Projecting the future of rainfall extremes: better classic than trendy

[1*]Theano Iliopoulou and [1]Demetris Koutsoyiannis

[1]Department of Water Resources, Faculty of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zografou, Greece

* Corresponding author. Tel.: +30 6978580613

E-mail address: tiliopoulou@hydro.ntua.gr

Highlights

- Trend identification is revisited through a model selection framework employing BIC
- Trends, local-, and global-mean models are used to project rainfall indices
- The predictive performance of all models is assessed using 30-year moving windows
- Trends have the worst performance and local-mean models the best

Abstract

Non-stationarity approaches have been increasingly popular in hydrology fitting linear trends in observation periods and evaluating their significance using standard hypothesis testing. Here we reframe the problem of trend identification as a model selection task devising a methodological framework in which trend models are compared to alternative simpler models, namely the local- and the global-mean models, based on their performance in predicting future climatic-scale (30-year) annual maxima, annual totals, wet-day average and probability dry. We evaluate the historical predictive performance of all models through a split-sample calibration-validation technique progressively scanning the whole record by 30-year moving windows. Model selection is based on the BIC, and contrasted to RMSE evaluation. Trends perform worse in all metrics, exhibiting on average the largest prediction errors, the largest variance of errors and the lowest percentage of selection as best model for the future. The worst performance is found for the annual maxima, followed by the probability dry. On the contrary, local-mean models perform consistently well for most cases, especially for persistent indices as the annual totals and probability dry. On these grounds, we argue that multi-model approaches and split-sample validation provide better insights disfavouring trend modelling because of inferior prediction skill compared to simpler approaches.

Keywords: trends, rainfall extremes, rainfall totals, probability dry, climate projections, non-stationarity

1. Introduction

"A trend is a trend is a trend / But the question is, will it bend? /
Will it alter its course / Through some unforeseen force /
And come to a premature end?"

(Sir Alec Cairncross, 1969, signing as "*Stein Age Forecaster*")

In the past decades there has been a plethora of trend analyses in rainfall studies (Bunting et al., 1976; Haylock and Nicholls, 2000, 2000; Rotstayn and Lohmann, 2002; Modarres and da Silva, 2007; Ntegeka and Willems, 2008; Kumar et al., 2010), and it could be argued that relevant studies are still on the rise (4 370 results for the key phrase "rainfall trends" from Google Scholar since 2015; e.g. Biasutti, 2019; Degefu et al., 2019; Folton et al., 2019; Khan et al., 2019; Papalexiou and Montanari, 2019; Quadros et al., 2019; Rahimi and Fatemi, 2019). This boom of trend studies and related results has brought aside it a growing discourse on the appropriate modelling approach. There has been an ongoing debate between stationary vs nonstationary methods, with the former representing a well-established hydrological practice (Montanari and Koutsoyiannis, 2014; Koutsoyiannis and Montanari, 2015) and the latter reflecting the attempts of the scientific community to find a new way to respond to change and uncertainty (Milly et al., 2008; Craig, 2010; Milly et al., 2015), concepts which however are already fully represented in the stationarity framework (Koutsoyiannis and Montanari, 2007; Serinaldi and Kilsby, 2018).

Recently, critiques of trend modelling have been growing on the grounds of empirical evidence (Cohn and Lins, 2005), theoretical consistency (Koutsoyiannis and Montanari, 2015), modelling efficiency (Montanari and Koutsoyiannis, 2014), as well as meaningfulness of the results (Serinaldi et al., 2018). Still, it is currently commonplace to perform trend fitting analyses, mostly with the aim to validate or invalidate the general speculation of intensification of extremes.

In this research, we examine the trend fitting framework from a different perspective, through the evaluation of its modelling qualities and namely, its predictive powers for a record. For this purpose, we introduce a multi-model framework for the evaluation of the results, adding the simpler models of local and global mean in the pool of candidates, and we base the reasoning on the statistical performance of the models evaluated from extensive calibration-validation analysis from multiple climatic-length moving windows of 30 years. While split-sample techniques (Klemeš, 1986) and multi-model approaches (Georgakakos et al., 2004; Duan et al., 2007) are certainly not new in hydrology, in the field of trend modelling these concepts are usually disregarded, with the research question typically revolving around the mere estimation of the statistical significance of a given trend for a specific time window.

Therefore, to escape the various limitations of statistical significance testing for trend modelling (Cohn and Lins, 2005; Serinaldi et al., 2018), we introduce a model selection framework, in which all models' performance in moving windows is simultaneously evaluated by the same criteria, and the linear trend constitutes just one of the candidate models. This analysis seeks to answer the following key questions:

- How well are the rainfall statistics of the most recent climatic period predicted by the candidate models based on the trend calibrated to the prior 30 year period?
- What is the historical all-record predictive performance of linear trend modelling compared to local mean and global mean models and how is this performance affected by the presence of dependence in the rainfall series?

The first question is driven by omnipresent scientific concerns for climate change during the last decades (e.g. Houghton et al., 1991; Parmesan and Yohe, 2003; Oreskes, 2004; Solomon et al., 2007; McCarl et al., 2008; Moss et al., 2010; Craig, 2010; Pachauri et al., 2014; Kellogg, 2019), while the second introduces a methodological framework for assessing model predictions, including ones derived from extrapolating trends, based on objective criteria from model selection theory.

2. Dataset

Our dataset includes the 60 longest available daily rainfall records collected from global datasets, i.e. the Global Historical Climatology Network Daily database (Menne et al., 2012), the European Climate Assessment and Dataset (Klein Tank et al., 2002), as well as third parties listed in the acknowledgments sections. It is an update of the previous dataset explored in Iliopoulou et al. (2018) of long rainfall records surpassing 150 years of daily values. A brief summary of the stations' properties is given in the Appendix (Table A1), while the geographic location of the rain gauges is shown in Figure 1. The length of the timeseries provides insight into rainfall variability and is critical for the investigation of linear trends in multiple time windows and the statistical evaluation of their predictive performance.

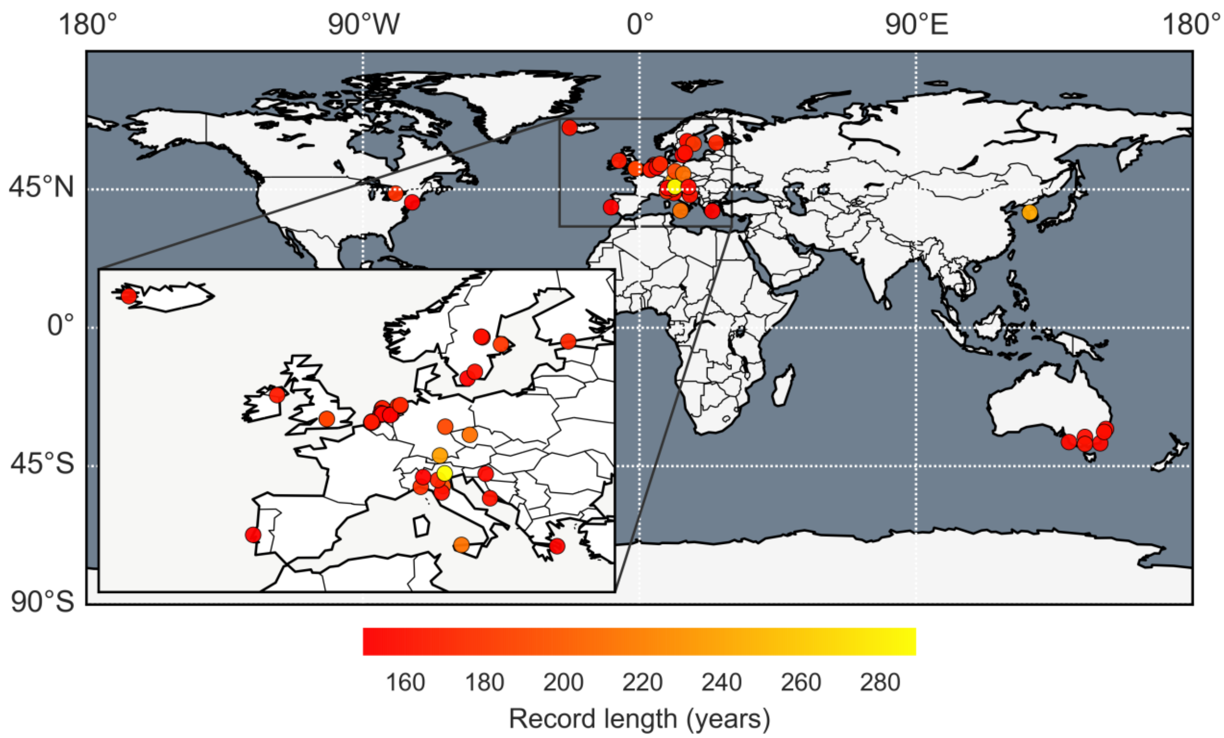

Figure 1. Map of the 60 stations with longest records used in the analysis.

3. Methodological framework

3.1 Selected indices of rainfall extremes and initial quality control

We examine four statistical indices of rainfall: annual maxima (AM), annual totals (AT), wet-day average rainfall (WDAV) and probability dry (PD). As wet, we consider any day with rainfall surpassing the threshold of 1 mm, while we consider values below this threshold as dry days taken into account for the PD estimation. We employ the following criteria for missing values. For the annual maxima we use a methodology proposed by Papalexiou and Koutsoyiannis (2013), according to which an annual maximum in a year with missing values is not accepted if (a) it belongs to the lowest 40% of the annual maxima values and (b) 30% or more of the observations for that year are missing. For the rest of the indices, we do not compute the yearly index in years with more than 15% of missing values. In general, most records have very low percentages of missing values (Table A1), which in most cases are clustered in the beginning of the records and excluded at the second step of the quality control for the trend analysis, which is introduced later in the relevant sections.

3.2 Predictive models: Linear Trend, Local and Global Mean

Let $\underline{x}_i$ be a stationary stochastic process in discrete time i, i.e. a collection of random variables $\underline{x}_i$, and x:={x_1, … x_n} a single realization (observation) of the latter, i.e. a timeseries. We assume that in time $i = O \leq n$ a hypothetical observer makes a forecast based on a subset of the historical information. Namely from the entire available information that we have (the observed sample {x_1,….x_n}) we assume that the hypothetical observer knows only the subset x={x_1,….x_O}. We define the climatic 'present' as the recently experienced 30-year period including the time O, i.e. the period [O-29,O], the climatic 'past' as all the previous years from the beginning of the record [1,O-30], and the climatic 'future' as the following 30 years [O+1,O+29].

To predict the unobserved periods, past or future, we employ three alternative models. The first is the typical linear trend model, encompassing two parameters, a slope and an intercept, fitted via least-squares regression, and the second is a local mean model, including one parameter, the mean of the present climatic period, extrapolated to the past and future periods. We also consider a third option, the global mean model which stands for the estimate of the average from the whole information available at the point O in time. Intentionally, we do not compute the global mean as the actual historic mean, although (assuming stationarity) it does converge to that as the sample grows larger, because we are interested in evaluating the models in realistic conditions in which one has limited information.

As we move the point of the observer in time the new information is used to update the three models, as shown in Fig. 2b. Note that in the global mean model no information is ever lost, only new information keeps adding, while in the other two models the new information is progressively taking the place of the past, which is ‘forgotten’ (progressively discarded), since the model is always fitted on 30 years values.

3.3 Validation schemes

To assess the performance of the linear trend and the two mean models, we formulate two distinct validation schemes, which are illustrated in Fig. 2. In the first (Fig.2a) we evaluate the models’ performance in explaining the variability of the recent 30 year period of each station based on calibration on the prior 30 year period. By this ‘static validation’ scheme we also intend to evaluate whether extremes have changed in a consistent manner in the second half of the 20th century, as they are commonly assumed, and we also extend this examination to the past data. In order to maximize the exploitation of the length of each record, we apply this evaluation to the most recent period of each station, even if the final dates of all records do not coincide. We favour separate treatment of each station, based on its record length, over a common (and arbitrary) trend evaluation window for all stations, since our focus is placed on the operational exploitation of records for predictive purposes and less on a summary of the trend results for a specific time period. In this regard, the climate change postulation drives the exploratory assessment of changes in the most recent periods, but it does not dictate the choice of a common evaluation window, for which there is no empirical support. However, the majority of the records span the whole 20th century, and extend beyond, with a few exceptions that are mentioned.

The second scheme (Fig.2b) takes a different perspective and examines the historical performance of the models assuming a hypothetical observer moving in time and making predictions as new information becomes available. We fit all three models on 30-year moving-window calibration following the procedure described in the previous section (Section 3.2). The first moving-window starts from the start of each station while the last moving-window is the last period for which there are available 30 years of validation data, i.e. 60 years prior to the end of the station, which coincides with the 30-year window that was exploited in the previous scheme.

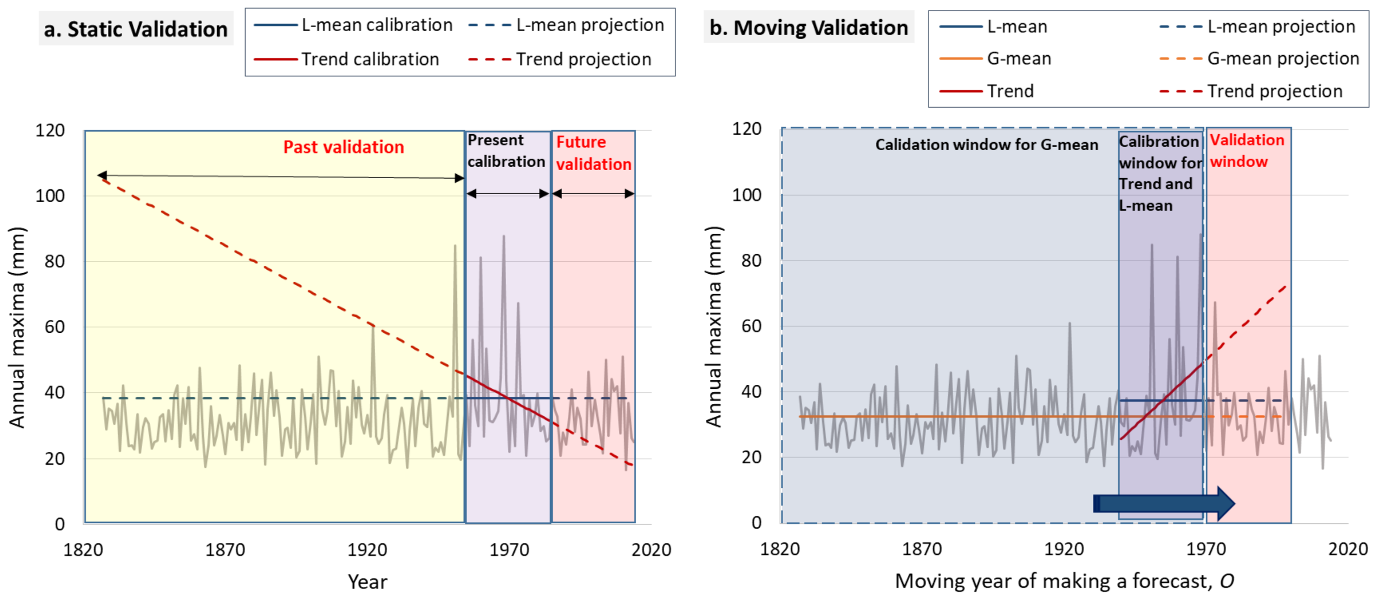

Figure 2. Explanatory sketch showing the two validation schemes (a. Static Validation and b. Moving Validation) for an example station.

Because the earlier periods of some records are affected by missing values we introduce further criteria for this step. To ensure consistency of measurements and minimize the impact of consecutive missing values in trend estimation, we only analyse periods with less than 5% of consecutive missing values of the yearly indices. Accordingly, we estimate the linear trend in a time window only if there exist at least 27 valid indices in each of the 30-year periods of calibration and validation and we compute the summary statistics only for stations having at least at least 60 time windows of valid information.

3.4 Beyond statistical significance to model selection criteria for trend evaluation

Statistical significance is now considered a poor and outdated scientific method for model evaluation and strong critiques against its blind use are increasingly communicated by the statistician community itself, part of it even calling for the abandonment of the concept entirely (Nuzzo, 2014; Wasserstein and Lazar, 2016; Amrhein and Greenland, 2018; Wasserstein et al., 2019; Amrhein et al., 2019). The American Statistical Association concludes that "*the widespread use of 'statistical significance' (generally interpreted as 'p ≤ 0.05') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process*" (Wasserstein and Lazar, 2016), while a relevant criticism of the concept of statistical significance dates back to Akaike's work (Akaike, 1969)*: "The main difficulty in applying this kind of procedures stems from the fact that they are essentially formulated in the form of a successive test of the whiteness of the series against multiple "alternatives." Actually one of the "alternatives" is just the model we are looking for and thus it is very difficult for us to get the feeling of the possible alternatives to set reasonable significance levels."*

Akaike has contributed to the introduction of information theory into model selection criteria (Akaike, 1974) which are prevailing worldwide in model inference (Anderson and Burnham, 2004) and are increasingly adopted in hydrology as well (e.g. Ye et al., 2008; Laio et al., 2009; Iliopoulou et al., 2018a). Yet the concept of statistical significance in the field of trend studies is still widely applied, although its misuse in hydrology has been further emphasized by seminal studies (e.g. Cohn and Lins, 2005; Koutsoyiannis and Montanari, 2007; Serinaldi et al., 2018) which have established the fact that for hydrological, non i.i.d. data the null hypothesis, which tacitly contains independence, is a priori wrong, and its rejection, if correctly interpreted, should point out to the wrong independence assumption. However, the common practice has been to misleadingly favour trends. For this reason we do not employ statistical hypothesis testing for the fitted trends and instead, we resort to model selection theory, regarding trends as models.

For the evaluation of the candidate models we consider two different criteria, the Root Mean Square Error, a conventional, standard metric of goodness of fit, applied here for illustration of its pitfalls and comparison purposes, and the Bayesian Information Criterion (Schwarz, 1978), an information theoretic approach, widely applied in various model selection problems, because it is capable of balancing the modelling pursuit of parsimony and goodness of fit (Burnham and Anderson, 2004). In both cases, the preferred model is the one yielding the lowest value of the criterion. The RMSE is defined as the square root of the mean square error of the predicted values $\hat{x}_i$ with respect to the observed x_i:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{n}} \tag{1}$$

while the Bayesian Information Criterion for linear regression is defined as (Priestley, 1981):

$$\mathrm{BIC} = n\ln(\mathrm{RSS}/n) + k\ln(n) \tag{2}$$

where n is the length of the data, k the number of the parameters and RSS stands for the residual sum of squares, $\mathrm{RSS} = \sum_{i=1}^{n}(\hat{x}_i - x_i)^2$. We note that the alternative well-established and closely related criterion, AIC is defined for linear regression as $\mathrm{AIC} = n\ln(RSS/n) + 2k$. The difference between the two is the manner by which the extra model parameters are penalized. In BIC this is done by the multiplication of the parameters by the term ln(n), while in AIC, simply by the number 2. This implies that asymptotically, as the number of observations increases, the goodness of fit term dominates over the parameterization term for both criteria, but at a much quicker rate for AIC. In essence for large n as in our case (when examining the past), AIC reduces to an evaluation of RSS. For this reason, we employ the BIC instead, that converges less quickly to the RSS evaluation.

For the application in future periods for which the models are extrapolated and not fitted, we apply the criteria under the hypothesis that the models represent the future period, i.e. both the RSS and the number of data are computed from the future record, even if the model is calibrated to the previous period. We stress that BIC is a measure of relative performance of the models and does not assume that the 'true' model is included in the set of alternatives (Ye et al., 2008), if one naively assumes that there exist a "true" model when dealing with complex natural phenomena. Therefore it is justified to use it as a criterion for predictive quality. Obviously, the direct comparison of the absolute BIC values for the past, present and future periods is not meaningful since these are obtained from different data, however the comparison the outcomes of the model selection for these periods is allowed.

3.5 Statistics of model performance

For the periods of calibration ('present') and validation ('future'), we estimate the following four performance measures for the three models: present RMSE and present BIC, future RMSE and future BIC. We evaluate the models' performance based on two different aspects; the quality of future performance, irrespective of their present performance, and the quality of future performance conditional on the 'present' performance being superior. Both qualities are judged by the outcome of the model selection process, i.e. the future performance is evaluated by the frequency (%) of future time windows in which each model outperforms the other two, while the conditional performance is obtained by the frequency (%) of present time windows that fulfil the following condition: *the presently chosen model is also found superior in the future*. Clearly, the second evaluation aims at examining the degree of evidence available at present for the selection of each model in the future, while the first evaluation considers only the frequency of good future results.

For the first type of the evaluation, the historical ‘future’ performance of the three models is also tracked by computing the following metrics: a) the average RMSE and BIC of each model computed from all future time windows, and b) the standard deviation of the RMSE computed from all future time windows. As a further insight into model performance, we assess the temporal properties of the propagation of the prediction error as well as the BIC value for specific case studies.

For the second evaluation, in the cases, where a model is best for the present but is eventually outperformed by another one in the future, we also report the % of time windows for the preferred alternative models. This evaluation provides insights into the consistency of the models’ performance, but also allows a critical comparison of the two evaluation settings. For instance, we note that by usual practice a model would not be considered for future application if its performance on the present was found inferior. By this rationale, global mean models are unlikely to be chosen from this set of candidates, as their performance in the present is always (by construction) worse than that of the local mean model. However, their future performance often proves superior to the other’s two, as shown next.

3.6 Effect of dependence on model selection

To evaluate how the presence of dependence in the data affects the results, we repeat all the model evaluations for the randomized counter-series of the original records. The latter are produced by randomly resampling (shuffling) the original data, in order to destroy the temporal dependence structure, while fully preserving the marginal properties. Accordingly, we produce 500 shuffled timeseries for each original record. Because in shuffling missing data present in the older time periods of the dataset contaminate the whole record with their random spread, for certain stations more iterations are needed since some shuffled timeseries are discarded due to the temporal distribution of missing values. From the sample of shuffled timeseries for each station, we compute the average metrics for the station, to which we compare the original metrics. We do not perform the random resampling for the stations that did not pass the missing value control at the preceding step of the analysis.

This examination scrutinizes two key parameters of the results: a) the chance of obtaining a trend by the two criteria under iid conditions, i.e. the false trend discovery rate by each criterion, and b) the effect of dependence in the historical data on the predictive relative performance of each model, as well as on the overall preference for a predictive model.

4. Results

4.1 Trend models performance in static validation

Results from the performance on trend modelling on the last 30 years of each station are shown in Figures 3-6 for all studied indices and are further summarized in Table 1. It can be seen that trends have performed poorly, compared to the local mean model, in explaining the most recent changes of extremes and even more poorly with the respect to capturing the past. Interestingly, the worst performance —both in future and past validation periods, is encountered in the annual maxima, followed by probability dry. For the vast majority of cases the local mean model, fitted with only one parameter, performs better. A visual examination of the 60 stations plots clearly suggests that climatic trends quickly reverse (thus suggesting a positive answer to the question in the motto…), while their extrapolation to past and future periods may tend to physically inconsistent results, e.g. negative amounts of rainfall. Apparently, the only cases where trends show a better performance are the ones where the fitted slope is very mild, thus not differing substantially from the local mean.

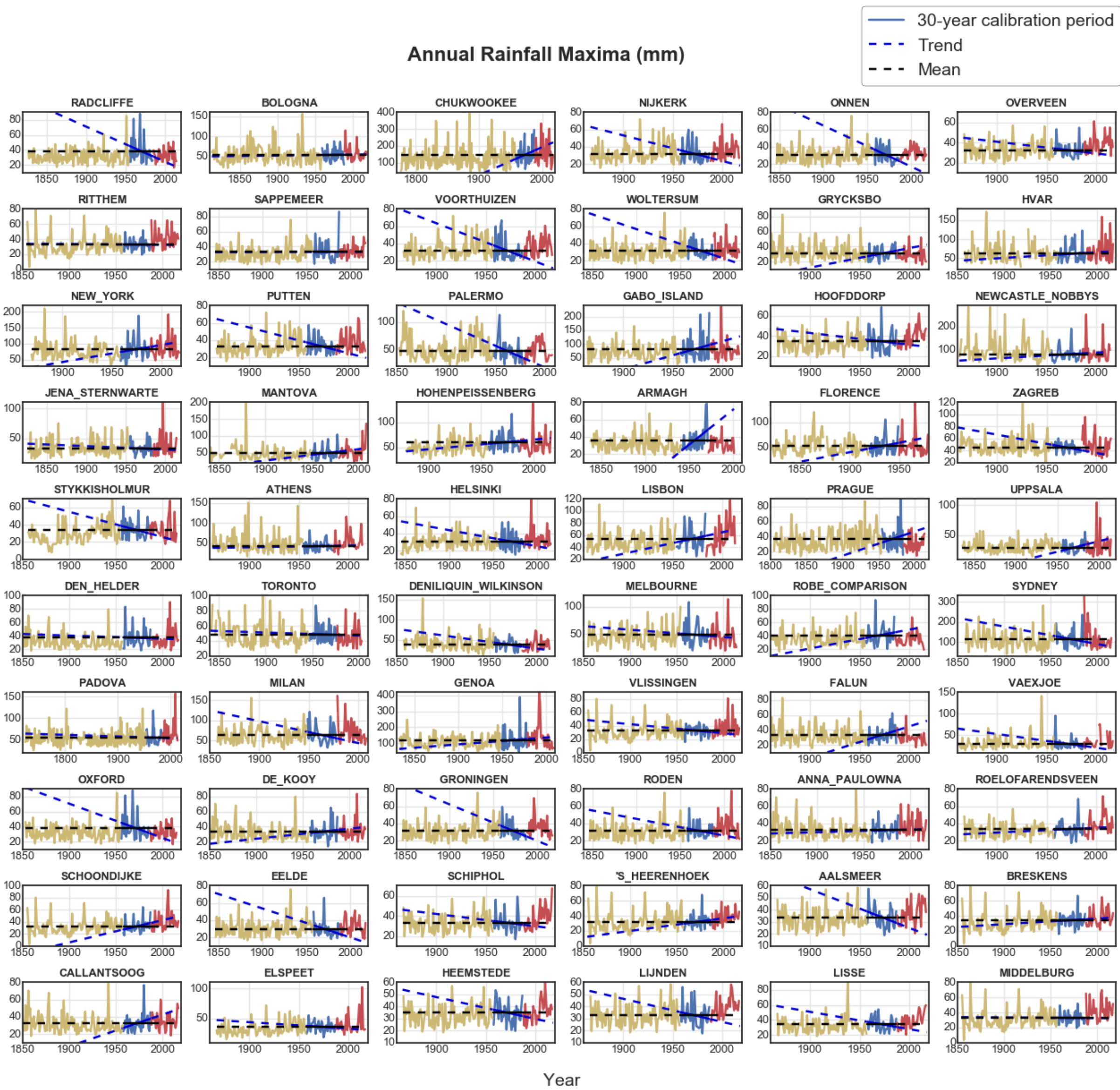

Figure 3. Trends vs the local mean in projecting annual maxima for the 60 longest rainfall stations.

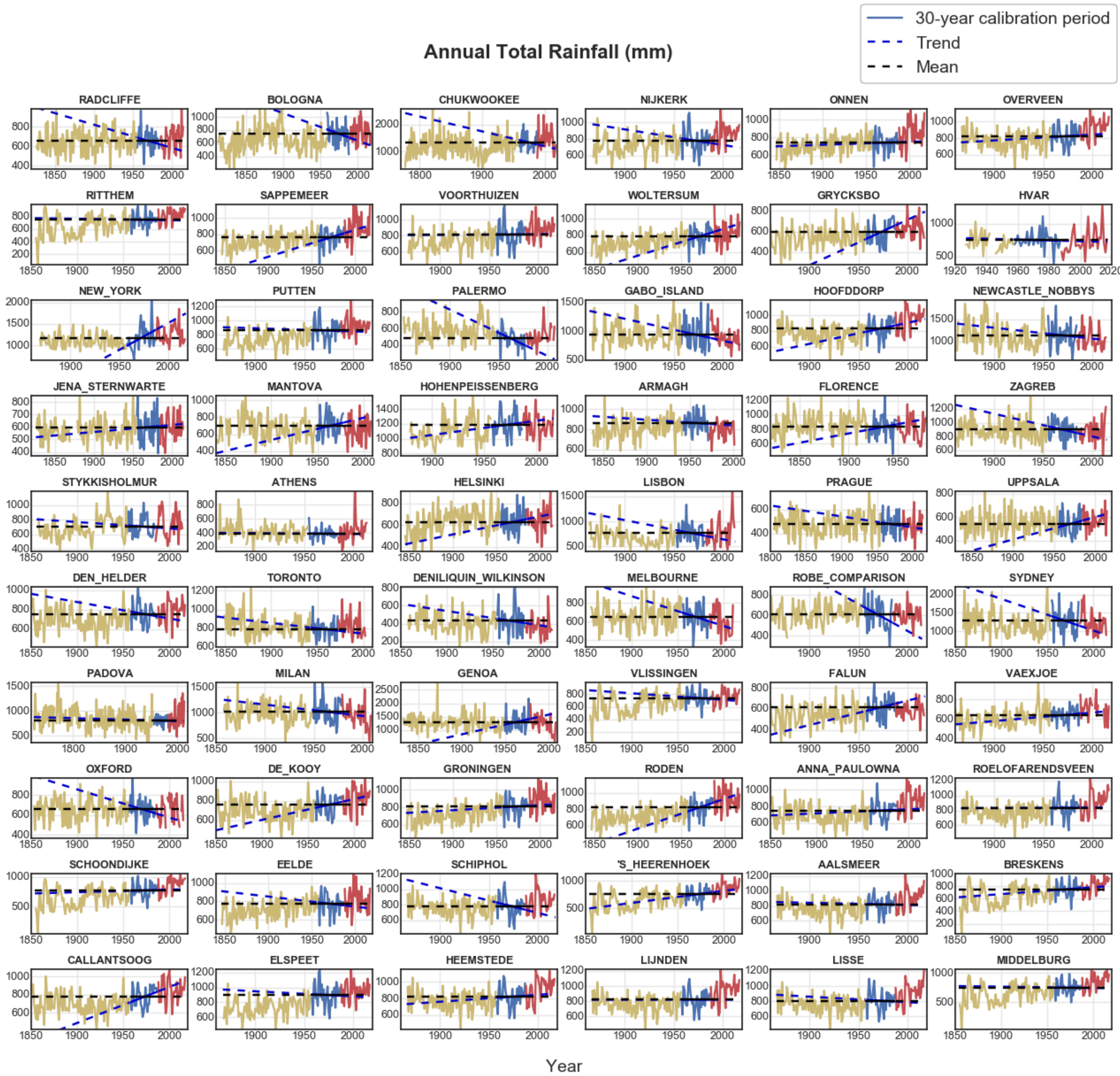

Figure 4. Trends vs the local mean in projecting annual totals for the 60 longest rainfall stations.

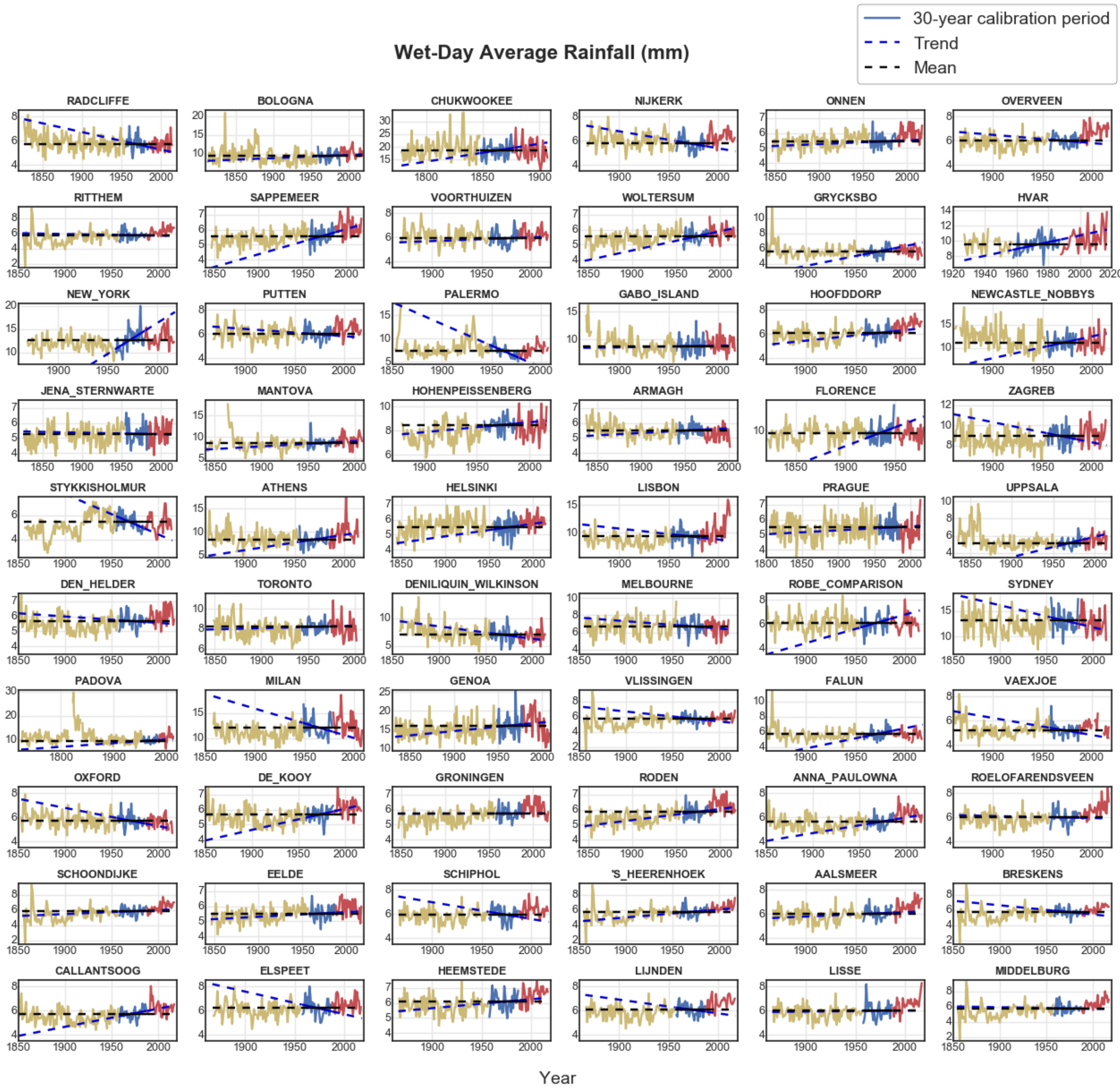

Figure 5. Trends vs the local mean in projecting wet-day average rainfall for the 60 longest rainfall stations.

Figure 6. Trends vs the local mean in projecting probability dry for the 60 longest rainfall stations.

Table 1 Results from first validation setting (static validation) considering the more recent 30-year period of each station as 'future validation' calibrated to its prior 30-year period, and the whole past record as 'validation in the past'. The numbers report the sum of the stations (total of 60) in which the Local Mean model outperforms the Linear Trend Model by each criterion.

	30-year Calibration		30-year Future validation		Past validation	
Index	By BIC	By RMSE	By BIC	By RMSE	By BIC	By RMSE
AM	52	0	56	45	58	57
AT	57	0	44	35	49	43
WDAV	52	0	42	35	51	49
PD	57	0	50	42	49	46

4.2 Moving-window assessment of predictive performance

In this step, we further explore the predictive qualities of the models based on the statistical analysis of the whole record, by adding in the pool of candidates the third alternative model, the global mean. Figures 7-10 show the distribution of the average error of each model resulting from the moving window application described in Section 3.3, both for the original and the shuffled data. It is evident that there is a marked inferior performance of the linear trend, which naturally is more pronounced for the shuffled data, where in reality there is no trend signal. The other two models perform similarly, with the local mean model however clearly prevailing in the cases of annual totals and probability dry (Fig. 8, 10). Interestingly, a preference for the local mean model is not identified in the shuffled data. An important insight is derived when the standard deviation of the RMSE is examined from the whole record moving window application of the three models (Fig. 11). In this case, it is evident that for the original data the local mean model prevails in all cases to the global mean model, which however is in turn preferred when the data are shuffled. The direct explanation is that there is dependence in time, which is destroyed by shuffling. A user unfamiliar with dependence may misinterpret it as a trend and perhaps this explains why trend claims have been so common lately. Expectedly, the linear trend model shows a much worse performance in the variance examination than revealed by the average RMSE evaluation. The average values of the metrics are summarized in Table 2.

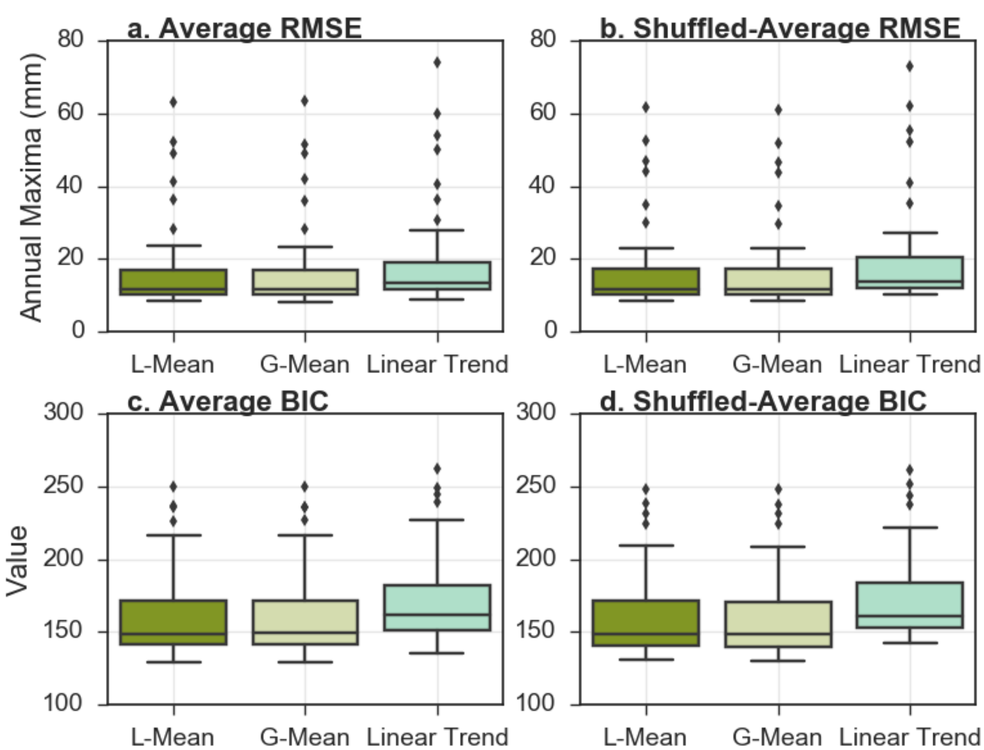

Figure 7. Average RMSE and BIC value for the local (L-) mean, global (G-) mean and Linear Trend models applied for the annual maxima prediction of both the original and the shuffled data. The band inside the box reports the median of the distribution, the lower and upper ends of the box represent the 1st and 3rd quartiles, respectively, and the whiskers extend to the most extreme value within 1.5 IQR (interquartile range) from the box ends; outliers are plotted as points.

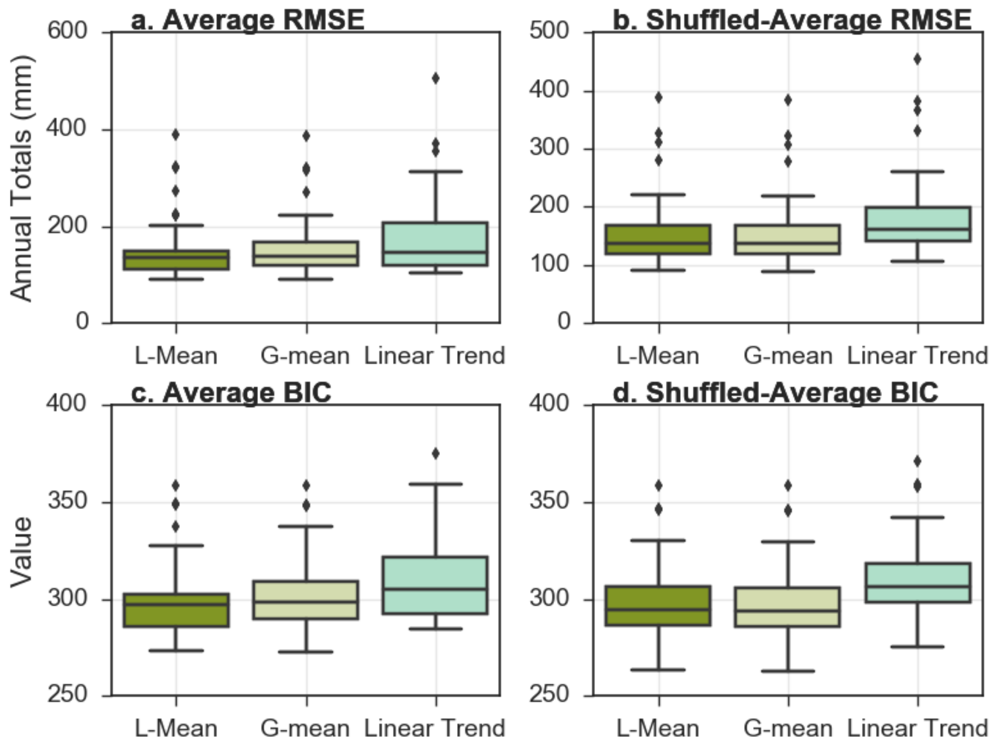

Figure 8. Average RMSE and BIC value for the local (L-) mean, global (G-) mean and Linear Trend models applied for the annual totals prediction of both the original and the shuffled data. The band inside the box reports the median of the distribution, the lower and upper ends of the box represent the 1st and 3rd quartiles, respectively, and the whiskers extend to the most extreme value within 1.5 IQR (interquartile range) from the box ends; outliers are plotted as points.

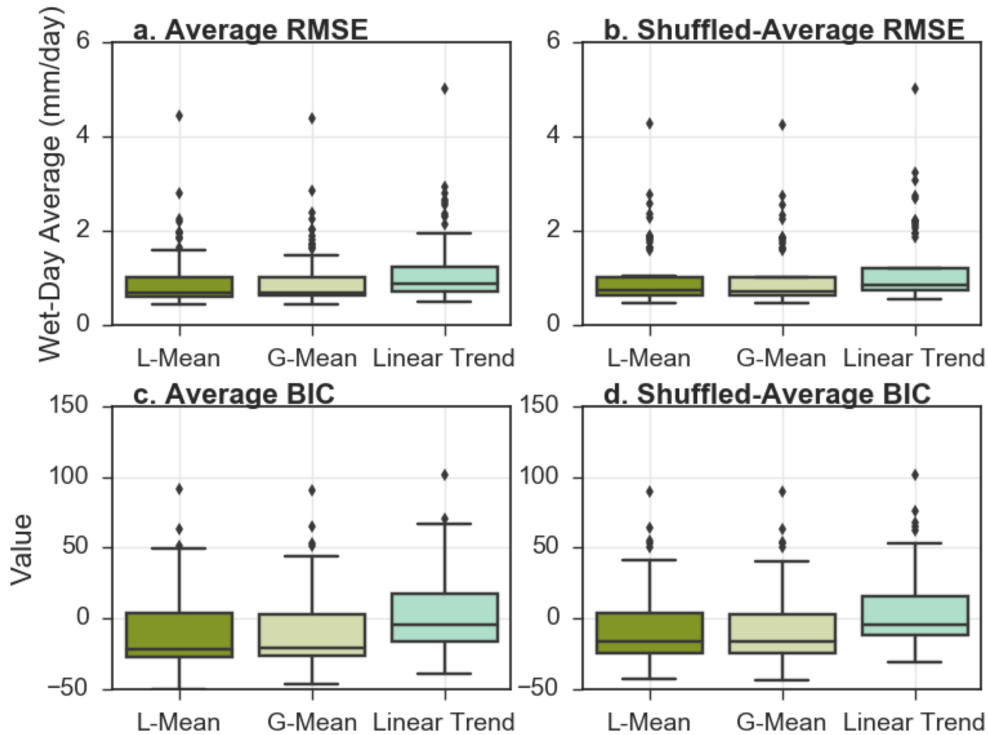

Figure 9. Average RMSE and BIC value for the local (L-) mean, global (G-) mean and Linear Trend models applied for the wet-day average prediction of both the original and the shuffled data. The band inside the box reports the median of the distribution, the lower and upper ends of the box represent the 1st and 3rd quartiles, respectively, and the whiskers extend to the most extreme value within 1.5 IQR (interquartile range) from the box ends; outliers are plotted as points.

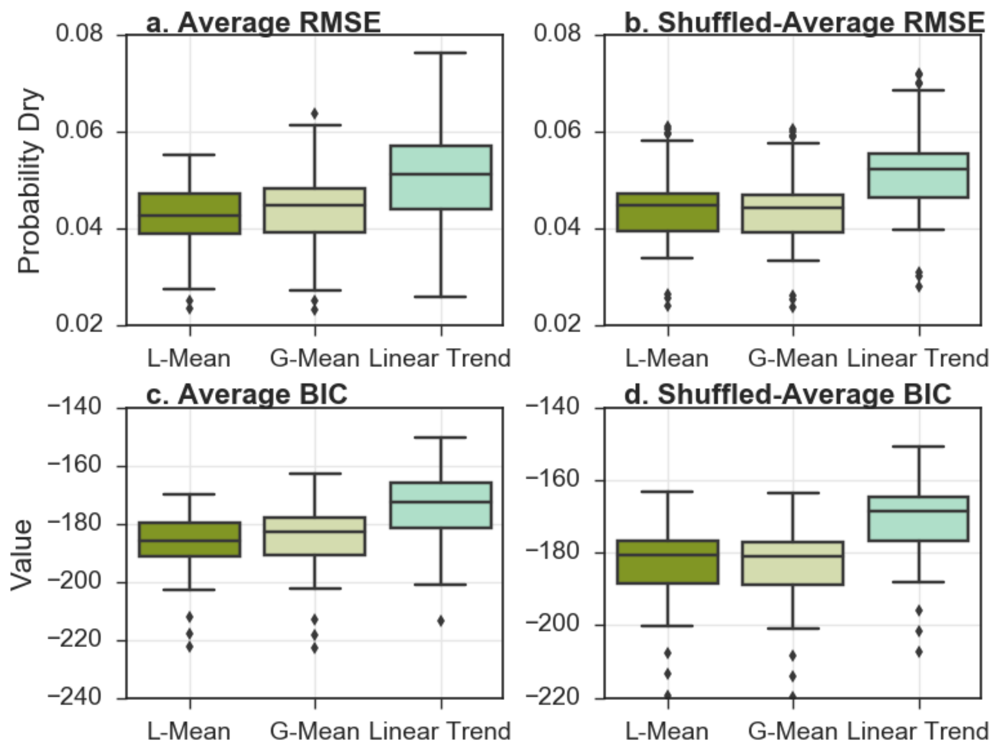

Figure 10. Average RMSE and BIC value for the local (L-) mean, global (G-) mean and Linear Trend models applied for the probability dry prediction of both the original and the shuffled data. The band inside the box reports the median of the distribution, the lower and upper ends of the box represent the 1st and 3rd quartiles, respectively, and the whiskers extend to the most extreme value within 1.5 IQR (interquartile range) from the box ends; outliers are plotted as points.

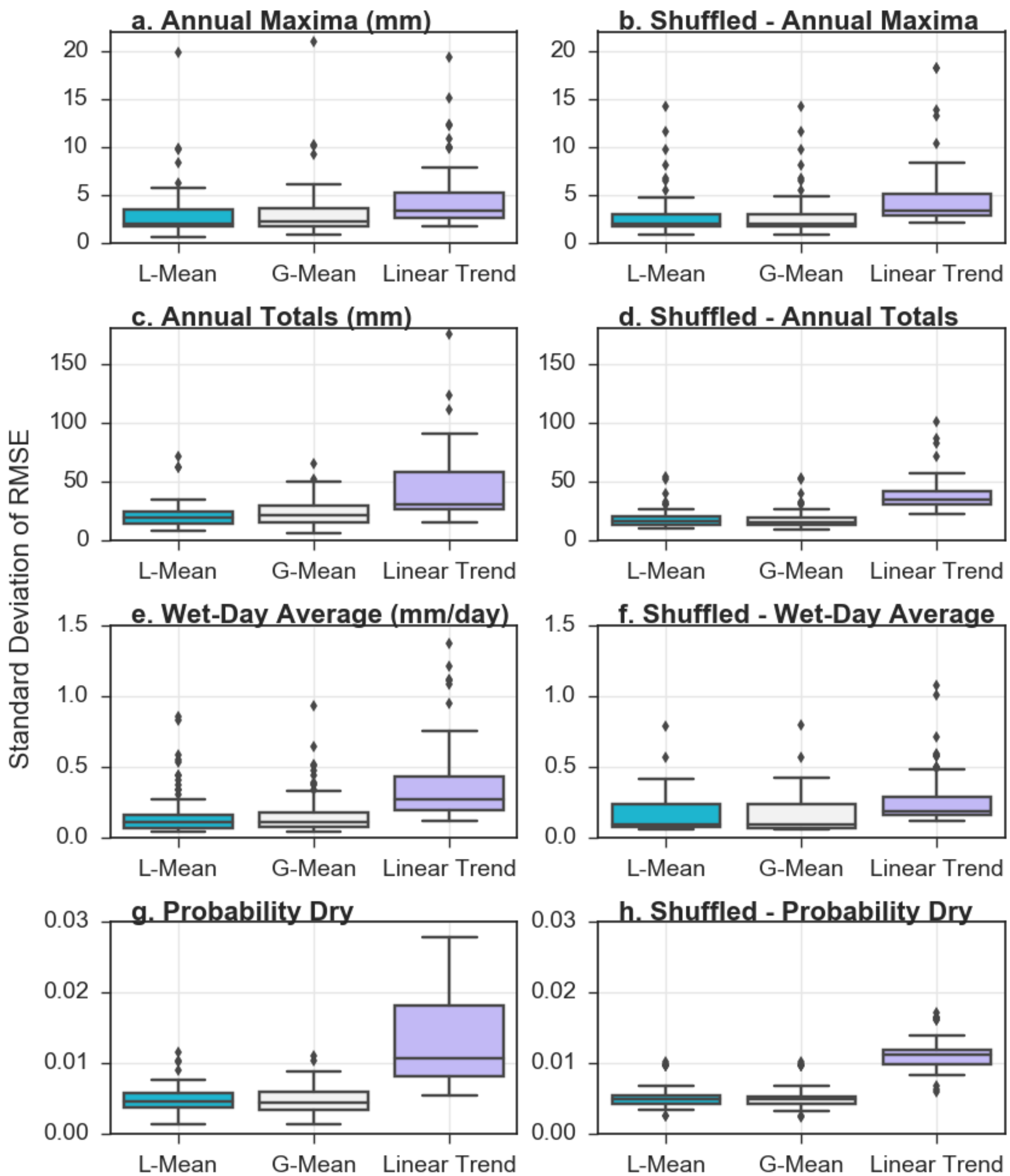

Figure 11. Standard deviation of RMSE for the local (L-) mean, global (G-) mean and Linear Trend models applied for all the indices of both the original and the shuffled data. The band inside the box reports the median of the distribution, the lower and upper ends of the box represent the 1st and 3rd quartiles, respectively, and the whiskers extend to the most extreme value within 1.5 IQR (interquartile range) from the box ends; outliers are plotted as points.

From the historical evaluation of the models’ performance, described in Section 3.6, it is at first evident that possible differences in the marginal distributions among stations have no impact on the selected models for the shuffled data. Namely, the percentages of model selection in the future for the shuffled data are approximately stable also for all indices, revolving around the values of 40% for the L-Mean model, 58% for the G-Mean and 2% for the linear trend according to the BIC, as seen in Table 2. For the random data, the false trend discovery rate by RMSE rises to an average of 15%. The worst performance of trend modelling is again identified for the annual maxima, selected on average by BIC only 3.16% of the future windows (Table 2), while the local mean model clearly dominates on average in the rest three cases of the annual totals (54.57%), wet-day average (53.8%) and probability dry (58.71%). Percentages however considerably fluctuate among the different stations (Fig.12-15), obviously as a result of their different temporal dependence properties, in contrast to the almost stable results obtained from their shuffled counter-series. Qualitatively similar model selection results are obtained from the RMSE evaluation, yet with the marked difference of the percentages of the trend model being inflated, both for the original and (erroneously) for the shuffled data (15%), accompanied by an analogous decrease in the rate of selection of the mean models.

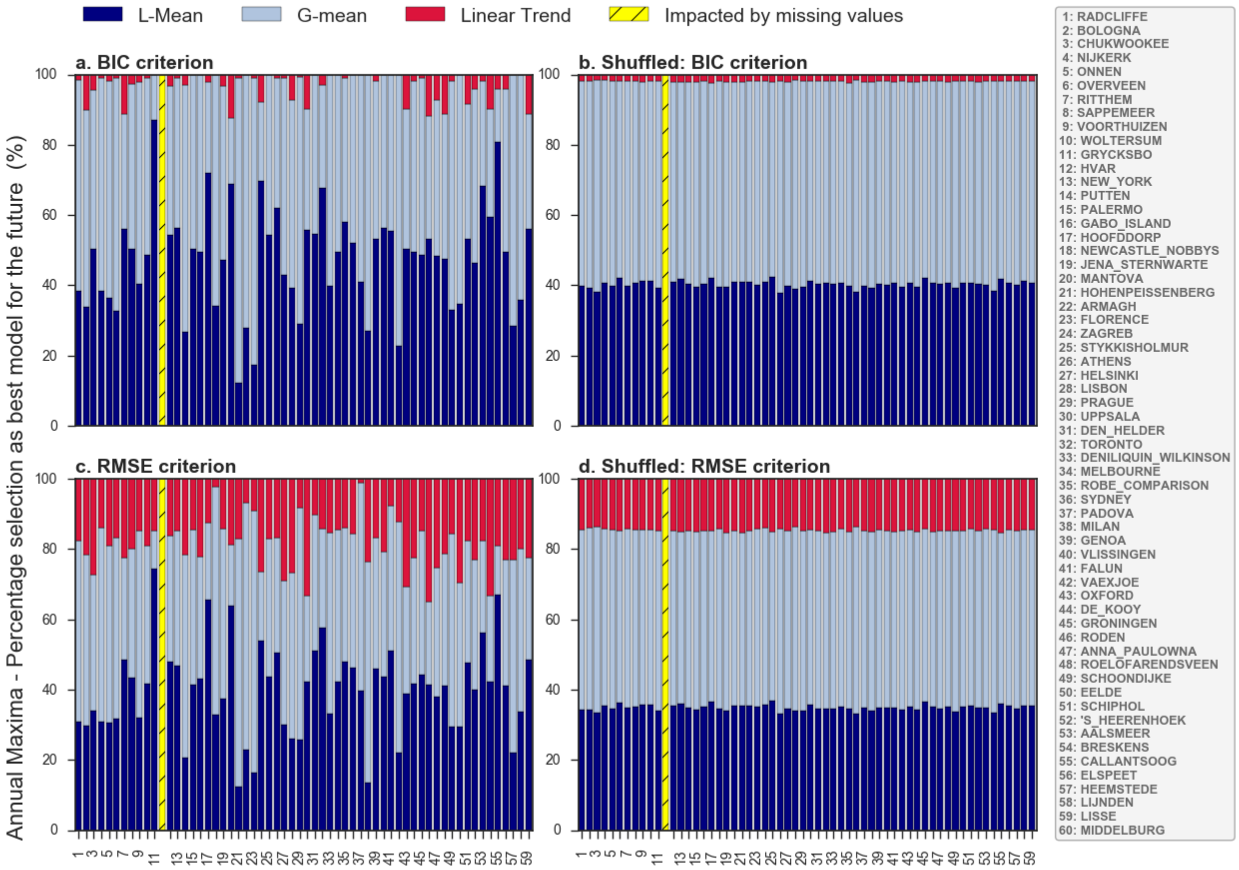

Figure 12. Percentage of selection as best model for the future based on RMSE and BIC criteria for annual maxima of both the original and the shuffled data.

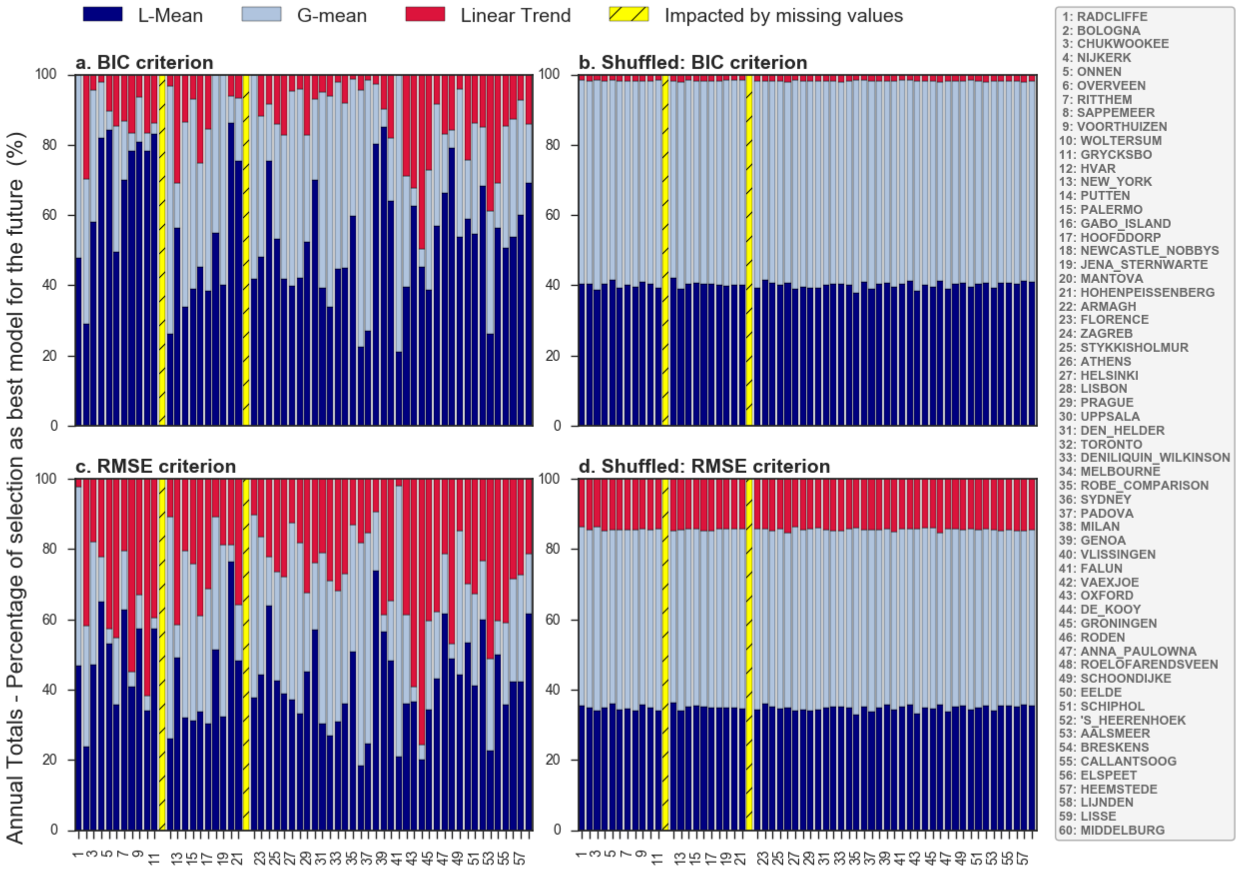

Figure 13. Percentage of selection as best model for the future based on RMSE and BIC criteria for annual totals of both the original and the shuffled data.

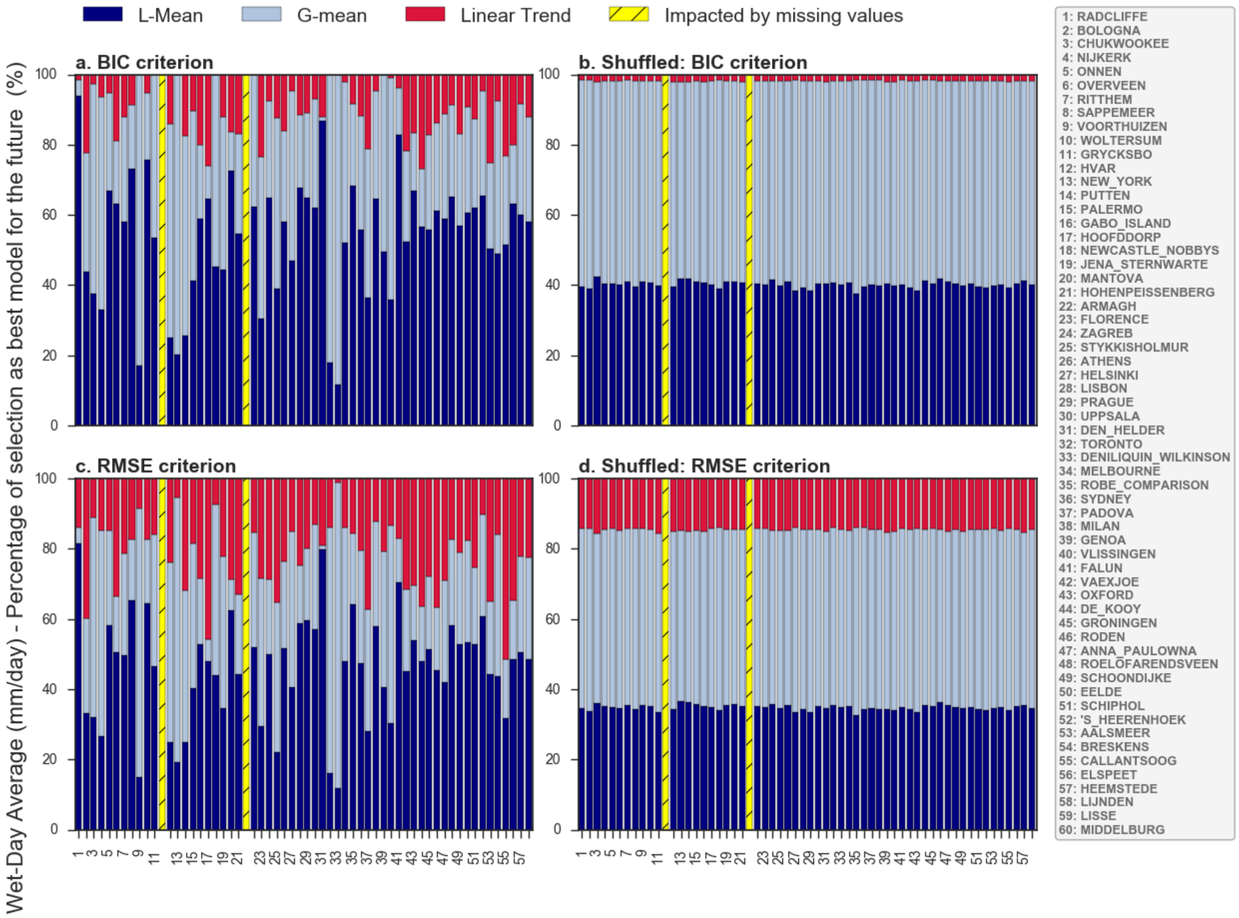

Figure 14. Percentage of selection as best model for the future based on RMSE and BIC criteria for wet-day average of both the original and the shuffled data.

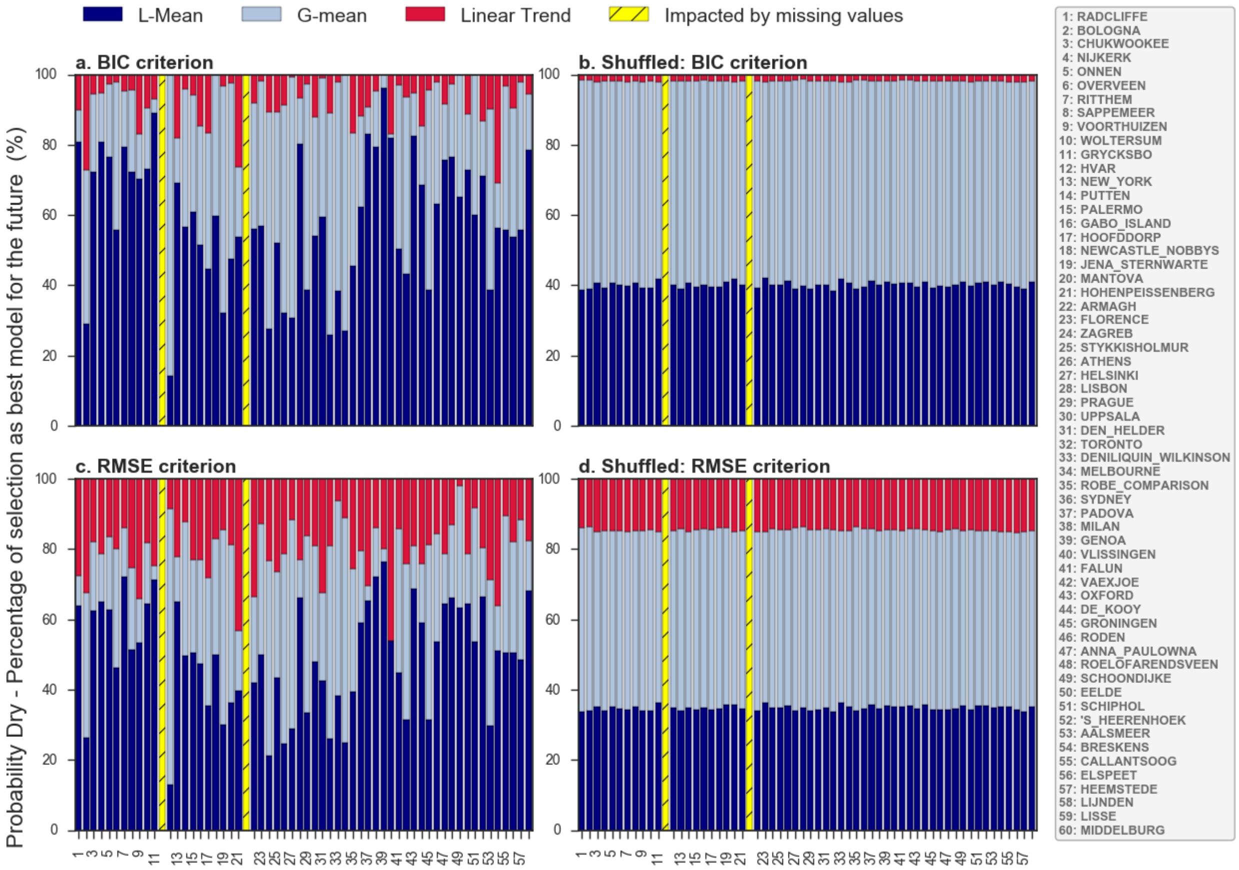

Figure 15. Percentage of selection as best model for the future based on RMSE and BIC criteria for probability dry of both the original and the shuffled data.

Table 2 Average of the metrics of the three models (local (L-) mean, global (G-) mean and Linear Trend) from all stations and for all four indices. Values from shuffling are reported in parentheses in red color.

	Annual Maxima			**Annual Totals**		
	L-mean	G-mean	Trend	L-mean	G-mean	Trend
Average RMSE	16.00 (16.11)	15.95 (15.95)	18.76 (18.99)	149.07 (153.8)	154.27 (152.38)	174.7 (180.57)
St. Deviation RMSE	17.83 (15.55)	19.11 (15.6)	34.80 (34.27)	24.79 (20.42)	26.13 (20.18)	55.08 (41.9)
% selection by RMSE	39.78 (34.9)	41.62 (50.43)	18.6 (14.67)	42.81 (34.8)	27.94 (50.76)	29.25 (14.44)
% selection by BIC	47.53 (40.34)	49.31 (57.72)	3.16 (1.94)	54.57 (40.15)	32.79 (58.04)	12.64 (1.81)
% conditional selection by BIC	47.56 (40.07)	-	0.07 (0.13)	54.15 (39.88)	-	3.71 (0.09)
	Wet-Day Average			**Probability Dry**		
	L-mean	G-mean	Trend	L-mean	G-mean	Trend
Average RMSE	0.98 (1.05)	1 (1.04)	1.2 (1.24)	0.04 (0.04)	0.04 (0.04)	0.05 (0.05)
St. Deviation RMSE	0.26 (0.21)	0.25 (0.21)	0.49 (0.35)	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)
% selection by RMSE	45.9 (34.76)	31.53 (50.64)	22.57 (14.6)	49.6 (34.74)	29.67 (50.66)	20.73 (14.6)
% selection by BIC	53.8 (40.22)	35.32 (57.94)	10.88 (1.84)	58.71 (40.16)	33.46 (57.99)	7.83 (1.85)
% conditional selection by BIC	53.43 (39.95)	-	3.23 (0.09)	58.63 (39.89)	-	2.62 (0.08)

4.3 Conditional evaluation of predictive performance

In this step, we place our focus on tracking the future performance of the models that are originally chosen as optimal for the present, following the rationale explained in Section 3.5. We also report the alternative outcomes in case a model is outperformed by its alternatives in the future, albeit being superior in the present. Obviously, we can perform this evaluation only for the linear trend model and the L-mean model since the global mean model is never preferred to the other two in present time. Yet, as it can be seen it is very often selected as the best model for the future. We do not report the results by the RMSE evaluation, because these are identical to the results previously reported. The reason is that by the RMSE evaluation, the trend model is always selected as best for the present —because of its two parameters, and therefore the only unknown quantity for the estimation of the conditional percentage is the future evaluation by RMSE, already reported in the previous figures.

Results however from the BIC evaluation allow the objective comparison of both metrics in conditional mode. The latter are striking in terms of the very poor future performance of the trend model in the cases when it is found appropriate for the present. For the annual maxima, the trends performance almost universally deteriorates in the future (Fig. 16a), which is also true for the rest of the indices (Fig. 17a-19a) with just a few exceptions. On the contrary, the L-mean model shows a considerably robust performance in the future, as in the majority of cases continues to be the best model, being outperformed in some cases only from the emergence of the global mean model as a better choice, namely for the annual maxima (Fig. 20a). Because by BIC evaluation, both in the present and in the future, the L-mean dominates, its average percentage is very close to its future percentage (Table 2), although not identical as in the case of trends evaluated by RMSE. Less expected however, is the future deterioration of the trend's performance in the cases when it is selected in the present. The few time-windows in which the trend model performs relatively well in the future (e.g. Fig. 13a for annual totals) do not in general correspond to the present windows in which it also performs well (e.g. Fig. 17a), which explains the difference in the two statistics.

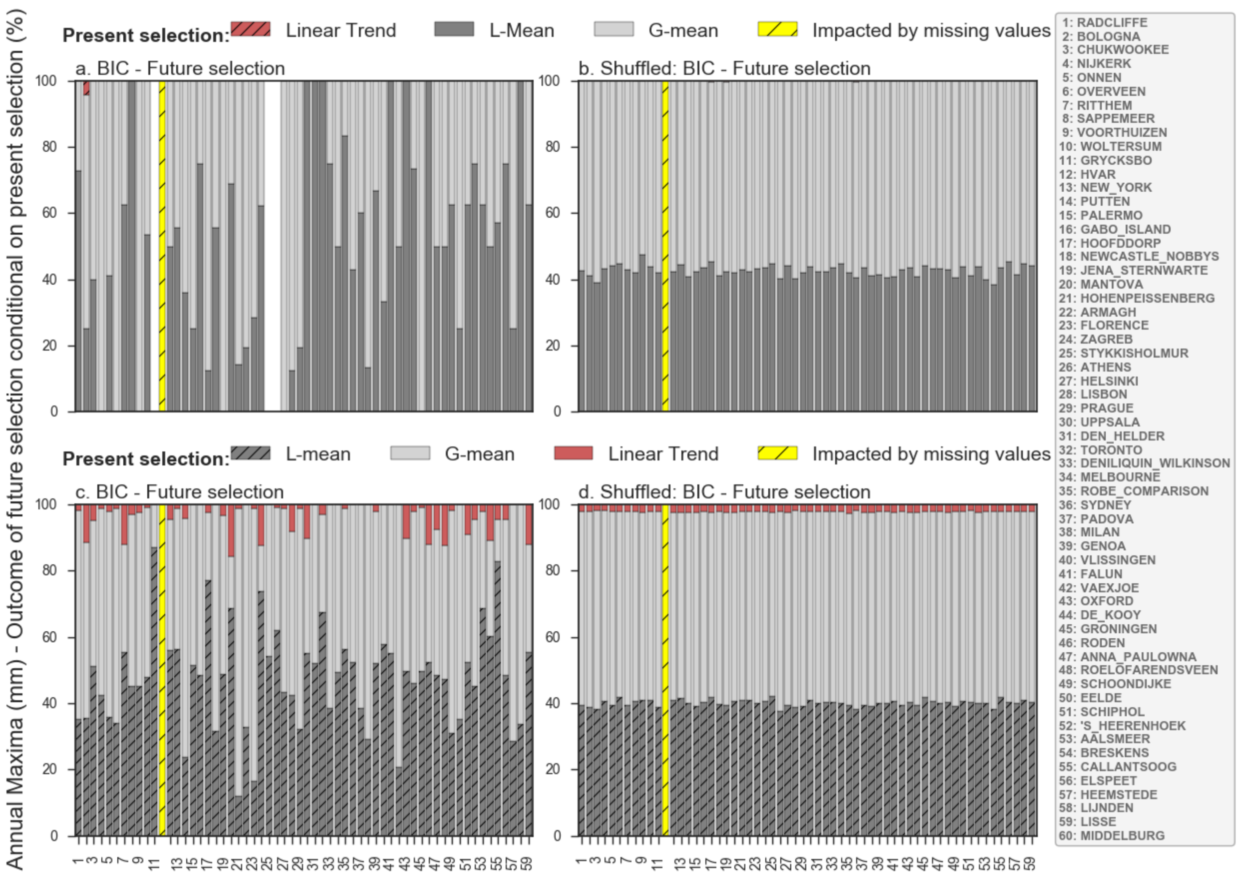

Figure 16. Percentage of future performance of each model conditional on its selection on the present based on RMSE and BIC evaluation for annual maxima of both the original and the shuffled data. White straps correspond to cases of zero selections for the present.

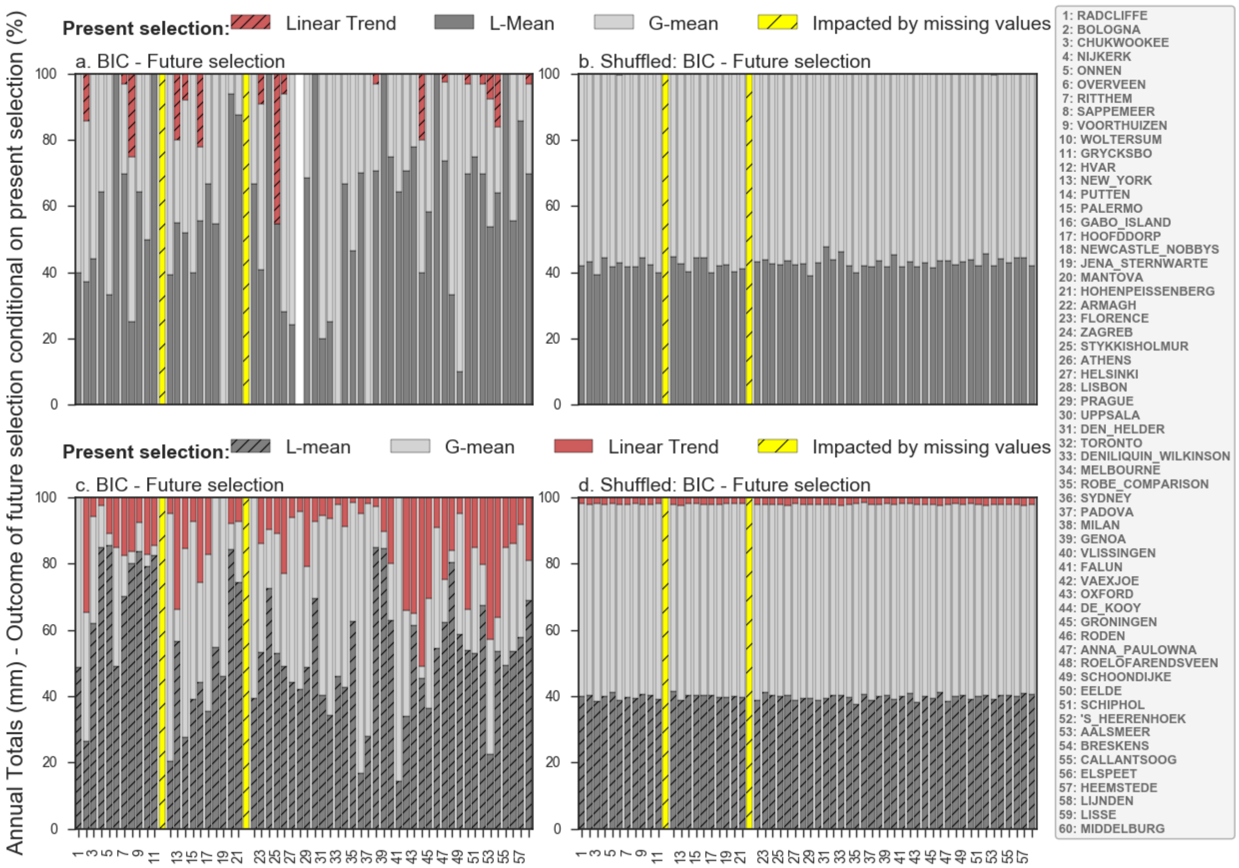

Figure 17. Percentage of selection as best model for the future for each model conditional on its selection on the present based on RMSE and BIC evaluation for annual totals of both the original and the shuffled data. White straps correspond to cases of 0 present selection.

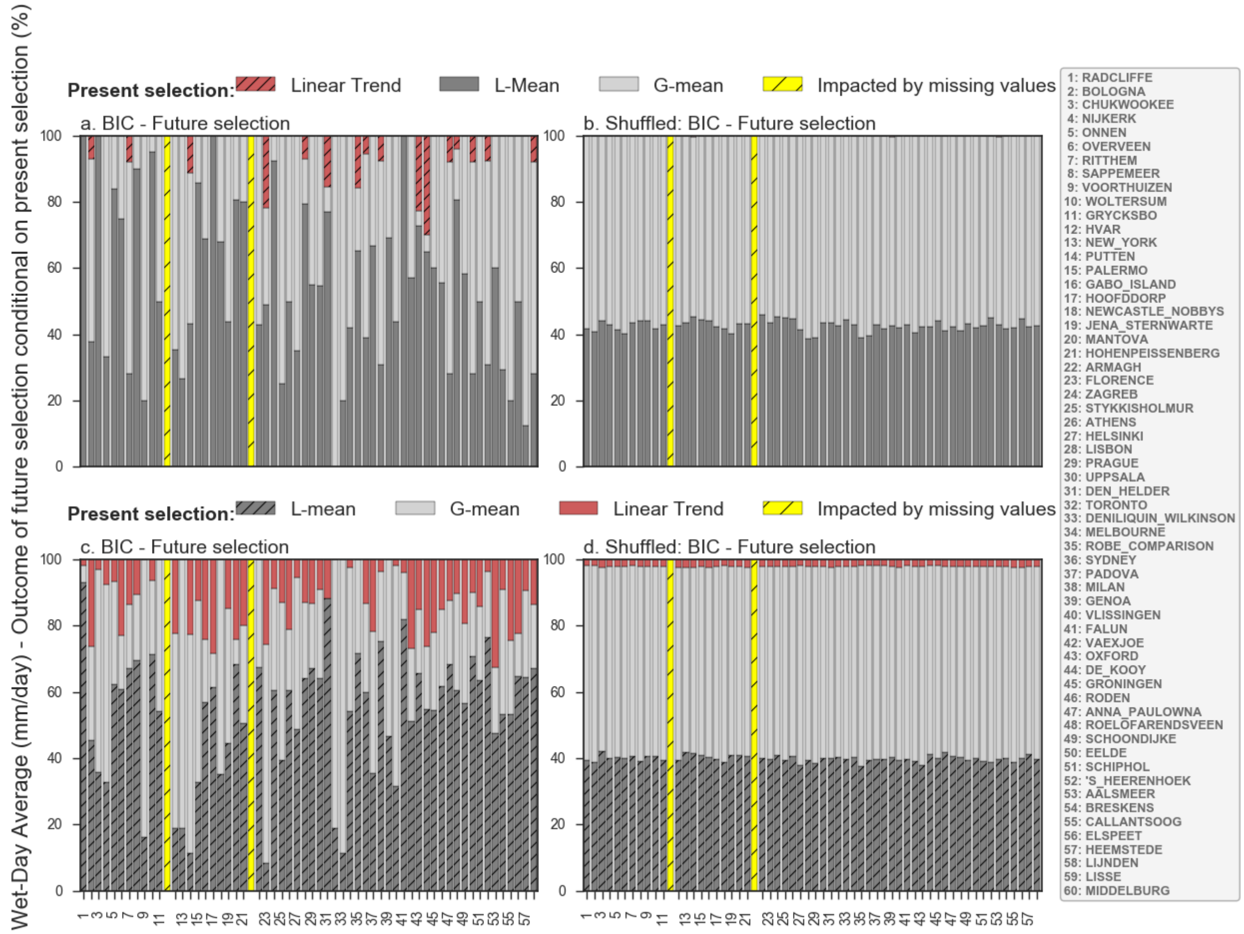

Figure 18. Percentage of selection as best model for the future for each model conditional on its selection on the present based on RMSE and BIC evaluation for wet-day average of both the original and the shuffled data. White straps correspond to cases of 0 present selection.

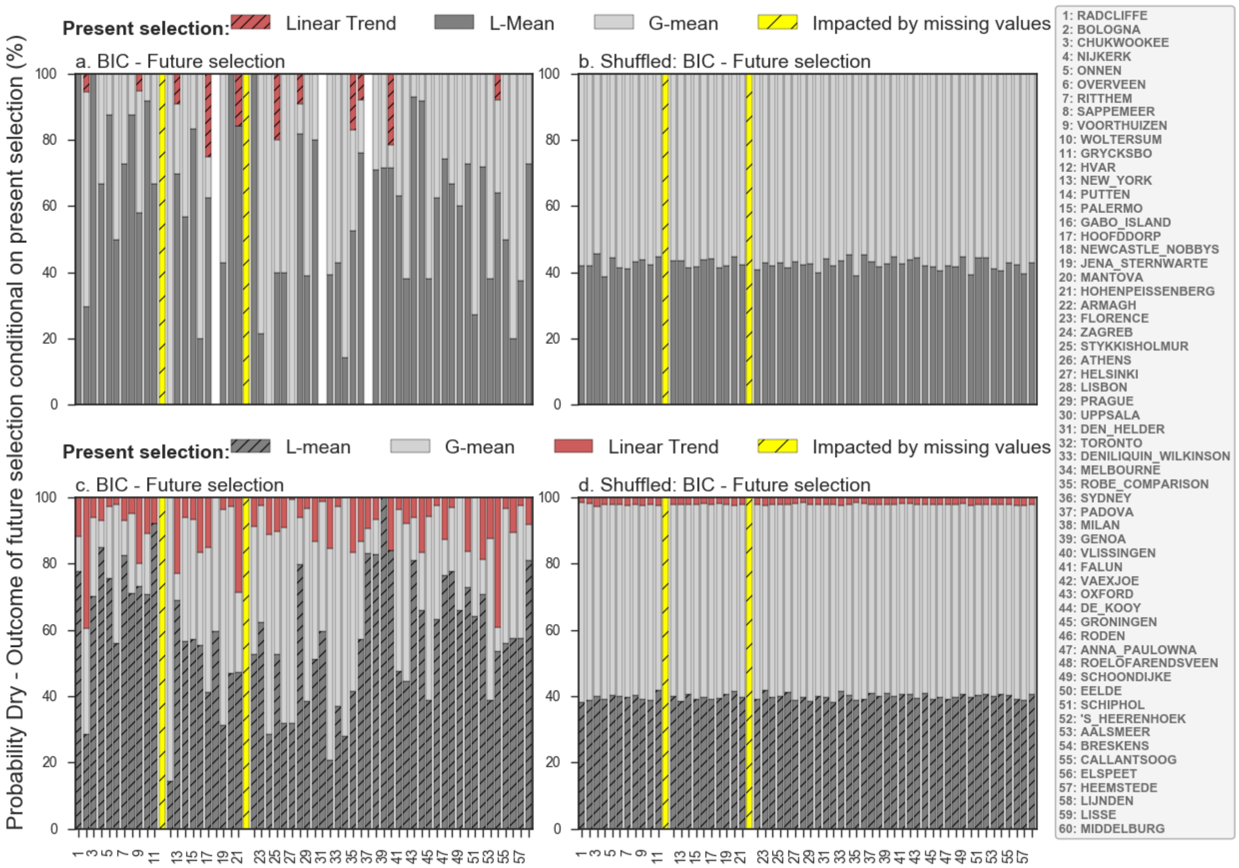

Figure 19. Percentage of selection as best model for the future for each model conditional on its selection on the present based on RMSE and BIC evaluation for probability dry of both the original and the shuffled data. White straps correspond to cases of 0 present selection.

4.4 Temporal propagation of error in projections

For some of the longest and uninterrupted stations of our dataset, i.e. one of the oldest daily stations worldwide, the Chukwookee station (Jhun and Moon, 1997) in Korea (241 years), the Prague station in Czech Republic (211 years) and the Radcliffe station in the UK (188 years) we also evaluate certain temporal patterns of the models' performance. We examine all indices for Radcliffe, and two indices for Chukwookee, i.e. annual maxima and totals. We find that a known change in instrumentation that occurred in the Korean data in 1908 (Lee and Kim, 2018) affected the statistics of probability dry and wet-day average resulting in inhomogeneity between the two periods. Therefore, for these indices, we choose to evaluate instead the next longer station, that of Prague.

All models show pronounced periods of clustering of errors that illustrate the strong variability of rainfall climatology manifested throughout the observed records. However for the majority of time, the mean models are at the lower front of the errors, with the L-mean model showing overall superior performance in some cases (Fig. 22-23), owing to the presence of the persistence, also known as Hurst-Kolmogorov (HK) dynamics (Koutsoyiannis, 2011; Dimitriadis, 2017) in the generating process. It is also clear that the linear trend model results in higher errors, and higher variance of errors as already discussed. It is however additionally showcased that the trend model is trapped for long periods in areas of poor predictions that also quickly deteriorate, taking longer to converge to the other two models in areas of lower errors (e.g. Fig. 21). This is attributed to the fact that the trend model is sensitive to the presence of extreme observations in the calibration period, which historically, did not continue in the climatic 'future'.

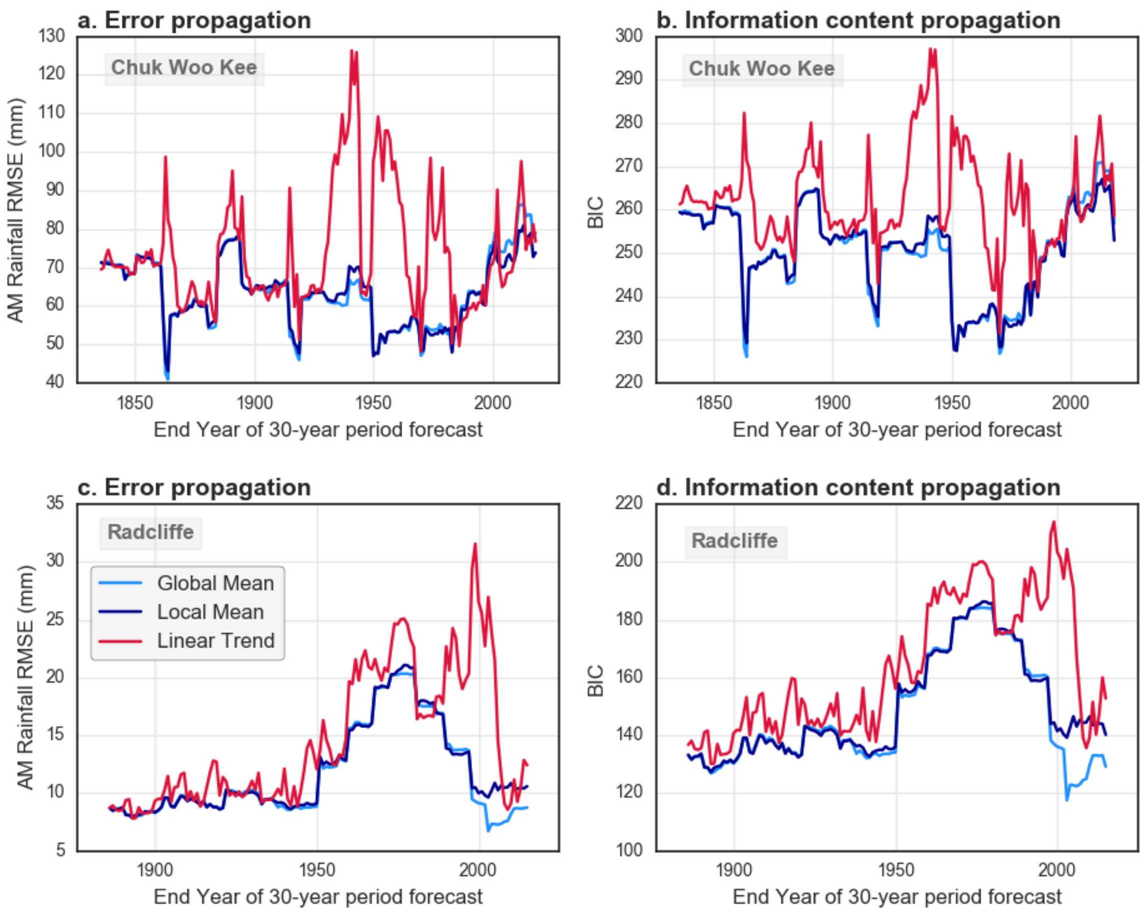

Figure 20. Propagation of RMSE and BIC value for the three models along their application to 30 year moving windows of the records for annual maxima.

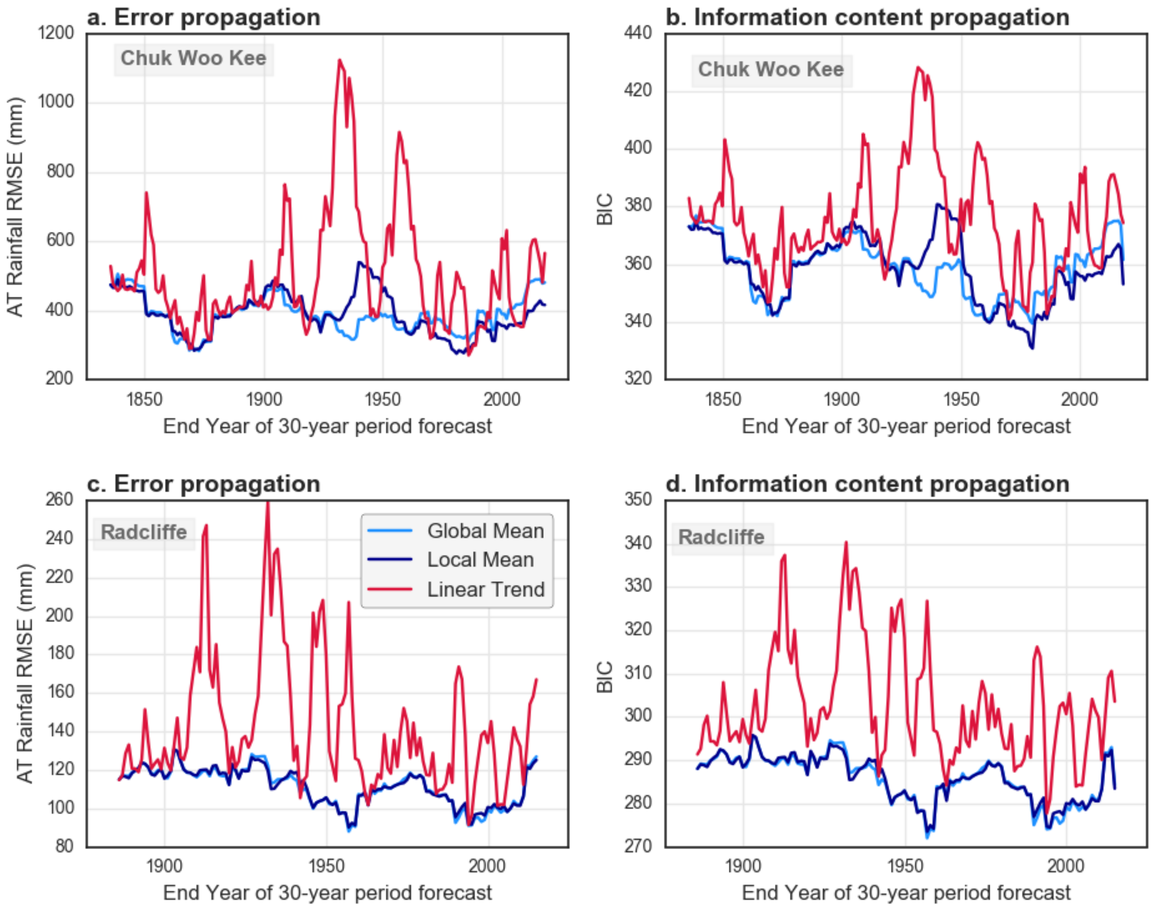

Figure 21. Propagation of RMSE and BIC value for the three models along their application to 30 year moving windows of the records for annual totals.

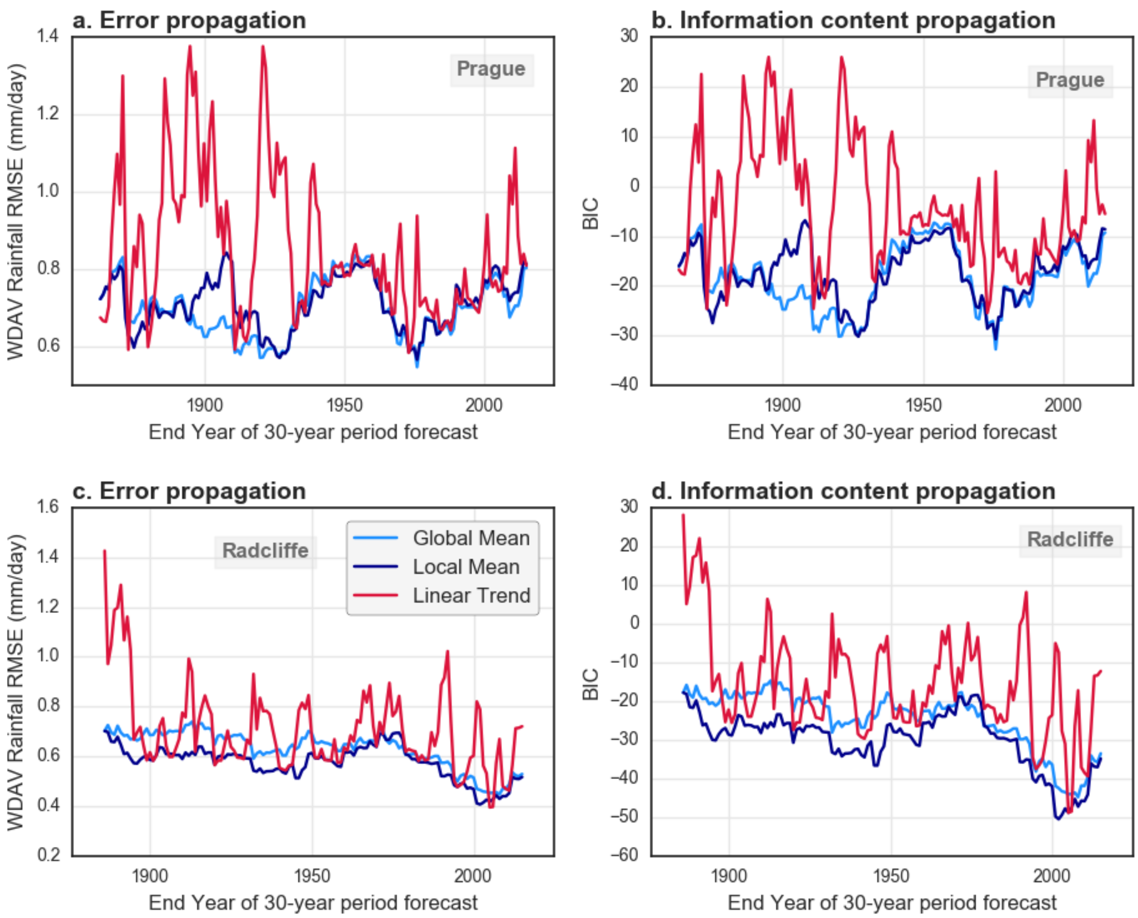

Figure 22. Propagation of RMSE and BIC value for the three models along their application to 30 year moving windows of the records for wet-day average.

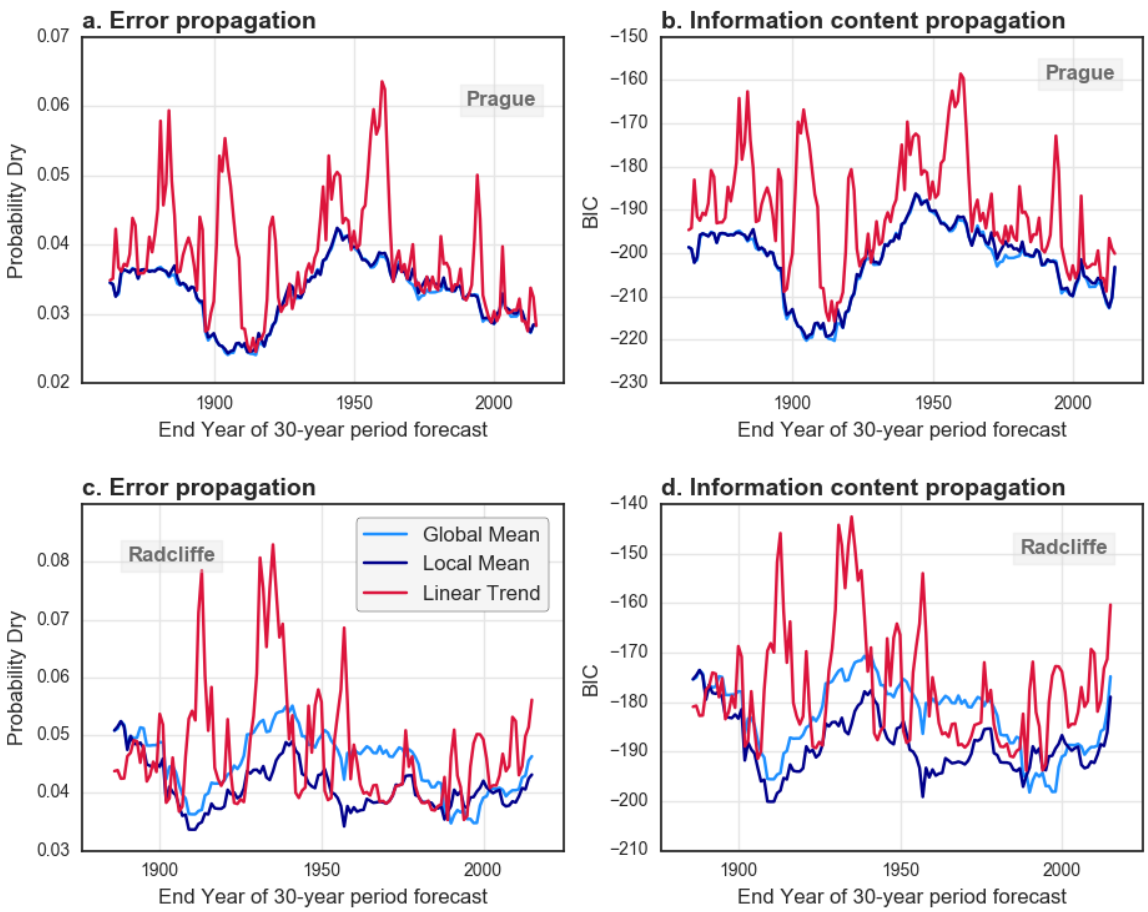

Figure 23. Propagation of RMSE and BIC value for the three models along their application to 30 year moving windows of the records for probability dry.

5. Discussion and Conclusions

Applications of trend modelling are increasingly encountered in contemporary hydrological studies aiming to identify deterministic signals of change. In the standard methodological framework, trends are fitted on a given observation period and subsequently validated based on the outcome of the binary type hypothesis testing applied with predefined significance levels; a practice however highly criticised on theoretical and empirical grounds (Section 3.4). This research reframes the problem of trend evaluation, this time as a model selection problem, focusing on the assessment of the predictive qualities of trends when compared to simpler alternative models, as the global and local mean. The performance of models is evaluated by their future projections of four rainfall indices: annual maxima, annual totals, wet-day average rainfall and probability dry, estimated from the 60 longest rainfall records having over 150 years of daily data. The candidate models are judged by the Bayesian Information Criterion (BIC) both in calibration and future validation mode, following a process in which the whole observational record is progressively scanned by climatic moving-windows of 30 years. The outcomes of the model selection in the future periods are subsequently processed in order to derive the statistics of the historical performance. This model selection framework also considers metrics of model properties that reveal the temporal behaviour of the performance, as the percentage of time-windows of selection, the variance of error, and the error propagation behaviour. These expose aspects of the historical performance which are often masked by the summary statistics, such as the average error. Its evaluations are also contrasted to the ones obtained from application of a simple performance metric given by the RMSE.

Results consistently disfavour trend modelling for all rainfall indices. It is revealed by all evaluations that the mean models outperform the trend models for all four indices and formulated metrics, namely the average RMSE and BIC, the standard deviation of the RMSE, the percentage of selection as best model for the future, as well as the percentage of selection as best model for the future conditional on selection as best model for the present. From a statistical point of view, trends appear to be the worst modelling choice among the three models, while it is found that even if they are selected in the present, it is highly likely that they will fail in the future.

Examination of the error propagation patterns in the longest stations highlight the fact that their future performance is easily trapped in prolonged periods of high errors due to their sensitivity to the existence of extreme observations in the calibration period. Interestingly, among the four indices, the worst performance of trends is found in the case of annual maxima followed by the probability dry. These findings also hold true for the examined recent 30 year period of most stations that show no signs of prevailing trends.

It appears that linear trend projections exhibit the typical characteristics of overfitting, i.e. sharp deterioration of future performance as opposed to excellent performance in calibration in terms of RMSE, together with increased variance of errors. The BIC provides a solid theoretical framework for partly addressing overfitting, showing very good results for the randomized (shuffled) data where the 'trend detection rate' in the future is very small by BIC (2%), contrary to the one by RMSE (15%).

We should stress that we have fitted trends in all given periods, without discrimination, while for the exploratory analysis of the recent years, we have chosen the most recent 30 year period of each station, with no prior visual examination. In many common cases, where trends are identified from the data in hindsight, i.e. the trends are fitted in periods where they are also found significant, the start period and/or the length of the trend should also be considered a parameter of the trend model in the BIC evaluation. This should be done in order to compensate for hindsight bias, in essence for the fact that in retrospect, past events appear more predictable that they were in the actual time of occurrence.

The comparison to the shuffled data reveals the key role of the dependence properties, else long-term persistence or HK dynamics (Koutsoyiannis, 2011; O'Connell et al., 2016; Dimitriadis, 2017), of the indices on the choice of the best future model. Remarkably, indices known for their persistence properties, such as annual totals (Iliopoulou et al., 2018b; Tyralis et al., 2018) and probability dry (Koutsoyiannis, 2006) show a marked preference for the Local-Mean model, while others where persistence is less manifested, as annual maxima (Iliopoulou and Koutsoyiannis, 2019) the performance of the global and the local mean model are comparable; still the variance of the errors being smaller for the latter. Therefore, the local mean model appears to be a good candidate model for most projections of examined rainfall properties. A thorough treatment of the theoretical basis and practical formulation of local mean models in relation to the persistence properties of the parent process is given by Koutsoyiannis (2020).

Attempting to make forecasts in complex systems, such as mechanisms generating rainfall extremes, is undoubtedly a challenging task and it is expected for models to fail and predictions to be invalidated. Yet the recent inclination in discovering trends has led to wildly overlooking modelling practices, as split-sample validation, long established in hydrology. The present research poses a counter-argument against the trendy use of trend models for hydroclimatic projections on the basis of their inferior predictive performance as evaluated from classical model selection theory. Considering a pool of alternative models instead, is critical for decision-making under uncertainty and risk policies, and in this regard we have shown that local mean models are simpler and better alternatives.

Acknowledgments

We greatly thank the Radcliffe Meteorological Station, the Icelandic Meteorological Office (Trausti Jónsson), the Czech Hydrometeorological Institute, the Finnish Meteorological Institute, the National Observatory of Athens, the Department of Earth Sciences of the Uppsala University and the Regional Hydrologic Service of the Tuscany Region (servizio.idrologico@regione.toscana.it) for providing the required data for each region respectively. We are also grateful to Professor Ricardo Machado Trigo (University of Lisbon) for providing the Lisbon timeseries, to Professor Marco Marani (University of Padua) for providing the Padua timeseries and to Professor Joo-Heon Lee (Joongbu University) for providing the

Seoul timeseries. All the above data were freely provided after contacting the acknowledged sources. The remaining timeseries are publicly available by the data providers in the ECA&D project (http://www.ecad.eu), and in the GHCN-Daily database (https://data.noaa.gov/dataset/global-historical-climatology-network-daily-ghcn-daily-version-3). The analyses were performed in the Python 2.6 (Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org) using the contributed packages pandas, scipy and seaborn.

References

Akaike, H., 1974. A new look at the statistical model identification, in: Selected Papers of Hirotugu Akaike. Springer, pp. 215–222.

Akaike, H., 1969. Fitting autoregressive models for prediction. Annals of the institute of Statistical Mathematics 21, 243–247.

Amrhein, V., Greenland, S., 2018. Remove, rather than redefine, statistical significance. Nature Human Behaviour 2, 4.

Amrhein, V., Greenland, S., McShane, B., 2019. Scientists rise up against statistical significance. Nature 567, 305. https://doi.org/10.1038/d41586-019-00857-9

Anderson, D.R., Burnham, K., 2004. Model selection and multi-model inference. Second. NY: Springer-Verlag 63.

Biasutti, M., 2019. Rainfall trends in the African Sahel: Characteristics, processes, and causes. Wiley Interdisciplinary Reviews: Climate Change e591.

Bunting, A., Dennett, M.D., Elston, J., Milford, J.R., 1976. Rainfall trends in the west African Sahel. Quarterly Journal of the Royal Meteorological Society 102, 59–64.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research 33, 261–304.

Burt, T.P., Howden, N.J.K., 2011. A homogenous daily rainfall record for the Radcliffe Observatory, Oxford, from the 1820s. Water Resour. Res. 47, W09701. https://doi.org/10.1029/2010WR010336

Cairncross, A., 1969. Economic forecasting. The Economic Journal 79, 797–812.

Cohn, T.A., Lins, H.F., 2005. Nature’s style: Naturally trendy. Geophysical Research Letters 32.

Craig, R.K., 2010. Stationarity is dead-long live transformation: five principles for climate change adaptation law. Harv. Envtl. L. Rev. 34, 9.

Degefu, M.A., Alamirew, T., Zeleke, G., Bewket, W., 2019. Detection of trends in hydrological extremes for Ethiopian watersheds, 1975–2010. Regional Environmental Change 1–11.

Dimitriadis, P., 2017. Hurst-Kolmogorov dynamics in hydrometeorological processes and in the microscale of turbulence.

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Advances in Water Resources 30, 1371–1386.

Folton, N., Martin, E., Arnaud, P., L’Hermite, P., Tolsa, M., 2019. A 50-year analysis of hydrological trends and processes in a Mediterranean catchment. Hydrology and Earth System Sciences 23, 2699–2714.

Georgakakos, K.P., Seo, D.-J., Gupta, H., Schaake, J., Butts, M.B., 2004. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles.

Journal of Hydrology, The Distributed Model Intercomparison Project (DMIP) 298, 222–241. https://doi.org/10.1016/j.jhydrol.2004.03.037

Haylock, M., Nicholls, N., 2000. Trends in extreme rainfall indices for an updated high quality data set for Australia, 1910–1998. International Journal of Climatology: A Journal of the Royal Meteorological Society 20, 1533–1541.

Houghton, J.T., Jenkins, G.J., Ephraums, J.J., 1991. Climate change.

Iliopoulou, T., Koutsoyiannis, D., Montanari, A., 2018a. Characterizing and modeling seasonality in extreme rainfall. Water Resources Research 54, 6242–6258.

Iliopoulou, T., Papalexiou, S.M., Markonis, Y., Koutsoyiannis, D., 2018b. Revisiting long-range dependence in annual precipitation. Journal of Hydrology 556, 891–900.

Iliopoulou, T., Koutsoyiannis, D., 2019. Revealing hidden persistence in maximum rainfall records. Hydrological Sciences Journal (*accepted*).

Jhun, J.G., Moon, B.K., 1997. Restorations and analyses of rainfall amount observed by Chukwookee. J. Korean Meteor. Soc 33, 691–707.

Kellogg, W.W., 2019. Climate change and society: consequences of increasing atmospheric carbon dioxide. Routledge.

Khan, N., Pour, S.H., Shahid, S., Ismail, T., Ahmed, K., Chung, E.-S., Nawaz, N., Wang, X., 2019. Spatial distribution of secular trends in rainfall indices of Peninsular Malaysia in the presence of long-term persistence. Meteorological Applications.

Klein Tank, A.M.G., Wijngaard, J.B., Können, G.P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. International journal of climatology 22, 1441–1453.

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal 31, 13–24.

Koutsoyiannis, D., 2011. Hurst-Kolmogorov Dynamics and Uncertainty. JAWRA Journal of the American Water Resources Association 47, 481–495.

Koutsoyiannis, D., 2006. An entropic-stochastic representation of rainfall intermittency: The origin of clustering and persistence. Water Resources Research 42.

Koutsoyiannis, D., Montanari, A., 2015a. Negligent killing of scientific concepts: the stationarity case. Hydrological Sciences Journal 60, 1174–1183.

Koutsoyiannis, D., Montanari, A., 2015b. Negligent killing of scientific concepts: the stationarity case. Hydrological Sciences Journal 60, 1174–1183.

Koutsoyiannis, D., Montanari, A., 2007. Statistical analysis of hydroclimatic time series: Uncertainty and insights. Water resources research 43.

Koutsoyiannis, D., 2020. Stochastics of hydroclimatic extremes: a cool look at risk (*in preparation*).

Kumar, V., Jain, S.K., Singh, Y., 2010. Analysis of long-term rainfall trends in India. Hydrological Sciences Journal–Journal des Sciences Hydrologiques 55, 484–496.

Kutiel, H., Trigo, R.M., 2014. The rainfall regime in Lisbon in the last 150 years. Theoretical and applied climatology 118, 387–403.

Laio, F., Di Baldassarre, G., Montanari, A., 2009. Model selection techniques for the frequency analysis of hydrological extremes. Water Resources Research 45.

Lee, J.-H., Kim, H., 2018. Long-term Rainfall Observation Records through Korean History and its Application for Modern Hydro-Meteorological Science, in: EGU General Assembly Conference Abstracts. p. 5717.

Marani, M., Zanetti, S., 2015. Long-term oscillations in rainfall extremes in a 268 year daily time series. Water Resources Research 51, 639–647.

McCarl, B.A., Villavicencio, X., Wu, X., 2008. Climate change and future analysis: is stationarity dying? American Journal of Agricultural Economics 90, 1241–1247.

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G., 2012. An Overview of the Global Historical Climatology Network-Daily Database. J. Atmos. Oceanic Technol. 29, 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1

Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., 2008. Stationarity is dead: Whither water management? Science 319, 573–574.

Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., Dettinger, M.D., Krysanova, V., 2015. On critiques of "Stationarity is dead: Whither water management?" Water Resources Research 51, 7785–7789.

Modarres, R., da Silva, V. de P.R., 2007. Rainfall trends in arid and semi-arid regions of Iran. Journal of arid environments 70, 344–355.

Montanari, A., Koutsoyiannis, D., 2014. Modeling and mitigating natural hazards: Stationarity is immortal! Water Resources Research 50, 9748–9756.

Moss, R.H., Edmonds, J.A., Hibbard, K.A., Manning, M.R., Rose, S.K., Van Vuuren, D.P., Carter, T.R., Emori, S., Kainuma, M., Kram, T., 2010. The next generation of scenarios for climate change research and assessment. Nature 463, 747.

Ntegeka, V., Willems, P., 2008. Trends and multidecadal oscillations in rainfall extremes, based on a more than 100-year time series of 10 min rainfall intensities at Uccle, Belgium. Water Resources Research 44.

Nuzzo, R., 2014. Scientific method: statistical errors. Nature News 506, 150.

O'Connell, P.E., Koutsoyiannis, D., Lins, H.F., Markonis, Y., Montanari, A., Cohn, T., 2016. The scientific legacy of Harold Edwin Hurst (1880–1978). Hydrological Sciences Journal 61, 1571–1590.

Oreskes, N., 2004. The scientific consensus on climate change. Science 306, 1686–1686.

Pachauri, R.K., Allen, M.R., Barros, V.R., Broome, J., Cramer, W., Christ, R., Church, J.A., Clarke, L., Dahe, Q., Dasgupta, P., 2014. Climate change 2014: synthesis report. Contribution of Working Groups I. II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change 151.

Papalexiou, S.M., Koutsoyiannis, D., 2013. Battle of extreme value distributions: A global survey on extreme daily rainfall. Water Resources Research 49, 187–201.

Papalexiou, S.M., Montanari, A., 2019. Global and Regional Increase of Precipitation Extremes under Global Warming. Water Resources Research.

Parmesan, C., Yohe, G., 2003. A globally coherent fingerprint of climate change impacts across natural systems. Nature 421, 37.

Priestley, M.B., 1981. Spectral analysis and time series. Academic press London.

Quadros, L.E. de, Mello, E.L. de, Gomes, B.M., Araujo, F.C., 2019. Rainfall trends for the State of Paraná: present and future climate. Revista Ambiente & Água 14.

Rahimi, M., Fatemi, S.S., 2019. Mean versus Extreme Precipitation Trends in Iran over the Period 1960–2017. Pure and Applied Geophysics 1–19.

Rotstayn, L.D., Lohmann, U., 2002. Tropical rainfall trends and the indirect aerosol effect. Journal of Climate 15, 2103–2116.

Schwarz, G., 1978. Estimating the dimension of a model. The annals of statistics 6, 461–464.

Serinaldi, F., Kilsby, C.G., 2018. Unsurprising Surprises: The Frequency of Record-breaking and Overthreshold Hydrological Extremes Under Spatial and Temporal Dependence. Water Resources Research 54, 6460–6487.

Serinaldi, F., Kilsby, C.G., Lombardo, F., 2018. Untenable nonstationarity: An assessment of the fitness for purpose of trend tests in hydrology. Advances in Water Resources 111, 132–155.

Solomon, S., Qin, D., Manning, M., Averyt, K., Marquis, M., 2007. Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC. Cambridge university press.

Tyralis, H., Dimitriadis, P., Koutsoyiannis, D., O'Connell, P.E., Tzouka, K., Iliopoulou, T., 2018. On the long-range dependence properties of annual precipitation using a global network of instrumental measurements. Advances in Water Resources 111, 301–318.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 70, 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R.L., Schirm, A.L., Lazar, N.A., 2019. Moving to a world beyond "p< 0.05." Taylor & Francis.

Ye, M., Meyer, P.D., Neuman, S.P., 2008. On model selection criteria in multimodel analysis. Water Resources Research 44. https://doi.org/10.1029/2008WR006803

Appendix

Table A1. Properties (name, source, latitude, longitude, start year, end year, record length and missing values percentage) of the 60 longest stations used in the analysis sorted by decreasing length. For the global datasets, the European Climate Assessment dataset (ECA; http://www.ecad.eu) and the Global Historical Climatology Network Daily database (GHCND; https://data.noaa.gov/dataset/global-historical-climatology-network-daily-ghcn-daily-version-3), the station identifier is also reported. Asterisks (*) in the "end year" column denote data that have been continued from a second source. The country of each station is abbreviated in parentheses aside its name.

NAME	SOURCE	LAT	LON	START YEAR	END YEAR	RECORD LENGTH	MISSING %
PADOVA (IT)	Marani and Zanetti (2015)	45.87	11.53	1725	2013	289	5.04
CHUK-WOO-KEE, SEOUL (KR)	Jhun and Moon (1997) and Korea Meteorological Agency	37.53	127.02	1777	2017*****	241	0.00
HOHENPEISSENBERG (DE)	ECA: 48 HOHENPEISSENBERG DE	47.80	11.01	1781	2017	237	25.56
PALERMO (IT)	GHCND:ITE00105250	38.11	13.35	1797	2008	212	17.16
PRAGUE (CZ)	Czech Hydrometeorological Institute	50.05	14.25	1804	2014	211	0.20
BOLOGNA (IT)	GHCND:ITE00100550 and Dext3r of ARPA Emilia Romagna, Rete di monitoraggio RIRER (http://www.smr.arpa.emr.it/dext3r/)	44.50	11.35	1813	2018*****	206	0.00
JENA STERNWARTE GM (DE)	GHCND:GM000004204	50.93	11.58	1826	2015	190	5.47
RADCLIFFE (UK)	Radcliffe Meteorological Station (Burt and Howden, 2011)	51.76	-1.26	1827	2014	188	0.05
UPPSALA (SE)	Department of Earth Sciences of the Uppsala University	59.86	17.63	1836	2014	179	0.02
TORONTO (CA)	GHCND:CA006158350	43.67	-79.40	1840	2015	176	5.97
GENOA (IT)	GHCND:ITE00100552	44.41	8.93	1833	2008	176	0.00
ONNEN (NL)	ECA :2491 ONNEN NL	53.15	6.67	1846	2018	173	1.10
SAPPEMEER (NL)	ECA:2507 SAPPEMEER NL	53.17	6.73	1846	2018	173	1.10
WOLTERSUM (NL)	ECA:2553 WOLTERSUM NL	53.27	6.72	1846	2018	173	1.14
GRONINGEN (NL)	ECA:147 GRONINGEN NL	53.18	6.60	1846	2018	173	1.10
RODEN (NL)	ECA:516 RODEN NL	53.15	6.43	1846	2018	173	1.10
EELDE (NL)	ECA:164 EELDE NL	53.12	6.58	1846	2018	173	1.10
HELSINKI (FI)	Finnish Meteorological Institute	60.17	24.93	1845	2015	171	0.33
MANTOVA (IT)	GHCND:ITE00100553	45.16	10.80	1840	2008	169	5.75
DEN_HELDER (NL)	ECA:146 DEN_HELDER NL	52.93	4.75	1850	2018	169	1.13
DE_KOOY (NL)	ECA:145 DE_KOOY NL	52.92	4.78	1850	2018	169	1.13
ANNA_PAULOWNA (NL)	ECA:521 ANNA_PAULOWNA NL	52.87	4.83	1850	2018	169	1.13

CALLANTSOOG (NL)	ECA:2382 CALLANTSOOG NL	52.85	4.70	1850	2018	169	1.13
RITTHEM (NL)	ECA:2503 RITTHEM NL	51.47	3.62	1854	2018	165	1.16
VLISSINGEN (NL)	ECA:166 VLISSINGEN NL	51.44	3.60	1854	2018	165	1.16
SCHOONDIJKE (NL)	ECA:572 SCHOONDIJKE NL	51.35	3.55	1854	2018	165	1.16
'S_HEERENHOEK (NL)	ECA:2350 'S_HEERENHOEK NL	51.47	3.77	1854	2018	165	1.16
BRESKENS (NL)	ECA:2377 BRESKENS NL	51.40	3.55	1854	2018	165	1.16
MIDDELBURG (NL)	ECA:2474 MIDDELBURG NL	51.48	3.60	1854	2018	165	1.16
ARMAGH (UK)	GHCND:UK000047811	54.35	-6.65	1838	2001	164	0.26
OXFORD (UK)	GHCND:UK000056225	51.77	-1.27	1853	2015	163	0.42
HVAR (HR)	ECA:1686 HVAR HR	43.17	16.45	1857	2018	162	7.74
MELBOURNE REGIONAL OFFICE (AS)	GHCND:ASN00086071	-37.81	144.97	1855	2015	161	1.29
STYKKISHOLMUR (IS)	Icelandic Meteorological Office	65.08	-22.73	1856	2015	160	1.00
GRYCKSBO_D (SE)	ECA:6456 GRYCKSBO_D SE	60.69	15.49	1860	2018	159	0.62
FALUN (SE)	GHCND:SW000010537	60.62	15.62	1860	2018	159	0.89
VAEXJOE (SE)	GHCND:SWE00100003	56.87	14.80	1860	2018	159	4.13
FLORENCE (IT)	Regional Hydrologic Service of the Tuscany Region	43.80	11.20	1822	1979	158	2.00
SYDNEY OBSERVATORY HILL (AS)	GHCND:ASN00066062	-33.86	151.21	1858	2015	158	0.48
DENILIQUIN WILKINSON ST (AS)	GHCND:ASN00074128	-35.53	144.95	1858	2014	157	1.37
ZAGREB GRIC (HR)	GHCND:HR000142360	45.82	15.98	1860	2015	156	1.54
ROBE COMPARISON (AS)	GHCND:ASN00026026	-37.16	139.76	1860	2015	156	3.66
GABO ISLAND LIGHTHOUSE (AS)	GHCND:ASN00084016	-37.57	149.92	1864	2018	155	3.36
NEWCASTLE NOBBYS SIGNAL STATIO (AS)	GHCND:ASN00061055	-32.92	151.80	1862	2015	154	2.55
OVERVEEN (NL)	ECA:2497 OVERVEEN NL	52.40	4.60	1866	2018	153	1.25

HOOFDDORP (NL)	ECA:151 HOOFDDORP NL	52.32	4.70	1866	2018	153	1.25
ROELOFARENDSVEEN (NL)	ECA:540 ROELOFARENDSVEEN NL	52.22	4.62	1866	2018	153	1.29
SCHIPHOL (NL)	ECA:593 SCHIPHOL NL	52.32	4.79	1866	2018	153	1.25
AALSMEER (NL)	ECA:2351 AALSMEER NL	52.27	4.77	1866	2018	153	1.25
HEEMSTEDE (NL)	ECA:2430 HEEMSTEDE NL	52.35	4.63	1866	2018	153	1.25
LIJNDEN_(NH) (NL)	ECA:2466 LIJNDEN_(NH) NL	52.35	4.75	1866	2018	153	1.25
LISSE (NL)	ECA:2467 LISSE NL	52.27	4.55	1866	2018	153	1.29
NIJKERK (NL)	ECA:2484 NIJKERK NL	52.23	5.47	1867	2018	152	0.75
VOORTHUIZEN (NL)	ECA:2542 VOORTHUIZEN N	52.18	5.62	1867	2018	152	0.75
PUTTEN_(GLD) (NL)	ECA: 551 PUTTEN_(GLD) NL	5.62	14.00	1867	2018	152	0.75
ATHENS (GR)	National Observatory of Athens	37.97	23.72	1863	2014	152	0.66
ELSPEET (NL)	ECA:2404 ELSPEET NL	52.28	5.78	1867	2018	152	0.75
LISBON (PT)	Kutiel and Trigo (2014)	39.20	-9.25	1863	2013	151	1.06
MILAN (IT)	GHCND:ITE00100554	45.47	9.19	1858	2008	151	0.12
NEW_YORK_CNTRL_PK_TWR (US)	GHCND: USW00094728	40.78	-73.97	1869	2018	150	0.51