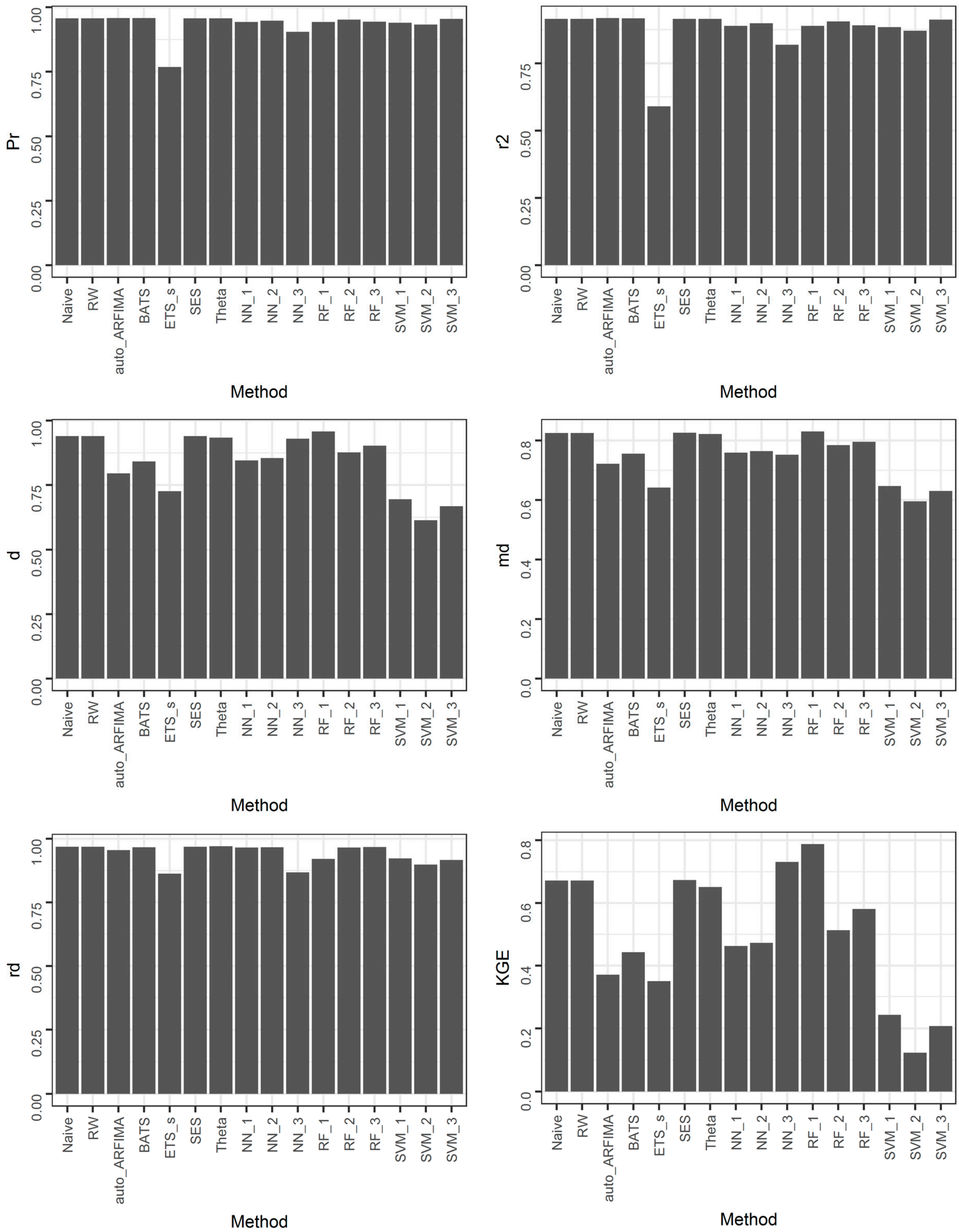

Figure 20. Barplots for the comparative assessment of the forecasting methods regarding their performance within the CS_77 case study (part 3).

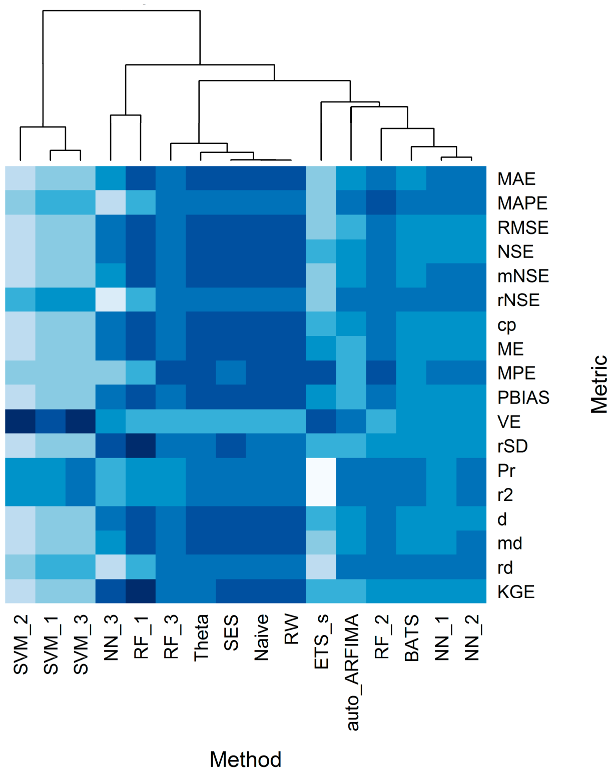

Figure 21. Heatmap for the comparative assessment of the forecasting methods within the CS_77 case study according to the values of the metrics and the conditions listed on Table 9.

The CS_77 case study validates the core findings of the present study, while focusing on a particularly interesting individual case that would have stayed hidden otherwise. As in the simulation experiments, here again none of the forecasts is found to be better or worse than the rest in respect to all the metrics simultaneously and, consequently, we cannot decide on a uniformly best or worst forecasting method, not even for the single case under discussion. Additionally, it seems that the forecasts resulting from both the main categories of forecasting methods are rather equally competent and subject to limitations. Furthermore, any ranking of the forecasting methods would require the prior selection of a metric of interest, while the clustering presented in Figure 21 would be totally meaningless beyond this specific case study. Moreover, away from the expected results, Naïve and RW are despite their indisputable simplicity amongst the best forecasting methods for this case. This is a quite remarkable outcome.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Finally, in Figure 22 we present the results of the examined variations of CS_77. Clearly, the relative performance of the forecasting methods is completely different across the five related experiments, a fact rather indicating that the best forecasting method might depend largely on the forecasting case itself and, thus, might vary, even when trying to predict in different time moments a specific process at a specific location, given a slightly different amount of observations at each experiment. This is another worth-considered outcome.

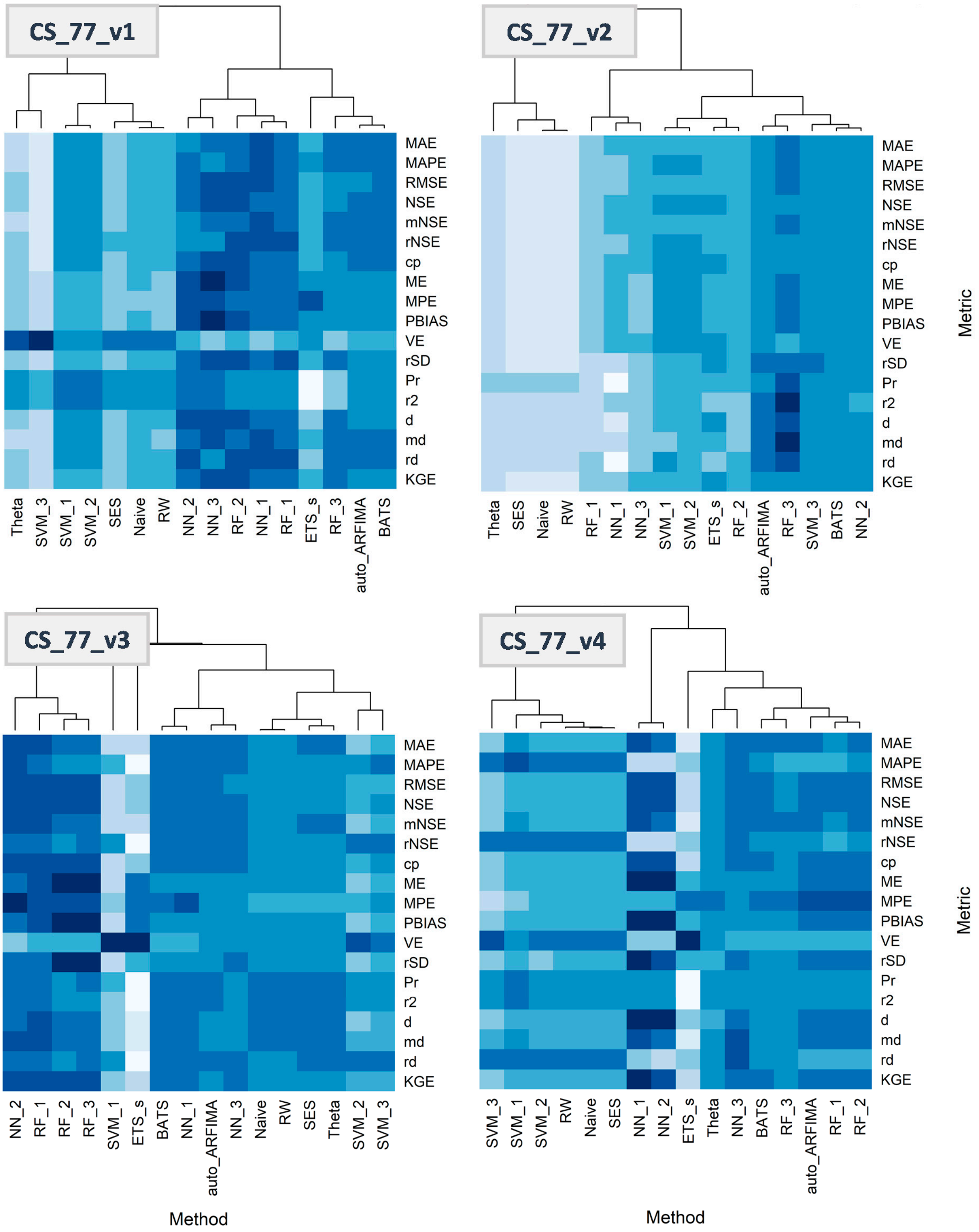

Figure 22. Heatmaps for the comparative assessment of the forecasting methods within the variations of the CS_77 case study according to the values of the metrics and the conditions listed on Table 9.

3.3 A collection of 92 real-world case studies

Being familiar with the results presented in Sections 3.1 and 3.2, the reader will not be surprised by the results of the collection of 92 diverging real-world case studies.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

However, since such collections are rare in the literature, we encourage the reading of the report entitled "Qualitative comparison of the forecasting methods in 92 real-world case studies" of the Supplementary material. Although we specifically focus on CS_77 (and its variations) herein, all of the real-world case studies conducted could be used to confirm the findings of Section 3.1 in an equally satisfactory manner, while the entire collection highlights the individuality of each case. The same applies to the multiple-case study conducted by Papacharalampous et al. (2017).

4. Discussion

4.1 Contribution in hydrology and beyond

The present study aims to contribute in hydrology (and beyond) in two direct ways and indirectly. Our generalized findings, derived in Section 3.1 and validated in Sections 3.2 and 3.3, can provide new insights into the nature of hydrological time series forecasting. This first direct contribution can be summarized with the following context-independent research questions and answers, which might be of scientific and/or practical interest:

- *Do the ML methods exhibit different forecasting performance from the stochastic? No, stochastic and ML methods can share a quite similar forecasting performance.*
- *Is it possible for two forecasting methods resulting from the implementation of the same algorithm to exhibit a far distant performance? Yes, it is possible for stochastic and ML algorithms.*
- *To which extent might the performance of a specific forecasting method differ across the various cases of time series? To an extent smaller or larger depending on various factors, such as the forecasting method, the criterion of interest and the process under investigation.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

- *Do sophisticated forecasting methods necessarily provide better forecasts than simple forecasting methods? Not necessarily. It depends on various factors, such as the compared forecasting methods, the criterion of interest, the process under investigation and the forecasting case itself.*
- *Are there forecasting methods standing out because of their good or bad performance? Yes, a forecasting method can be regularly better/worse than others with respect to specific metrics. However, this does not apply to all the forecasting methods.*
- *Is it possible to name several advantages/disadvantages of the forecasting methods? Yes, these advantages/disadvantages of the forecasting methods are related to their good/moderate/bad performance with respect to specific metrics.*
- *Furthermore, is the classification of the forecasting methods possible? Yes, but only to some extent. This classification could be based on the similar or contrasting performance of the forecasting methods with respect to the various metrics.*
- *Moreover, is a general ranking of the forecasting methods possible? No, any ranking of the forecasting methods would require the a priori selection of a stochastic process and a criterion of our interest, as well as the application of a simplification procedure (e.g. use of the median values of the selected metric) and, thus, would not be general.*
- *Finally, can we decide on a universally best forecasting method? Alternatively, can we decide on a best forecasting method regarding all the criteria set? No, none of the forecasting methods is better or worse than the rest with respect to all the metrics simultaneously.*

Admittedly, the qualitative form of the results of the simulation experiments facilitates their handy examination and, thus, eases the delivery of the generalized findings. The latter are entirely consistent with the evidence provided by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Papacharalampous et al. (2017). Furthermore, the limitations accompanying time series forecasting emphasized by Koutsoyiannis et al. (2008) and Papacharalampous et al. (2017) are highlighted here as well. In addition to this contribution, the second direct contribution is the context-independent quantitative information about the performance of 20 popular forecasting methods on linear processes according to 18 metrics. The forecasting methods are easily implemented, since they are available in code form in the Supplementary material, with the prospect of being of practical value in future applications.

The following procedure for the selection of the most appropriate forecasting method for a particular case study can be suggested, using the results of this study. First, we fit the 12 different models listed in Table 1 to the time series of our interest. Second, we use one information criterion of our choice (e.g. the AIC) or more, to select an optimal model for our data, which together with the length of our time series will nominate the closest simulation experiment for our case. Next, we decide on the metrics that will determine our final selection according to the criterion of interest. Last, we choose the forecasting method to be used depending on its expected performance regarding this specific criterion. It would be interesting to also run the code of the case study on our data and compare its results to the expected. The indirect contribution of this paper is the material for the evaluation of the metrics' utility.

One could claim that there may be an undiscovered ML technique, which will be better than the existing ones. As commented in Hong and Fan (2016) the number of original techniques is countable and exhausted, therefore researchers combine them (the so-called hybrid techniques) to introduce "new" techniques. Most of them (and the accompanying papers) are useless, however researchers test them in manipulated datasets to ensure publication and introduce "superior alternatives", "powerful tools"

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

and similarly described methods. Regarding the "myth of the best method" and more details on the invention of new techniques the interested reader is referred to Hong and Fao (2016). The present study aims at developing a detailed framework for assessing such techniques.

Another important contribution of the present study is related to the so-called "no free lunch theorem" (Wolpert 1996). According to Wolpert (1996), in the space of all possible problem instances, there is not a model, which will always perform better than the other models, in the absence of significant information for the problem at hand. The present empirical study shows that even in the finite space of simple (simulated time series) and real-world case studies problems examined here there is not an optimal solution for time series forecasting. Finding the best algorithm mostly depends on our knowledge of the system. For example, using ARFIMA models for forecasting the ARFIMA simulated time series is obviously the best choice, due to the prior known information about the system. The other methods are equivalent in performance since they cannot incorporate this knowledge. In the specific class of geophysical processes forecasting finding information about the examined system could be possible, e.g. with the application of principles of physics, such as the maximum entropy principle. Obviously, the knowledge of the system is not simply equivalent to the knowledge of its statistical properties, e.g. the mean, variance, the autocorrelation function etc., but should be deeper. Therefore, the frequently met in the literature blind use of forecasting methods is not suggested. Furthermore, it seems that major advancements in time series forecasting performance of ML methods can be achieved by incorporating appropriate exogenous variables in the model, while the potential for improving their performance in univariate time series forecasting seems limited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Regarding the extent to which the conclusions are generalizable, we note that the stationarity assumption and the reasoning concerning its appropriateness in the modelling of geophysical properties in Koutsoyiannis and Montanari (2015) is consistent with the no free lunch theorem. Therefore claiming that the results are not generalizable because of the restriction to the case of stationary stochastic processes is unfounded. In particular, if we cannot explain the behaviour of a geophysical process based on a deterministic mechanism, then the most appropriate models are stationary. Even in cases of deterministic systems, stochastic approaches are appropriate (Koutsoyiannis 2010). This is a frequently met case in modelling of geophysical processes (i.e. there is not an adequate explanation for the behaviour of the geophysical process), proving that our results are generalizable.

4.2 On the methodological approach

The above section highlights the efficiency of our methodological approach in producing generalized results. Moreover, the real-world case studies emphasize on important issues, which exhibit greater interest when presented on real-world data, while they also reinforce the findings of the simulation experiments. Someone who examines both the results of the simulation experiments and the case studies has a more complete picture of the underlying phenomena than whom considering only the results of the simulation experiments. On the other hand, although a case study can provide interesting insights, its results should definitely be used with caution. Furthermore, in addition to the use of simulated processes, which has proved pivotal in delivering the pursued generalization, the use of an adequate number of forecasting methods and metrics in the present study is also of crucial importance. Using fewer forecasting methods and fewer metrics would have led to a very different overall picture, particularly if those fewer metrics corresponded to fewer criteria. Besides, the comparison is rather the only available research method for any evaluation and, consequently, the bigger its scale the more generalized the derived results. For this specific reason, the novel methodological approach of the present study is considered appropriate for the assessment of forecasting methods.

In fact, our methodology enables the assessment of the failure risk or, alternatively worded, the available opportunities for success that accompany the use of a specific forecasting method to a significant extent, while it also leads to the recognition of several advantages/disadvantages characterizing the latter. This knowledge is fundamental to the forecasters and the users of the forecasts, since a specific forecasting method can be both useful and useless, depending on the forecasting task. Regarding the limitations of the present study, we model processes using the widely adopted ARMA and ARFIMA models, which, however, are not always useful in modelling real-world cases. The latter may require the use of non-normal rather than normal variables. Additionally, we do not investigate an optimum way for selecting the lagged predictor variables used to build the regression matrix in forecasting using ML algorithms neither we focus our research on hyperparameter optimization.

4.3 Recommendations for further research

As a continuation of the present study, we propose the conduct of several properly designed large-scale simulation experiments with the specific aim to facilitate the comparative assessment of a sufficient number of forecasting methods regarding their one-step ahead forecasting properties, which are also of practical interest. We recommend the computation of both the absolute errors and the errors within the proposed study, since they provide assessment regarding the Type 1 and Type 2 accuracy criteria respectively. Moreover, methodological elements of the present study, such as the simulated stochastic processes and the forecasting methods, could also be adopted for the comparison under discussion. In fact, the methodological approach followed here is highly appropriate for the assessment of any forecasting method, while there are several interesting metrics that could be incorporated in the comparisons, e.g. metrics used in forecasting competitions, if considered necessary.

Also, the understanding of the theoretical properties of the forecasting methods, motivated by scientific questions as well as engineering challenges, presupposes systematic and focused on each of them research. Additionally, the investigation of the capabilities that each metric provides regarding the quantification of the forecasting methods' performance would be useful. Simulations combined with statistical analysis can likewise constitute a robust approach to the subject under consideration. Moreover, the results of the present study could also be evaluated in this specific light. Other essential context-independent questions to be answered in the future concern the variation over time of the error values in multi-step ahead forecasting and the effect of the hyperparameter optimization and the lagged predictor variables selection in time series forecasting using ML algorithms. Regarding the latter, Tyralis and Papacharalampous (2017) investigated the performance of RF in one-step ahead time series forecasting as a function of the number of lagged predictor variables and found that less recent lagged predictor variables result in better performance. A possible reason is that increasing the number of predictor variables results in reducing the length of the training set. This may explain the difference in the performance of ML algorithms between simple regression problems (in which the predictor variables are exogenous) and time series forecasting problems. Finally and above all, the intensification of the research on probabilistic forecasting (e.g. Tyralis and Koutsoyiannis 2014) and its

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

effective exploitation by the users (e.g. Ramos et al. 2010, Ramos et al. 2013) should be thoroughly considered.

5. Summary and conclusions

In the present study we conduct an extensive comparison between several stochastic and machine learning methods for the multi-step ahead forecasting of hydrological processes by performing large-scale computational experiments based on simulations. The purpose is to provide generalized results, while the respective comparisons in the literature are usually based on case studies. The time series are generated by linear stationary stochastic processes, which are widely used for the modelling of hydrological processes. Additionally, we conduct 92 case studies using mean monthly time series of streamflow processes and particularly focus on one of them to reinforce the findings and highlight important facts. Despite this specific focus, the results concern all natural processes that could be modelled by stationary processes and all possible time scales and, thus, constitute a theoretical contribution in hydrology and beyond. This specific contribution is twofold; including a hopefully useful in future studies quantification of the performance of several popular forecasting methods in respect to a sufficient number of metrics and, perhaps more importantly, context-independent answers to several theoretical questions that have undoubtedly attracted the attention in the field of hydrological time series forecasting.

The most significant outcome is that none of the forecasting methods is uniformly better or worse than the rest. In other words, there is no best or worst forecasting method regarding all the criteria set simultaneously. This is particularly important, because it reveals that the forecast quality is subject to certain limitations. This is also consistent with the no free lunch theorem, albeit the theorem refers to an infinite space of problems instances, while here we examined a finite space of problems. The empirical investigation showed that in the given finite space (simulated time series and real world cases studies of streamflow time series) is still satisfied. Nevertheless, there are forecasting methods regularly better or worse than others with respect to specific metrics, while there are also forecasting methods sharing a quite similar performance. The latter occur due to the theoretical properties, which accompany by design each forecasting method. This specific observation allows the wording of several advantages/disadvantages of the forecasting methods. Moreover, it appears that, although a general ranking of the forecasting methods is not possible, their classification based on their similar or contrasting performance in the various metrics is possible to some extent. The findings also suggest that more sophisticated methods do not necessarily provide better forecasts compared to simpler methods. Finally, it is pointed out that the machine learning methods do not differ dramatically from the stochastic, except for the fact that the former are computationally intensive. In fact, both the main categories seem to be equally competent in time series forecasting as also subjects to the same limitations. These limitations might be a good reason for probabilistic forecasting.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abudu S, Cui C, King JP, Abudukadeer K (2016) Comparison of performance of statistical models in forecasting monthly streamflow of Kizil River, China. Water Science and Engineering 3(3):269-281. doi:10.3882/j.issn.1674-2370.2010.03.003

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

Ahmed NK, Atiya AF, GayarAn NE, El-Shishiny H (2010) An empirical comparison of machine learning models for time series forecasting. Econometric Reviews 29(5-6):594-621. doi:10.1080/07474938.2010.481556

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Asefa T, Kemblowski M, Lall U, Urroz G (2005) Support vector machines for nonlinear state space reconstruction: Application to the Great Salt Lake time series. Water Resources Research 41:W12422. doi:10.1029/2004WR003785

Assimakopoulos V, Nikolopoulos K (2000) The theta model: a decomposition approach to forecasting. International Journal of Forecasting 16(4):521-530. doi:10.1016/S0169-2070(00)00066-2

Atiya AF, El-Shoura SM, Shaheen SI, El-Sherif MS (1999) A comparison between neural-network forecasting techniques-case study: river flow forecasting. IEEE Transactions on Neural Networks 10(2):402-409. doi:10.1109/72.750569

Ballini R, Soares S, Andrade MG (2001) Multi-step-ahead monthly streamflow forecasting by a neurofuzzy network model. IFSA World Congress and 20th NAFIPS International Conference:992-997. doi:10.1109/NAFIPS.2001.944740

Belayneh A, Adamowski J, Khalil B, Ozga-Zielinski B (2014) Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. Journal of Hydrology 508:418-429. doi:10.1016/j.jhydrol.2013.10.052

Bontempi G (2013) Machine Learning Strategies for Time Series Prediction. European Business Intelligence Summer School, Hammamet, Lecture. 2013. Available online: https://pdfs.semanticscholar.org/f8ad/a97c142b0a2b1bfe20d8317ef58527ee329a.pdf (accessed on 14 October 2017)

Bontempi G, Taieb SB, Le Borgne YA (2013) Machine learning strategies for time series forecasting. In: Aufaure MA, Zimányi E (eds) Business Intelligence. Springer Berlin Heidelberg, pp 62-77. doi:10.1007/978-3-642-36318-4_3

Box GEP, Jenkins GM (1968) Some recent advances in forecasting and control. Journal of the Royal Statistical Society. Series C (Applied Statistics) 17(2):91-109. doi:10.2307/2985674

Breiman L (2001a) Random Forests. Machine Learning 45(1):5–32. doi:10.1023/A:1010933404324

Breiman L (2001b) Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199-231

Carlson RF, MacCormick AJA, Watts DG (1970) Application of linear random models to four annual streamflow series. Water Resources Research 6(4):1070-1078. doi:10.1029/WR006i004p01070

Chen J, Li M, Wang W (2012) Statistical uncertainty estimation using random forests and its application to drought forecast. Mathematical Problems in Engineering 2012. doi:10.1155/2012/915053

Cheng CT, Xie JX, Chau KW, Layeghifard M (2008) A new indirect multi-step-ahead prediction model for a long-term hydrologic prediction. Journal of Hydrology 361(1-2):118-130. doi:10.1016/j.jhydrol.2008.07.040

Cortez P (2010) Data mining with neural networks and support vector machines using the R/rminer tool. In: Perner P (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 (2016) rminer: Data Mining Classification and Regression Methods. R package version 1.4.2. https://CRAN.R-project.org/package=rminer

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3):273-297. doi:10.1007/BF00994018

Criss RE, Winston WE (2008) Do Nash values have value? Discussion and alternate proposals. Hydrological Processes 22:2723-2725. doi:10.1002/hyp.7072

Gable GG (1994) Integrating case study and survey research methods: an example in information systems. European Journal of Information Systems 3(2):112-126. doi:10.1057/ejis.1994.12

De Gooijer JG, Hyndman RJ (2006) 25 years of time series forecasting. International Journal of Forecasting 22(3):443-473. doi:10.1016/j.ijforecast.2006.01.001

Guo J, Zhou J, Qin H, Zou Q, Li Q (2011) Monthly streamflow forecasting based on improved support vector machine model. Expert Systems with Applications 38(10): 13073-13081. doi:10.1016/j.eswa.2011.04.114

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377(1-2):80-91. doi:10.1016/j.jhydrol.2009.08.003

Fraley C, Leisch F, Maechler M, Reisen V, Lemonte A (2012) fracdiff: Fractionally differenced ARIMA aka ARFIMA(p,d,q) models. R package version 1.4-2. https://CRAN.R-project.org/package=fracdiff

He Z, Wen X, Liu H, Du J (2014) A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flown in the semiarid mountain region. Journal of Hydrology 509:379-386. doi:10.1016/j.jhydrol.2013.11.054

Hong T, Fan S (2016) Probabilistic electric load forecasting: A tutorial review. International Journal of Forecasting 32(3):914-938. doi:10.1016/j.ijforecast.2015.11.011

Hong WC (2008) Rainfall forecasting by technological machine learning models. Applied Mathematics and Computation 200(1): 41-57. doi:10.1016/j.amc.2007.10.046

Hu J, Liu J, Liu Y, Gao C (2001) EMD-KNN model for annual average rainfall forecasting. Journal of Hydrologic Engineering 18(11):1450-1457. doi:10.1061/(ASCE)HE.1943-5584.0000481

Humphrey GB, Maier HR, Wu W, Mount NJ, Dandy GC, Abrahart RJ, Dawson CW (2017) Improved validation framework and R-package for artificial neural network models. Environmental Modelling & Software 92:82-106. doi:10.1016/j.envsoft.2017.01.023

Hyndman RJ, Athanasopoulos G (2013) Forecasting: principles and practice. OTexts: Melbourne, Australia. http://otexts.org/fpp/

Hyndman RJ, O'Hara-Wild M, Bergmeir C, Razbash S, Wang E (2017) forecast: Forecasting functions for time series and linear models. R package version 8.0. https://CRAN.R-project.org/package=forecast

Hyndman RJ, Billah B (2003) Unmasking the Theta method. International Journal of Forecasting 19(2):287-290. doi:10.1016/S0169-2070(01)00143-1

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. International Journal of Forecasting 22(4):679-688. doi:10.1016/j.ijforecast.2006.03.001

Hyndman RJ, 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

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) Forecasting with exponential smoothing: The state space approach. Springer - Verlag Berlin Heidelberg, pp 3-7. doi:10.1007/978-3-540-71918-2

Jain SK, Das A, Srivastava DK (1999) Application of ANN for reservoir inflow prediction and operation. Journal of Water Resources 125(5):263-271. doi:10.1061/(ASCE)0733-9496(1999)125:5(263)

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9): 1-20

Khan MS, Coulibaly P (2006) Application of support vector machine in lake water level prediction. Journal of Hydrologic Engineering 11(3):199-205. doi:10.1061/(ASCE)1084-0699(2006)11:3(199)

Kim TW, Valdés JB (2003) Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. Journal of Hydrologic Engineering 8(6):319-328. doi:10.1061/(ASCE)1084-0699(2003)8:6(319)

Kişi Ö (2004) River flow modeling using artificial neural networks. Journal of Hydrologic Engineering 9(1):60-63. doi:10.1061/(ASCE)1084-0699(2004)9:1(60)

Kişi Ö (2007) Streamflow forecasting using different artificial neural network algorithms. Journal of Hydrologic Engineering 12(5):532-539. doi:10.1061/(ASCE)1084-0699(2007)12:5(532)

Kişi Ö, Cimen M (2011) A wavelet-support vector machine conjunction model for monthly streamflow forecasting. Journal of Hydrology 399(1-2):132-140. doi:10.1016/j.jhydrol.2010.12.041

Kişi Ö, Cimen M (2012) Precipitation forecasting by using wavelet-support vector machine conjunction model. Engineering Applications of Artificial Intelligence 25(4): 783-792. doi:10.1016/j.engappai.2011.11.003

Kişi Ö, Shiri J, Nikoofar B (2012) Forecasting daily lake levels using artificial intelligence approaches. Computers & Geosciences 41:169-180. doi:10.1016/j.cageo.2011.08.027

Kitanidis PK, Bras RL (1980) Real time forecasting with a conceptual hydrologic model: 2. Applications and results. Water Resources Research 16(6):1034-1044. doi:10.1029/WR016i006p01034

Koutsoyiannis D (2007) Discussion of "Generalized regression neural networks for evapotranspiration modelling". Hydrological Sciences Journal 52(4):832-835. doi:10.1623/hysj.52.4.832

Koutsoyiannis D (2010) HESS Opinions "A random walk on water". Hydrology and Earth System Sciences 14:585–601. doi:10.5194/hess-14-585-2010

Koutsoyiannis D (2011) Hurst-Kolmogorov Dynamics and Uncertainty. Journal of the American Water Resources Association 47(3):481-495. doi:10.1111/j.1752-1688.2011.00543.x

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

Krause P, Boyle DP, Bäse F (2005) Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5:89-97

Krzysztofowicz R (2001) The case for probabilistic forecasting in hydrology. Journal of Hydrology 249(1-4):2-9. doi:10.1016/S0022-1694(01)00420-6

Lambrakis N, Andreou AS, Polydoropoulos P, Georgopoulos E, Bountis T (2000) Nonlinear analysis and forecasting of a brackish karstic spring. Water Resources Research 36(4):875-884. doi:10.1029/1999WR900353

Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2(3):18-22

Lin JY, Cheng CT, Chau KW (2006) Using support vector machines for long-term discharge prediction. Hydrological Sciences Journal 51(4):599-612. doi:10.1623/hysj.51.4.599

Liong SY, Sivapragasam C (2002) Flood stage forecasting with support vector machines. Journal of the American Water Resources Association 38(1):173-186. doi:10.1111/j.1752-1688.2002.tb01544.x

Lu K, Wang L (2011) A novel nonlinear combination model based on support vector machine for rainfall prediction. Fourth International Joint Conference on Computational Sciences and Optimization:1343-1346. doi:10.1109/CSO.2011.50

Makridakis S, Hibon M (1987) Confidence intervals: An empirical investigation of the series in the M-competition. International Journal of Forecasting 3(3-4):489-508. doi:10.1016/0169-2070(87)90045-8

Makridakis S, Hibon M (2000) The M3-Competition: results, conclusions and implications. International Journal of Forecasting 16(4):451-476. doi:10.1016/S0169-2070(00)00057-1

Maier HR, Dandy GC (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

Mishra AK, Desai VR, Singh VP (2007) Drought forecasting using a hybrid stochastic and neural network model. Journal of Hydrologic Engineering 12(6):26-638. doi:10.1061/(ASCE)1084-0699(2007)12:6(626)

Montanari A, Rosso R, Taqqu MS (1997) Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. Water Resources Research 33(5):1035–1044. doi:10.1029/97WR00043

Montanari A, Rosso R, Taqqu MS (2000) A seasonal fractional ARIMA Model applied to the Nile River monthly flows at Aswan. Water Resources Research 36(5):1249–1259. doi:10.1029/2000WR900012

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 50(3):885-900

Murphy AM (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting 8:28-293. doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology 10(3):282-290. doi:10.1016/0022-1694(70)90255-6

Pai PF, Hong WC (2007) A recurrent support vector regression model in rainfall forecasting. Hydrological Processes 21:819-827. doi:10.1002/hyp.6323

Papacharalampous GA (2016) Theoretical and empirical comparison of stochastic and machine learning methods for hydrological processes forecasting. MSc thesis. http://www.itia.ntua.gr/en/docinfo/1670/

Papacharalampous GA, Tyralis H, Koutsoyiannis D (2017) Forecasting of geophysical processes using stochastic and machine learning algorithms. 10th World Congress of EWRA on Water Resources and Environment. "Panta Rhei":527-534

Pappenberger F, Ramos MH, Clok e HL, Wetterhall F, Alfieri L, Bogner K, Mueller A, Salamon P (2015) How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. Journal of Hydrology 522:697-713. doi:10.1016/j.jhydrol.2015.01.024

Patel SS, Ramachandran P (2015) A comparison of machine learning techniques for modeling river flow time series: the case of upper Cauvery river basin. Water Resources Management 29(2):589-602. doi:10.1007/s11269-014-0705-0

Peel MC, Chiew FHS, Western AW, McMahon TA (2000) Extension of unimpaired monthly streamflow data and regionalisation of parameter values to estimate streamflow in ungauged catchments, Report prepared for the National Land and Water Resources Audit

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raghavendra NS, Deka PC (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

Ramos MH, Mathevet T, Thielen J, Pappenberger F (2010) Communicating uncertainty in hydro-meteorological forecasts: mission impossible?. Meteorological Applications 17(2):223-235. doi:10.1002/met.202

Ramos MH, Van Andel SJ, Pappenberger F (2013) Do probabilistic forecasts lead to better decisions?. Hydrology and Earth System Sciences 17:2219-2232. doi:10.5194/hess-17-2219-2013

Sapankevych NI, Sankar R (2009) Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine 4(2):24-38. doi:10.1109/MCI.2009.932254

Shabri A, Suhartono (2012) Streamflow forecasting using least-squares support vector machines. Hydrological Sciences Journal 57(7):1275-1293. doi:10.1080/02626667.2012.714468

Shi Z, Han M (2007) Support vector echo-state machine for chaotic time-series prediction. IEEE Transactions on Neural Networks 18(2):359-372. doi:10.1109/TNN.2006.885113

Singh M, Singh R, Shinde V (2011) Application of software packages for monthly stream flow forecasting of Kangsabati River in India. International Journal of Computer Applications 20(3):7-14

Sivapragasam C, Liong SY, Pasha MFK (2001) Rainfall and runoff forecasting with SSA-SVM approach. Journal of Hydroinformatics 3(3):141-152

Shmueli G (2010) To explain or to predict?. Statistical Science 25(3):289-310. doi:10.1214/10-STS330

Taieb SB, Bontempi G, Atiya AF, 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

Thissen U, Van Brakel R, De Weijer AP, Melssena WJ, Buydens LMC (2003) Using support vector machines for time series prediction. Chemometrics and Intelligent Laboratory Systems 69(1-2):35-49. doi:10.1016/S0169-7439(03)00111-4

Tongal H, Berndtsson R (2016) Impact of complexity on daily and multi-step forecasting of streamflow with chaotic, stochastic, and black-box models. Stochastic Environmental Research and Risk Assessment:1-22. doi:10.1007/s00477-016-1236-4

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Tran HD, Muttil N, Perera BJC (2015) Selection of significant input variables for time series forecasting. Environmental Modelling & Software 64:156-163. doi:10.1016/j.envsoft.2014.11.018

Tyralis H (2016) HKprocess: Hurst-Kolmogorov Process. R package version 0.0-2. https://CRAN.R-project.org/package=HKprocess

Tyralis H, Koutsoyiannis D (2011) Simultaneous estimation of the parameters of the Hurst–Kolmogorov stochastic process. Stochastic Environmental Research and Risk Assessment 25(1):21-33. doi:10.1007/s00477-010-0408-x

Tyralis H, Koutsoyiannis D (2014) A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. Climate Dynamics 42(11):2867-2883. doi:10.1007/s00382-013-1804-y

Tyralis H, Papacharalampous GA (2017) Variable selection in time series forecasting using random forests. Algorithms 10(4):114. doi:10.3390/a10040114

Valipour M, Banihabib ME, Behbahani SMR (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

Vapnik VN (1995) The nature of statistical learning theory, first edition. Springer-Verlag New York. doi:10.1007/978-1-4757-3264-1

Vapnik VN (1999) An overview of statistical learning theory. IEEE Transactions on Neural Networks 10(5):988-999. doi:10.1109/72.788640

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, fourth edition. Springer-Verlag New York. doi:10.1007/978-0-387-21706-2

LeVeque RJ, Mitchell IM, Stodden V (2012) Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture. Computing in Science and Engineering 14(4):13-17

Wang WC, Chau KW, Cheng CT, Qiu L (2009) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. Journal of Hydrology 374(3-4):294-306. doi:10.1016/j.jhydrol.2009.06.019

Wei WWS (2006) Time Series Analysis, Univariate and Multivariate Methods, second edition. Pearson Addison Wesley

Weijs SV, Schoups G, Van de Giesen N (2010) Why hydrological predictions should be evaluated using information theory. Hydrology and Earth System Sciences 14:2545-2558. doi:10.5194/hess-14-2545-2010

Wolpert DH (1996) The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation 8(7):1341-1390. doi:10.1162/neco.1996.8.7.1341

Yaseen ZM, Allawi MF, Yousif AA, Jaafar O, Hamzah FM, El-Shafie A (2016) Non-tuned machine learning approach for hydrological time series forecasting. Neural Computing and Applications. doi:10.1007/s00521-016-2763-0

Yapo PO, Gupta HV, Sorooshian S (1996) Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. Journal of Hydrology 181(1-4):23-48. doi:10.1016/0022-1694(95)02918-4

Yin RK (1994) Case study research: Design and methods, second edition. SAGE Publications

Yu X, Liong SY (2007) Forecasting of hydrologic time series with ridge regression in feature space. Journal of Hydrology 332(3-4):290-302. doi:10.1016/j.jhydrol.2006.07.003

Yu X, Liong SY, Babovic V (2004) EC-SVM approach for real-time hydrologic forecasting. Journal of Hydroinformatics 6(3):209-223

Zaini N, Malek MA, Yusoff M (2015) Application of computational intelligence methods in modelling river flow prediction: A review. International Conference on Computer, Communications and Control Technology (I4CT):370-374. doi:10.1109/I4CT.2015.7219600

Zambrano-Bigiarini M (2014) hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3-8. https://CRAN.R-project.org/package=hydroGOF

Zhang GP (2001) An investigation of neural networks for linear time-series forecasting. Computers & Operations Research 28(12):1183-1202. doi:10.1016/S0305-0548(00)00033-2

Zhang GP, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks:: The state of the art. International Journal of Forecasting 14(1):35-62. doi:10.1016/S0169-2070(97)00044-7

Appendix A Definition of the stochastic processes

In Appendix A, we briefly present the mathematical background of the Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) and Autoregressive Fractionally Integrated Moving Average (ARFIMA) stochastic processes, while the reader is also referred to Wei (2006 pp.6-87, 489-494).

A time series in discrete time is defined as a sequence of observations x_1, x_2, … of a certain phenomenon, while the time t is stated as a subscript to each value x_t. A time series can be modelled by a stochastic process. The latter is a sequence of random variables $\underline{x}_1$, $\underline{x}_2$, ….

Let us consider a stochastic process of normally distributed random variables. The mean function μ_t of the stochastic process is defined by

$$\mu_t := \mathrm{E}[\underline{x}_t] \tag{A.1}$$

The standard deviation function σ_t of the stochastic process is defined by

$$\sigma_t := \sqrt{\mathrm{Var}[\underline{x}_t]} \tag{A.2}$$

The covariance function between $\underline{x}_{t_1}$ and $\underline{x}_{t_2}$, $\gamma(t_1,t_2)$ of the stochastic process is defined by

$$\gamma(t_1, t_2) := \mathrm{E}[(\underline{x}_{t_1} - \mu_{t_1})(\underline{x}_{t_2} - \mu_{t_2})] \tag{A.3}$$

The correlation function between $\underline{x}_{t_1}$ and $\underline{x}_{t_2}$, $\rho(t_1,t_2)$ of the stochastic process is defined by

$$\rho(t_1, t_2) := \gamma(t_1,t_2)/(\sigma_{t_1}\,\sigma_{t_2}) \qquad \text{(A.4)}$$

For a strictly stationery stochastic process Equations (A.5) - (A.8) must be satisfied:

$$\mu_t = \mu\ \forall\ t \in \{1, 2, \ldots\} \qquad \text{(A.5)}$$

$$\sigma_t = \sigma\ \forall\ t \in \{1, 2, \ldots\} \qquad \text{(A.6)}$$

$$\gamma(t_1, t_2) = \gamma(t_1 + k, t_2 + k)\ \forall\ t_1, t_2, k \text{ integers} \qquad \text{(A.7)}$$

$$\rho(t_1, t_2) = \rho(t_1 + k, t_2 + k)\ \forall\ t_1, t_2, k \text{ integers} \qquad \text{(A.8)}$$

In this case, let us consider that:

$$t_1 := t - k, t_2 := t \qquad \text{(A.9)}$$

Then we have:

$$\gamma(t_1, t_2) = \gamma(t - k, t) = \gamma(t, t + k) = \gamma_k \qquad \text{(A.10)}$$

$$\rho(t_1, t_2) = \rho(t - k, t) = \rho(t, t + k) = \rho_k \qquad \text{(A.11)}$$

Using the Equations (A.5) - (A.11), the autocovariance function γ_k and the autocorrelation function ρ_k of a stationary stochastic process are defined with Equations (A.12) and (A.13) respectively.

$$\gamma_k := \mathrm{E}[(\underline{x}_t - \mu)(\underline{x}_{t+k} - \mu)] \qquad \text{(A.12)}$$

$$\rho_k := \gamma_k / \sigma^2 \qquad \text{(A.13)}$$

For a stationary stochastic process the partial autocorrelation function P_k is defined by

$$P_k := \mathrm{Corr}[(\underline{x}_t, \underline{x}_{t+k} \mid \underline{x}_{t+1}, \ldots, \underline{x}_{t+k-1})] \qquad \text{(A.14)}$$

The partial autocorrelation function is the correlation between two random variables $\underline{x}_t$ and $\underline{x}_{t+k}$, with the linear dependency between the intervening variables $\underline{x}_{t+1}$, …, $\underline{x}_{t+k-1}$ removed.

A strictly stationary stochastic process $\{\underline{a}_t\}$ is called a white noise process, if it is a sequence of uncorrelated random variables. Let us consider hereinafter that the white noise is a normal variable with zero mean, unless mentioned otherwise, and standard deviation σ_a.

AR(p), MA(q), ARMA(p,q) models

The stochastic process $\{\underline{y}_t\}$ is defined by

$$\underline{y}_t := \underline{x}_t - \mu \tag{A.15}$$

Let us consider the operator B, which is defined by

$$\mathrm{B}^j \underline{x}_t = \underline{x}_{t-j} \tag{A.16}$$

Then the operator $\varphi_p(\mathrm{B})$ is defined by

$$\varphi_p(\mathrm{B}) := (1 - \varphi_1 \mathrm{B} - \ldots - \varphi_p \mathrm{B}^p) \tag{A.17}$$

The stochastic process $\{\underline{x}_t\}$ is an AR(p), if the following equation holds:

$$\varphi_p(\mathrm{B}) \underline{y}_t = \underline{a}_t \tag{A.18}$$

that can be written in the following form:

$$\underline{y}_t = \varphi_1 \underline{y}_{t-1} + \ldots + \varphi_p \underline{y}_{t-p} + \underline{a}_t \tag{A.19}$$

Let us also consider the operator $\theta_q(\mathrm{B})$, which is defined by

$$\theta_q(\mathrm{B}) := 1 + \theta_1 \mathrm{B} + \ldots + \theta_q \mathrm{B}^q \tag{A.20}$$

The stochastic process $\{\underline{x}_t\}$ is a MA(q), if the following equation holds:

$$\underline{y}_t = \theta_q(\mathrm{B}) \underline{a}_t \tag{A.21}$$

that can be written in the following form:

$$\underline{y}_t = \underline{a}_t + \theta_1\underline{a}_{t-1} + \ldots + \theta_q\underline{a}_{t-q} \qquad \text{(A.22)}$$

The stochastic process $\{\underline{x}_t\}$ is an ARMA(p,q), if the following equation holds:

$$\varphi_p(\mathrm{B})\underline{y}_t = \theta_q(\mathrm{B})\underline{a}_t \qquad \text{(A.23)}$$

that can be written in the following form:

$$\underline{y}_t = \varphi_1\underline{y}_{t-1} + \ldots + \varphi_p\underline{y}_{t-p} + \underline{a}_t + \theta_1\underline{a}_{t-1} + \ldots + \theta_q\underline{a}_{t-q} \qquad \text{(A.24)}$$

ARIMA(p,d,q) models

Let d be a natural number. Then the stochastic process $\{\underline{x}_t\}$ is an ARIMA(p,d,q), if the following equation holds:

$$\varphi_p(\mathrm{B})(1\text{-}\mathrm{B})^d\underline{x}_t = \theta_0 + \theta_q(\mathrm{B})\underline{a}_t \qquad \text{(A.25)}$$

If d = 0, then we have an ARMA(p,q) and for θ_0 we obtain:

$$\theta_0 = (1 - \varphi_1 - \ldots - \varphi_p)\mu \qquad \text{(A.26)}$$

If $d \geq 1$, then θ_0 is called deterministic trend term and is usually omitted from the model, unless it is truly required. This specific stochastic process is non-stationary (Wei 2006, pp.69).

ARFIMA(p,d,q) models

Let $d \in (-0.5, 0.5)$. The stochastic process $\{\underline{x}_t\}$ is an ARFIMA(p,d,q), if the following equation holds:

$$\varphi_p(\mathrm{B})(1\text{-}\mathrm{B})^d\underline{x}_t = \theta_q(\mathrm{B})\underline{a}_t \qquad \text{(A.27)}$$

In contrast to the stochastic process ARIMA(p,d,q), ARFIMA(p,d,q) is stationary (Wei 2006, pp.489). This specific stochastic process can be used to model processes that are characterized with long-range dependence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

Appendix B Definition of the metrics

In Appendix B we define the metrics used for the comparative assessment of the forecasting methods.

For the definitions we consider a time series of N observations. Let us also consider a model fitted to the first N - n observations of this specific time series and subsequently used to make predictions corresponding to the last n observations. Let x_1, x_2, ..., x_n represent the last n observations and f_1, f_2, ..., f_n represent the forecasts.

Assessment regarding the Type 1 accuracy

The mean absolute error (MAE) metric is defined by

$$\text{MAE} := (1/n) \sum_{i = 1}^{n} |f_i - x_i| \tag{B.1}$$

The mean absolute percentage error (MAPE) metric is defined by

$$\text{MAPE} := (1/n) \sum_{i = 1}^{n} |100(f_i - x_i)/x_i| \tag{B.2}$$

The root mean square error (RMSE) metric is defined by

$$\text{RMSE} := \sqrt{(1/n) \sum_{i = 1}^{n} (f_i - x_i)^2} \tag{B.3}$$

Let $\bar{x}$ be the mean of the observations, which is defined by

$$\bar{x} := (1/n) \sum_{i = 1}^{n} x_i \tag{B.4}$$

The Nash-Sutcliffe Efficiency (NSE) metric is defined by (Nash and Sutcliffe 1970)

$$\text{NSE} := 1 - (\sum_{i = 1}^{n} (f_i - x_i)^2 / \sum_{i = 1}^{n} (x_i - \bar{x})^2) \tag{B.5}$$

The modified Nash-Sutcliffe Efficiency (mNSE) metric is defined by (Krause et al. 2005)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2017 doi:10.20944/preprints201710.0133.v1

$$\text{mNSE} := 1 - \left(\sum_{i=1}^{n} |f_i - x_i| / \sum_{i=1}^{n} |x_i - \bar{x}|\right) \tag{B.6}$$

The relative Nash-Sutcliffe Efficiency (rNSE) metric is defined by (Krause et al. 2005)

$$\text{rNSE} := 1 - \left(\sum_{i=1}^{n} ((f_i - x_i)/ x_i)^2 / \sum_{i=1}^{n} ((x_i - \bar{x})/\bar{x})^2\right) \tag{B.7}$$

The persistence index (cp) metric is defined by (Kitanidis and Bras 1980)

$$\text{cp} := 1 - \left(\sum_{i=2}^{n} (f_i - x_i)^2 / \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2\right) \tag{B.8}$$

Assessment regarding the Type 2 accuracy

The mean error (ME) metric is defined by

$$\text{ME} := (1/n) \sum_{i=1}^{n} (f_i - x_i) \tag{B.9}$$

The mean percentage error (MPE) metric is defined by

$$\text{MPE} := (-1/n) \sum_{i=1}^{n} (100(f_i - x_i)/x_i) \tag{B.10}$$

The percent bias (PBIAS) metric is defined by (Yapo et al. 1996)

$$\text{PBIAS} := 100 \sum_{i=1}^{n} (f_i - x_i) / \sum_{i=1}^{n} (x_i) \tag{B.11}$$

The Volumetric Efficiency (VE) metric is defined by (Criss and Winston 2008)

$$\text{VE} := 1 - \left(\sum_{i=1}^{n} |f_i - x_i| / \sum_{i=1}^{n} x_i\right) \tag{B.12}$$

Assessment regarding the capture of the variance

Let s_x be the standard deviation of the observations, which is defined by

$$s_x := \sqrt{(1/(n-1)) \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{B.13}$$

Let $\bar{f}$ be the mean of the forecasts and s_f be the standard deviation of the forecasts, which are defined with Equations (B.14) and (B.15) respectively.

$$\bar{f} := (1/n) \sum_{i=1}^{n} f_i \tag{B.14}$$

$$s_f := \sqrt{(1/(n\text{-}1)) \sum_{i=1}^{n} (f_i - \bar{f})^2} \tag{B.15}$$

The ratio of standard deviations (rSD) metric is defined by (Zambrano-Bigiarini 2014)

$$\text{rSD} := s_f / s_x \tag{B.16}$$

Assessment regarding the correlation

The Pearson' s correlation coefficient (Pr) metric is defined by (Krause et al. 2005)

$$\text{Pr} := \left(\sum_{i=1}^{n} (x_i - \bar{x})(f_i - \bar{f})\right) / \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (f_i - \bar{f})^2\right)^{0.5} \tag{B.17}$$

The coefficient of determination (r2) metric is defined by (Krause et al. 2005)

$$\text{r2} := (\text{Pr})^2 \tag{B.18}$$

Assessment regarding the Type 1 accuracy and the capture of the variance

The index of agreement (d) metric is defined by (Krause et al. 2005)

$$\text{d} := 1 - \left(\sum_{i=1}^{n} (f_i - x_i)^2 / \sum_{i=1}^{n} (|f_i - \bar{x}| + |x_i - \bar{x}|)^2\right) \tag{B.19}$$

The modified index of agreement (md) metric is defined by (Krause et al. 2005)

$$\text{md} := 1 - \left(\sum_{i=1}^{n} |f_i - x_i| / \sum_{i=1}^{n} (|f_i - \bar{x}| + |x_i - \bar{x}|)\right) \tag{B.20}$$

The relative index of agreement (rd) metric is defined by (Krause et al. 2005)

$$\text{rd} := 1 - \left(\sum_{i=1}^{n} ((f_i - x_i) / x_i)^2 / \sum_{i=1}^{n} ((|f_i - \bar{x}| + |x_i - \bar{x}|) / \bar{x})^2\right) \tag{B.21}$$

Assessment regarding the Type 2 accuracy, the capture of the variance and the correlation

The Kling-Gupta efficiency (KGE) metric is defined by (Gupta et al. 2009)

$$\mathrm{KGE} := 1 - \sqrt{(\mathrm{Pr} - 1)^2 + ((s_f/s_x) - 1)^2 + ((\bar{f}/\bar{x}) - 1)^2} \qquad \text{(B.22)}$$

Appendix C Supplementary material

The supplementary material is available at: http://dx.doi.org/10.17632/fjr8244m35.1. We provide the fully reproducible reports together with their codes. We also provide a report entitled "Selected figures for the qualitative comparison of the forecasting methods'', which we suggest to be read alongside with Section 3.1. Finally, we provide a report entitled "Qualitative comparison of the forecasting methods in 92 real-world case studies".