A set of metrics for the effective evaluation of point forecasting methods used for hydrological tasks

Georgia Papacharalampous, Hristos Tyralis, and Demetris Koutsoyiannis

Asia Oceania Geosciences Society 14th Annual Meeting
Singapore, 06-11 August 2017
Session HS01: Challenges in Hydrologic Modeling

Presentation code: HS01-A001
Available online at: itia.ntua.gr/1718

Department of Water Resources and Environmental Engineering,
School of Civil Engineering, National Technical University of Athens
(papacharalampous.georgia@gmail.com)

1. Abstract

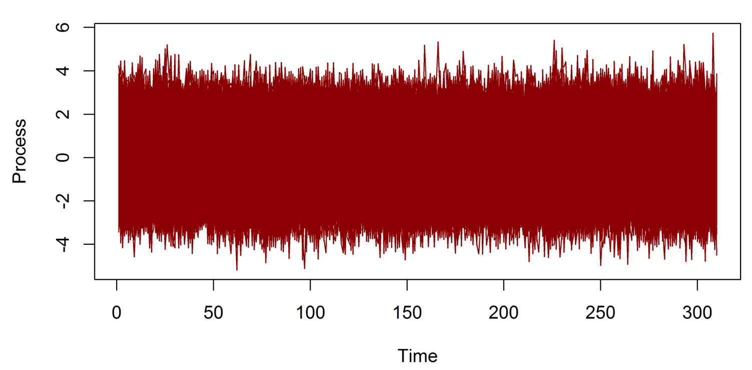
The selection of metrics for the evaluation of point forecasting methods can be challenging even for very experienced hydrologists. We conduct a large-scale computational experiment based on simulations to compare the information that 18 metrics proposed in the literature give about the forecasting performance. Our purpose is to provide generalized results; thus we use 2 000 simulated Autoregressive Fractionally Integrated Moving Average time series. We apply several forecasting methods and we compute the values of the metrics for each forecasting experiment. Subsequently, we measure the correlation between the values of each pair of metrics, separately for each forecasting method. Furthermore, we explore graphically the detected relationships. Finally, we propose a set of metrics that we consider to be suitable for the effective evaluation of point forecasting methods.

2. Introduction

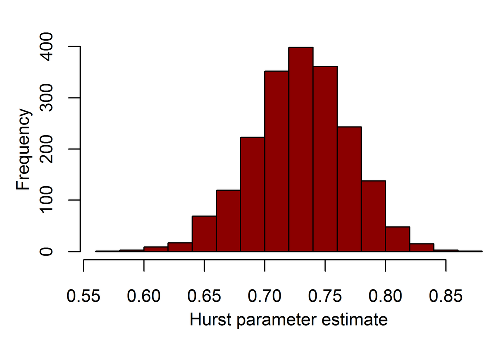
- According to Krause et al. (2005):
  - a) The selection of metrics for the evaluation of point forecasting methods can be challenging even for very experienced hydrologists.
  - b) While there are many available metrics in the literature most of the studies use only a few.
- In contrast with this second observation, Papacharalampous et al. (2017) recently used **18 metrics** (which do not share one-to-one relationships with each other) within an innovative methodological framework aiming at providing generalized results in the field of hydrology regarding the comparison between several stochastic and machine learning point forecasting methods.
- Most of these metrics are available in the R package hydroGOF (Zambrano-Bigiarini 2014). However, the results from this package are rounded at the second decimal digit. For this reason, a function for the computation of these 18 metrics was programmed from scratch in Papacharalampous et al. (2017).
- The present study is devoted to the metrics used in the latter study.
- Our **research question** is: *How close is the information that these 18 metrics provide regarding the performance of point forecasting methods?*
- To answer the above question, we conduct a large-scale computational experiment based on simulations.
- Using the results of this experiment we propose a **set of metrics** for the effective evaluation of point forecasting methods used for hydrological tasks.

3. Methodology outline

- We simulate 2 000 time series of 310 values according to the ARFIMA(0,0.30,0) process using the R package fracdiff (Fraley et al. 2012).
- To describe the long-term persistence of the simulated time series we estimate their Hurst parameter H using the R package HKprocess (Tyralis 2016, see also Tyralis and Koutsoyiannis 2011).



- We apply 4 forecasting methods (RW, auto_ARFIMA, ETS_s and Theta) on the time series using the R package forecast (Hyndman and Khandakar 2008, Hyndman et al. 2017). The code for their impementation can be found in Papacharalampous et al. (2017).
- Regarding the application of the forecasting methods, we split each time series into a fitting and a test set. The latter is the last 10 values. We fit the models to the fitting set and make predictions corresponding to the test set.
- Next, we compute the values of **18 metrics** on the test set. These metrics are also used in Papacharalampous et al. (2017). Their definitions are listed in **4** and **5**.
- We explore graphically the relationships between the metrics (see **6**, **7** and **Supplementary material**).
- Subsequently, we build a tool for the comparison of the information that the metrics provide regarding the performance of the forecasting methods (see **8** and **9**).
- We use this tool to deside on a set of metrics in **10.**



4. Metrics

- We consider a time series of N observations.
- We 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 $x_1, x_2, \ldots, x_n$ represent the last n observations and $f_1, f_2, \ldots, f_n$ represent the forecasts.
- Let $\bar{x}$ be the mean of the observations and s_x be the standard deviation of the observations.
- Let $\bar{f}$ be the mean of the forecasts and s_f be the standard deviation of the forecasts.

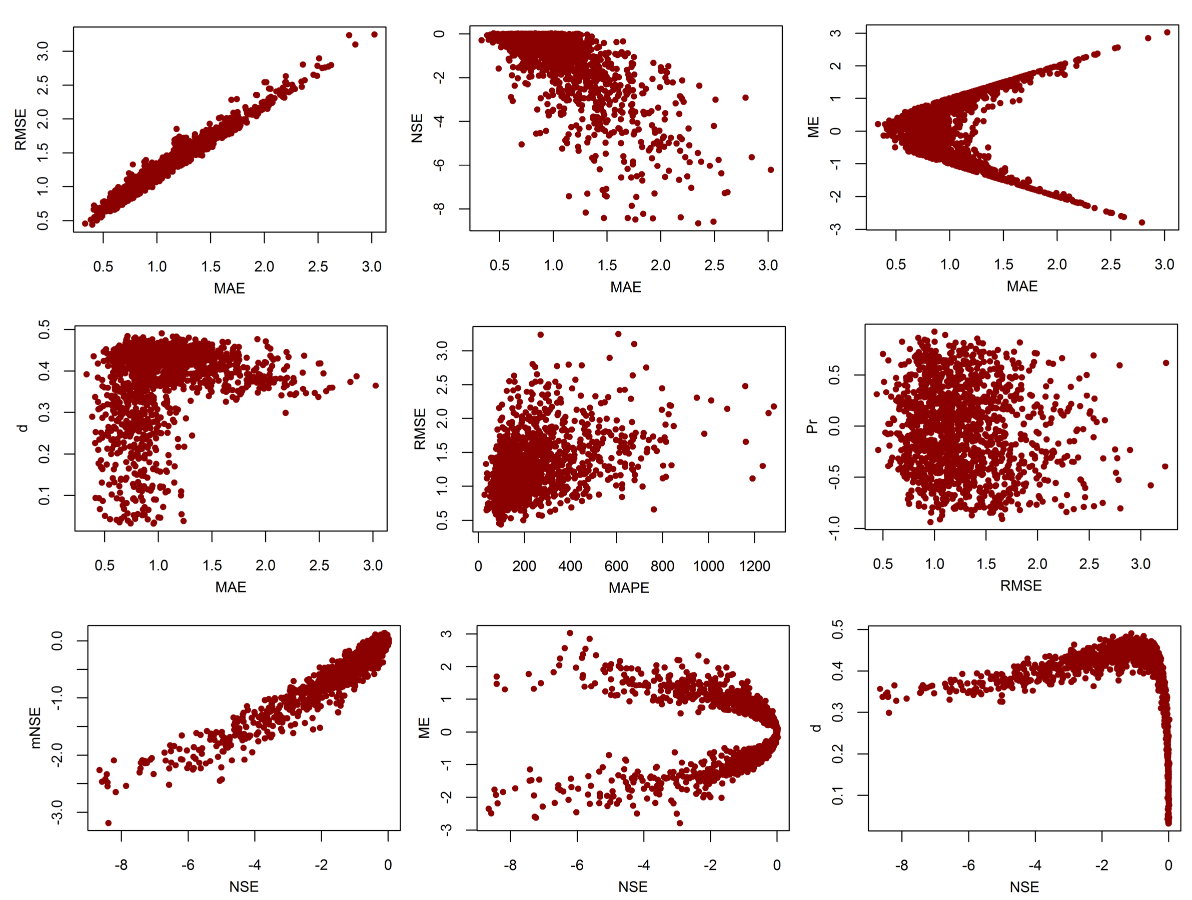
MAE *mean absolute error*	$\text{MAE} := (1/n) \sum_{i=1}^{n} \lvert f_i - x_i \rvert$
MAPE *mean absolute percentage error*	$\text{MAPE} := (1/n) \sum_{i=1}^{n} \lvert 100(f_i - x_i)/x_i \rvert$
RMSE *root mean square error*	$\text{RMSE} := \sqrt{(1/n) \sum_{i=1}^{n} (f_i - x_i)^2}$
NSE *Nash-Sutcliffe efficiency*	$\text{NSE} := 1 - (\sum_{i=1}^{n} (f_i - x_i)^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2)$
mNSE *modified Nash-Sutcliffe efficiency*	$\text{mNSE} := 1 - (\sum_{i=1}^{n} \lvert f_i - x_i \rvert / \sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert)$
rNSE *relative Nash-Sutcliffe efficiency*	$\text{rNSE} := 1 - (\sum_{i=1}^{n} ((f_i - x_i)/x_i)^2 / \sum_{i=1}^{n} ((x_i - \bar{x})/\bar{x})^2)$
cp *index of persistence*	$\text{cp} := 1 - (\sum_{i=2}^{n} (f_i - x_i)^2 / \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2)$
ME *mean error*	$\text{ME} := (1/n) \sum_{i=1}^{n} (f_i - x_i)$

5. Metrics

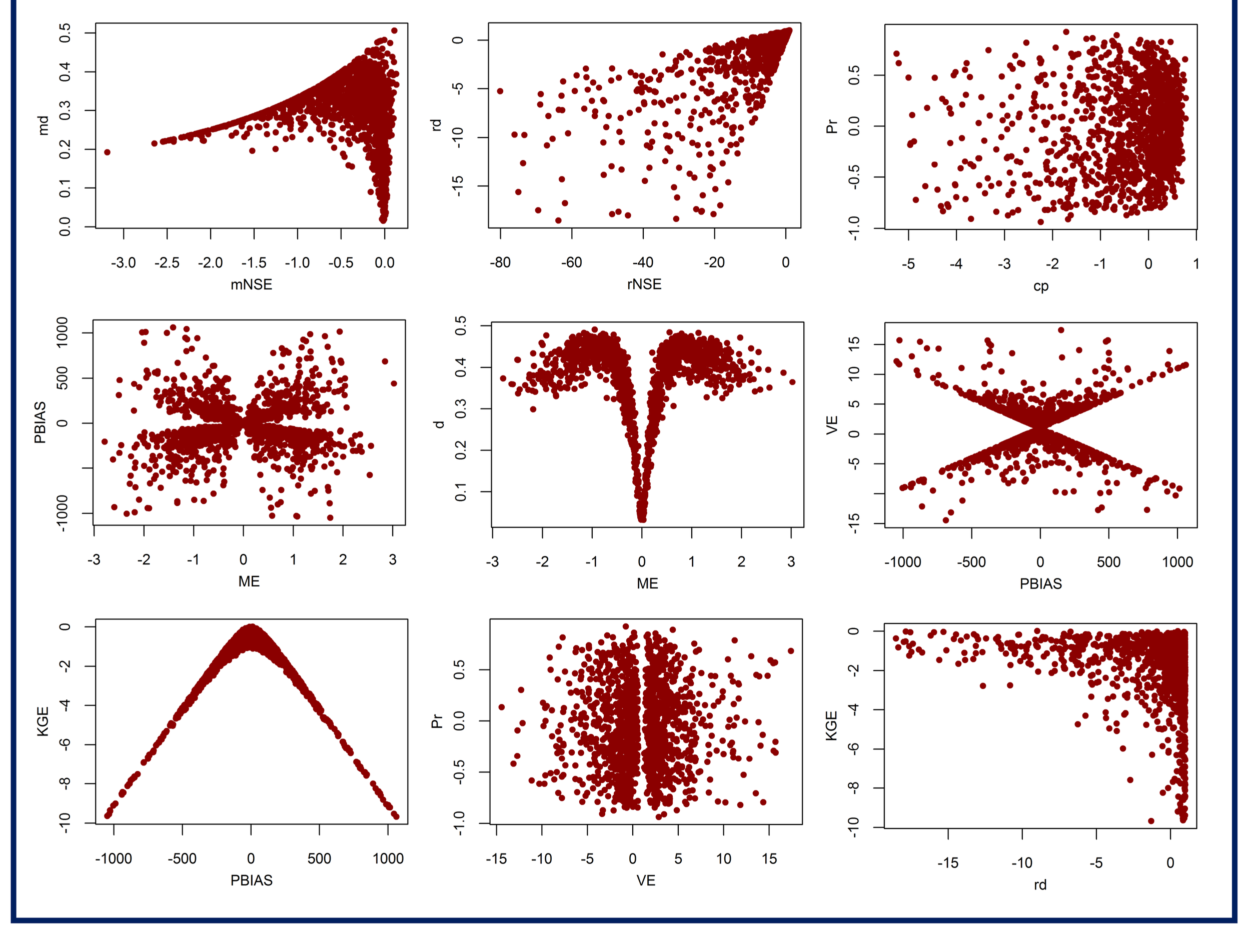
MPE *mean percentage error*	$\text{MPE} := (-1/n) \sum_{i=1}^{n} (100(f_i - x_i)/x_i)$
PBIAS *percent bias*	$\text{PBIAS} := 100 \sum_{i=1}^{n} (f_i - x_i) / \sum_{i=1}^{n} (x_i)$
VE *volumetric efficiency*	$\text{VE} := 1 - (\sum_{i=1}^{n} \lvert f_i - x_i \rvert / \sum_{i=1}^{n} x_i)$
rSD *ratio of standard deviations*	$\text{rSD} := s_f/s_x$
Pr *Pearson' s correlation coefficient*	$\text{Pr} := (\sum_{i=1}^{n} (x_i - \bar{x})(f_i - \bar{f})) / (\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (f_i - \bar{f})^2)^{0.5}$
r2 *coefficient of determination*	$\text{r2} := (\text{Pr})^2$
d *index of agreement*	$\text{d} := 1 - (\sum_{i=1}^{n} (f_i - x_i)^2 / \sum_{i=1}^{n} (\lvert f_i - \bar{x} \rvert + \lvert x_i - \bar{x} \rvert)^2)$
md *modified index of agreement*	$\text{md} := 1 - (\sum_{i=1}^{n} \lvert f_i - x_i \rvert / \sum_{i=1}^{n} (\lvert f_i - \bar{x} \rvert + \lvert x_i - \bar{x} \rvert))$
rd *relative index of agreement*	$\text{rd} := 1 - (\sum_{i=1}^{n} ((f_i - x_i)/x_i)^2 / \sum_{i=1}^{n} ((\lvert f_i - \bar{x} \rvert + \lvert x_i - \bar{x} \rvert)/\bar{x})^2)$
KGE *Kling-Gupta efficiency*	$\text{KGE} := 1 - \sqrt{(\text{Pr} - 1)^2 + ((s_f/s_x) - 1)^2 + ((\bar{f}/\bar{x}) - 1)^2}$

- See also: 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)

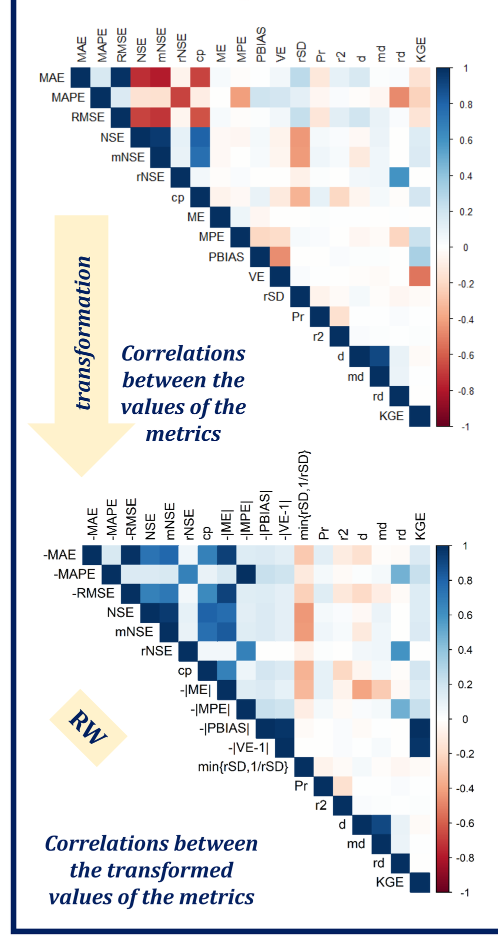
6. Graphical exploration of the relationships: RW



7. Graphical exploration of the relationships: RW



8. Building a tool: Use of a transformation

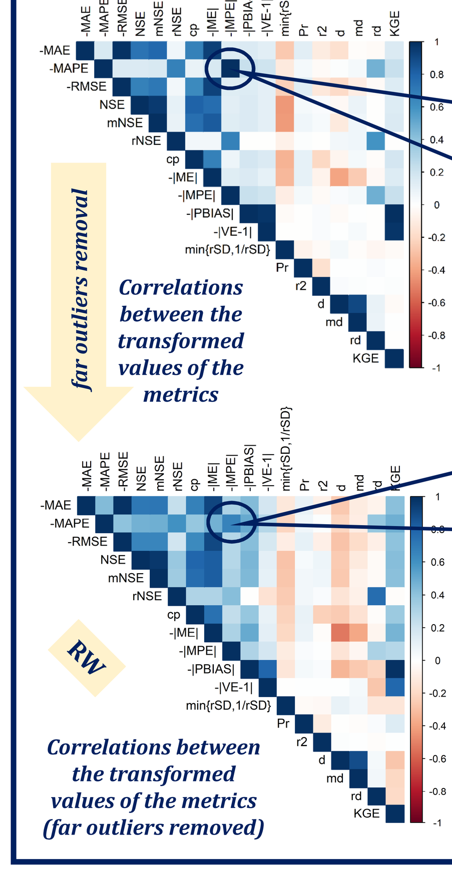


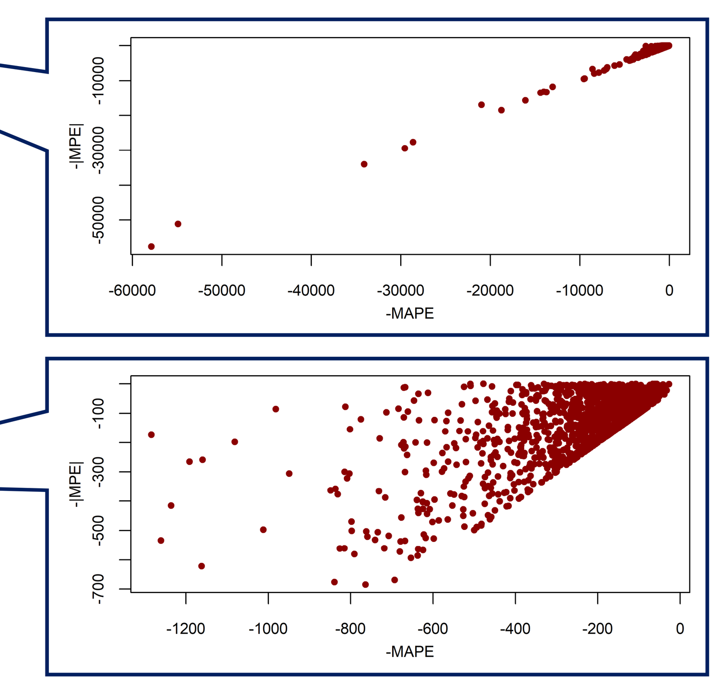
- We measure the correlations between the values of each pair of metrics (see upper figure).
- We transform the values of the metrics according to the following table. **The larger the transformed values the better the forecasts.**
- We measure the correlations between the transformed values (see lower figure).

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

9. Building a tool: Far outliers removal

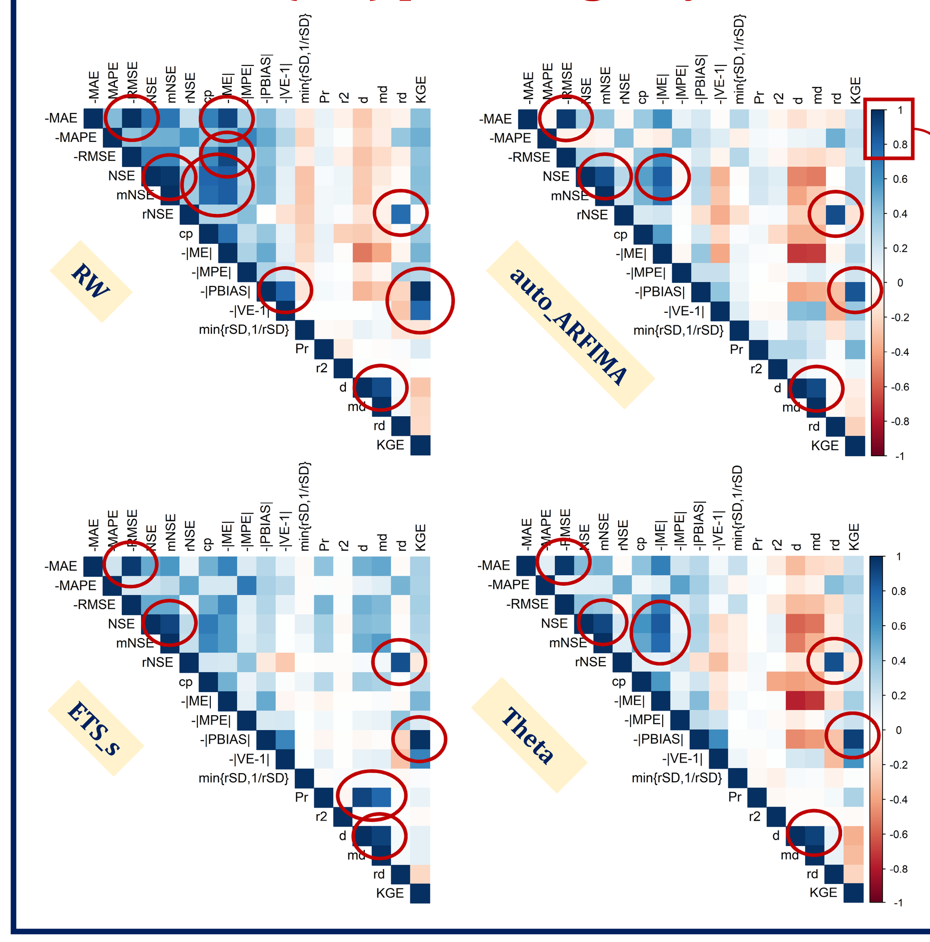
- We remove the far outliers from the datasets to avoid misleading calculation of the correlations (see figures below).





- We use this tool in **10** for the detection of the strong monotonic possitive relationships between the transformed values of the metrics (far outliers removed).

10. Metrics (not) providing very close information



Correlations between the transformed values of the metrics (far outliers removed)

strong monotonic possitive correlations (corr ≥ 0.75) between the transformed values of the metrics (far outliers removed)

metrics providing very close information regarding the performance of each point forecasting method examined

PROPOSED SET
{MAPE, RMSE, NSE, rNSE, cp, ME, MPE, VE, rSD, Pr, r2, d, KGE}

11. Contribution of the present study

Summary

- We conduct a large scale computational experiment based on simulations with the aim to compare the information that 18 metrics provide regarding the performance of point forecasting methods.
- We explore graphically the reltionships between the metrics.
- Subsequently, we build a tool for the comparison of the information provided.
- Finally, we use this tool to decide on a set of metrics for the effective evaluation of point forecasting methods.
- The proposed set is composed by the following 13 metrics:

MAPE, RMSE, NSE, rNSE, cp, ME, MPE, VE, rSD, Pr, r2, d, KGE

Recommendations for further research

- We recommend the analytical investigation of the relationships between the metrics.
- We also recommend the repetition of the experiment of this study using a sufficient number of real-world time series.

References

[1] Criss, R.E. and Winston, W.E., 2008. Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes*, 22, 2723-2725. doi:10.1002/hyp.7072

[2] Gupta, H.V., Kling, H., Yilmaz, KK., and Martinez, G.F., 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

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

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

[5] Fraley, C., Leisch, F., Maechler, M., Reisen, V., and 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

[6] Kitanidis, P.K. and Bras, R.L., 1980. Real time forecasting with a conceptual hydrologic model: 2. Applications and results. *Water Resources Research*, 16 (6), 1034-1044. doi:10.1029/WR016i006p01034

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

[8] Nash, J.E. and Sutcliffe, J.V., 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

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

[10] Tyralis, H., 2016. *HKprocess: Hurst-Kolmogorov Process. R package version 0.0-2*. https://CRAN.R-project.org/package=HKprocess

[11] Tyralis, H. and 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

[12] Yapo, P.O., Gupta, H.V., and 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

[13] 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