# 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

- Prediction-oriented evaluation of rainfall trends
- Trend and mean models are used to project 30 years of rainfall indices
- The predictive skill of the models is assessed by moving-window validation
- Trends have the worst performance and local mean models the best

## Abstract

Non-stationarity approaches have been increasingly popular in hydrology, being dominated by the practice of identifying linear trends in data through in-sample analysis. In this work, we reframe the problem of trend identification using the out-of-sample predictive performance of trends as a reference point. We devise a systematic methodological framework in which trends are compared to simpler mean models, based on their performance in predicting climatic-scale (30-year) annual rainfall indices, i.e. maxima, totals, wet-day average and probability dry, from long-term daily records. The models are calibrated in two different schemes: block-moving, i.e. fitted on the recent 30 years of data, obtaining the local trend and local mean, and global-moving, i.e. fitted on the whole period known to an observer moving in time, thus obtaining the global trend and global mean. The investigation of empirical records spanning over 150 years

suggests that a great degree of variability has been ever present in the rainfall process, leaving small potential for long-term predictability. The local mean model ranks first in terms of average predictive performance, followed by the global mean and the global trend, in decreasing order of performance, while the local trend model ranks last among the models, showing the worst performance overall. Parallel experiments from synthetic timeseries characterized by persistence corroborated this finding, suggesting that future long-term variability of persistent processes is better captured using parsimonious features of the past. In line with the empirical findings, it is shown that, prediction-wise, simple is preferable to trendy.

**Keywords**: trends, rainfall extremes, probability dry, out-of-sample validation, predictive performance, rainfall projections

## 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 (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). For a quantitative analysis of the relevant literature, the reader is referred to Appendix I. 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 recent 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 represented in the stationarity framework (Koutsoyiannis and Montanari, 2007; Serinaldi and Kilsby, 2018).

Deterministic trend modelling in hydrology has been examined —and mostly criticized, on different grounds, namely with respect to empirical evidence (McKitrick and Christy, 2019; 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). In this research, we examine the trend modelling framework from a new perspective, through the evaluation of its out-of-sample modelling qualities, namely, its predictive powers for a given record.

For this purpose, we introduce a validation framework for the evaluation of the results, adding simpler, mean models in the pool of candidates, and we base the reasoning of model selection on the statistical out-of-sample performance of the models. 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 explanatory performance, mostly by means of in-sample measures, as hypothesis testing (Shmueli, 2010). In this work, we extend the simple split-sample validation by introducing a moving window calibration and validation approach that progressively scans each record by sliding windows of climatic-length (30 years). In this manner, we obtain a sample of estimates of the models' predictive performance, instead of a single value.

By shifting the focus to the predictive modelling of linear trend, this analysis seeks to answer the following key questions: (a) how well are the rainfall statistics of the most recent climatic period predicted by the candidate models based on the linear trend calibrated to the prior 30 year period? and (b) how do the statistics of the predictive performance of linear trends compare to the ones derived from application of simple mean models? The first question is driven by the omnipresent scientific concerns regarding intensification of extremes 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), and is consciously biased in favour of a model capturing the variability of the most recent

period of data. The second question introduces the abovementioned methodological framework for validating model predictions, which is applied to the empirical long-term rainfall records as well as to synthetic series produced in order to mimic the natural long-term variability of the rainfall process.

## 2. Dataset

Our dataset is an update of the previous long-term dataset explored in Iliopoulou et al. (2018) of long rainfall records surpassing 150 years of daily values. It 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 in the Appendix II (Table A1), along with a brief summary of the stations' properties; the geographic location of the rain gauges is shown in Figure 1. The length of the timeseries provides rare insights into long-term rainfall variability and enables the statistical evaluation of the predictive performance of linear trends from multiple time windows.

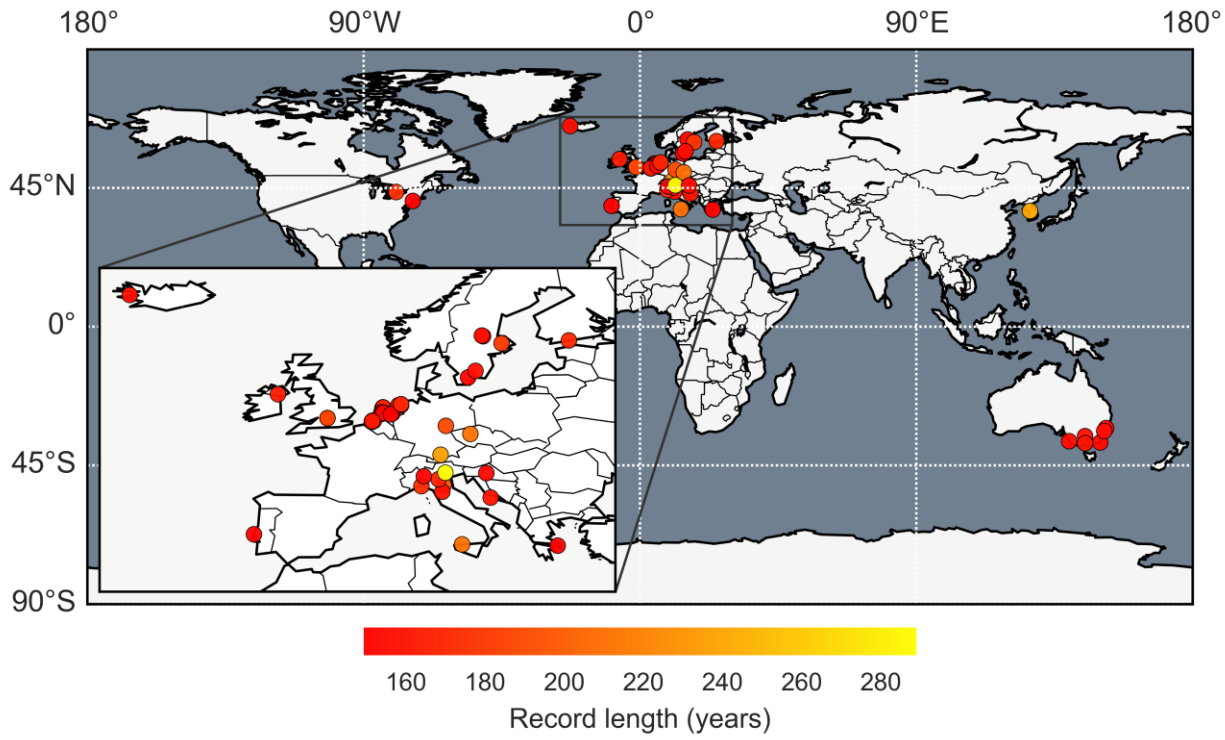



**Figure 1**. Map of the 60 stations with longest records used in the analysis.

## 3. Methodological framework

### 3.1 Overview of literature approaches to trend modelling: from explanatory trends to out-of-sample performance

It is well-known that studying the explanatory power of trends in hydroclimatic data is a very active research field; see the literature analysis included in the Appendix I for the rising use of relevant in-text words as well as in-title words from Google Scholar. Before discussing literature modelling strategies for trends, it is imperative to define the meaning of a trend per se. Although 'trends' are frequently used as a synonym of temporal 'changes' (Fig. AI3 provides a quantitative analysis on the use of both words) and their notion has sometimes been extended to encompass stochastic stationary models (Fatichi et al., 2009; Chandler and Scott, 2011), the general idea behind the trend concept, is that the expected value of a response variable $\underline{y}$ is

specified as a deterministic function of time *t*, $E\left[\underline{y}\right] = f(t)$. The function *f* may take different forms —the linear model being only the first one adopted, and the most widely used. Indeed, this definition of a trend can be traced back to the development of the field of econometrics in the early 20th century, when 'secular' trends, meaning long-term trends, were deemed to be a component of financial timeseries, along with seasonal variation, cycles and residual elements (Persons, 1922; Mitchell, 1930). Decomposition of a timeseries into components, one of them being a trend, continued to dominate the econometrics literature, although even at early times certain critiques were raised (Slutsky, 1927).

The most established technique to identify trends is hypothesis testing, i.e. a statistical inference technique that estimates the probability that an alternative hypothesis may hold true compared to the null one, characterizing the strength of evidence by significance levels. This is a dated scientific method for model evaluation, which has been in part misused. For instance, its misuse in hydrology has been showcased 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. Still, the common practice has been to misinterpret outcomes in favour of trends. Part of the statistician community argues against the concept (Nuzzo, 2014; Wasserstein and Lazar, 2016; Amrhein and Greenland, 2018; Wasserstein et al., 2019; Amrhein et al., 2019), with the main critique summarized in the statement of the American Statistical Association that "*the widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') 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). Other inference

techniques for assessing the plausibility of changes under an a priori assumed model are also used, most notably change point analysis (Hinkley, 1970), which attempts to identify points of abrupt changes in the data. This approach too, is very sensitive on a priori hypotheses about the expected degree of variability in the data (a brief discussion on the issue in provided in Chandler and Scott, 2011).

With a stronger focus on modelling power rather than confirmatory analysis, model selection criteria have been developed arising from Akaike's work (Akaike, 1969). Akaike has contributed to the introduction of information theory into model selection criteria (Akaike, 1974) which are now established 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). Information criteria are useful in that they try to achieve a better out-of-sample performance by prompting for parsimony when fitting the model to the calibration set. There is a vast literature on the asymptotic equivalence of information criteria and out-of-sample prediction measures under specific conditions (Stone, 1977; Shibata, 1980; Wei, 1992; Inoue and Kilian, 2006), which typically though imply large record lengths.

A discourse regarding the relative powers of the abovementioned 'in-sample' measures compared to the assessment of predictive or out-of-sample performance is active in numerous scientific fields (Breiman, 2001; Stein, 2002; Inoue and Kilian, 2006; Yarkoni and Westfall, 2017; Shmueli, 2010), while in fact, it has been argued that the distinction between the two approaches might only arise due to the different objectives of each study (Gauch, 2003; Inoue and Kilian, 2005). Obviously, predictive modelling dominates in operational fields concerned with short-term prediction, as numerical weather prediction (Lorenc, 1986), and in such domains,

it is widely acknowledged that the model yielding the best predictions, in non-stochastic terms, is not necessarily the 'true' one (Shmueli, 2010).

The premise of this work is that while explanatory performance of trends has been thoroughly explored in hydrological studies (e.g. Chandler and Scott (2011) provide a comprehensive review on the matter), much less attention has been given to the predictive performance of trend modelling. A simple explanation might lie in the fact that in many environmental studies trends have been employed as descriptors of changes or causal effects, and less as models for predictions, in spite of the fact that they strongly communicate expectations for the future by suggesting causal mechanisms (e.g. Fig. A2 on the combined use of the word 'trends' and 'projections'). The second reason could be related to the scarcity for long-term environmental data for out-of-sample validation. Therefore, our aim is to assess the relevance of long-term trend modelling in terms of point prediction, not examining elements of stochastic prediction and categorically, not engaging in the identification of a 'true' model for the data. We deem that this shift in point-of-view may provide contrasting insights to current literature with respect to the relevance of trends for operational applications.

### 3.2 Out-of-sample validation schemes

Cross-validation techniques are a systematic way to assess predictive power (Stone, 1974; Simonoff, 2012). The procedure typically entails multiple runs of validation schemes on random partitions of the original dataset and summarizes the model skill from the sample of all validation scores. Standard cross-validation is not straightforward to apply for timeseries data where the order of the data must be respected. Instead the use of a 'holdout' set for validation is frequently applied, e.g. in hydrology this is done by reserving some data for validation, while the rest are used for calibration (Klemeš, 1986). We consider an alternative approach respecting the data

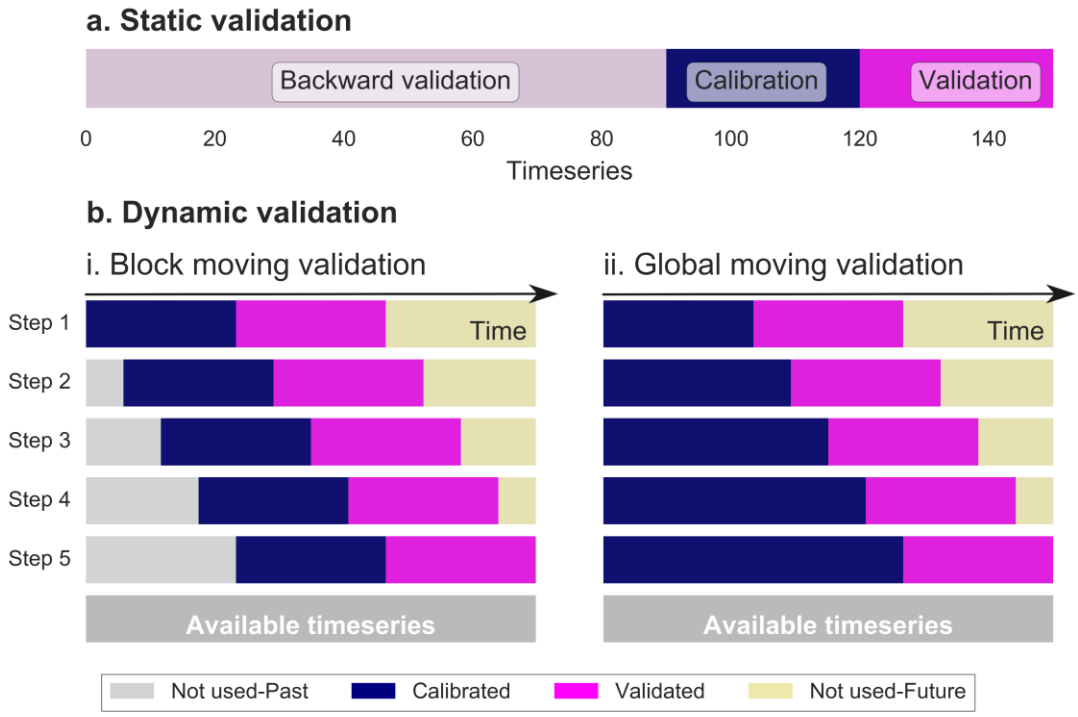
order, by performing calibration and validation in moving-window partitions of the original dataset, that constantly shift forward in time till the end of the record is reached. This approach is known as 'walk-forward' analysis in the field of econometrics (Kirkpatrick II and Dahlquist, 2010), and it is advantageous in that instead of a single measure of out-of-sample performance obtained by the 'split-sample' approach, a set of values is obtained, which can be statistically analysed. Further, it compensates for hindsight bias providing realistic estimates of historical predictability of changes by a given model. The statistics of the models' past performance can be considered a proxy of its future performance.

*3.2.1 Static calibration and validation*

We apply this type of analysis to the rainfall records by formulating two distinct calibration-validation schemes, which are illustrated in Fig. 2. In the first (Fig.2a) we evaluate the models' performance in capturing 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 intend to evaluate whether extremes have changed in a consistent manner in the second half of the 20th century, as they are commonly assumed. We also examine the performance of the models in backward validation, i.e. in predicting observations occurring before the calibration period (Fig. 2a). 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, since our focus is placed on the operational exploitation of records for predictive purposes and less on a summary of the results for a specific time period. However, the majority of the records span the whole 20th century, and extend beyond, with a few exceptions that are mentioned in Table A1.

*3.2.2 Dynamic calibration and validation*

198 The second scheme (Fig.2b) focuses on the historical performance of the models by the
199 'dynamic' (else, 'walk-forward') validation scheme introduced before. It assumes a hypothetical
200 observer moving in time and making predictions for the future 30-year period updating the
201 models as access to new information progressively becomes available. We formulate two
202 different schemes for making these predictions. In the first, which we call block-moving
203 calibration and validation, the models are calibrated on 30-year periods and validated by the next
204 'unobserved' 30 years, and this procedure is repeated by rolling the calibration and validation
205 origin in time (Fig.2bi). New information is gradually taking the place of the past information,
206 which is discarded by the 30-year sliding windows. The start of the first moving-window
207 coincides with the start of each station, while the start of the last calibration moving-window is
208 59 years prior to the end of the station, so that 30 years of validation data remain available. This
209 last validation window is the recent 30-year window that is exploited for validation in the static
210 scheme (Fig. 2a). The second scheme of the dynamic validation, which we call global-moving,
211 validates the models using sliding 30-year periods, exactly as in the prior scheme, but calibrates
212 the models on the whole available record, that is known at each time step to the observer.
213 Therefore, the origin of the calibration window remains stable, but the window gradually extends
214 in length as more data are assimilated in the model, while no data are discarded (Fig.2bii). This
215 scheme explores the potential of employing all available information to make a prediction for the
216 future. Since the validation periods are the same in both schemes, results between the two can be
217 directly compared.
218
219




220

221 **Figure 2**. Explanatory sketch showing the two calibration and validation schemes (a. Static
222 Validation and b. Dynamic Validation) for an example station.

223 For the evaluation of the candidate models we estimate the Root Mean Square Error, a standard
224 and established metric of goodness of fit (Sharma et al., 2019). The RMSE is defined as the
225 square root of the mean square error of the predicted values $\hat{x}_i$ with respect to the observed $x_i$:

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

226 where $n$ is the length of the data. We present the sample RMSE distribution of the models for
227 each station and we summarize the results by computing the average RMSE for each station and
228 its standard deviation. For the longest uninterrupted record of the station, we present a
229 comprehensive analysis including the temporal propagation of the errors.

230

### 3.3 Predictive models

Let $\underline{x_i}$ be a 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 $i \leq n$ the 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 subseries $x$ = ($x_1$, …, $x_i$).

To predict the unobserved periods, past or future, we employ two model structures. The first is the typical linear trend model, encompassing two parameters, a slope $b$ and an intercept $a$, whose mean $\mu$ is a deterministic linear function of time $t$:

$$\mu(t) = a + bt \tag{2}$$

The trend model is fitted via least-squares regression. Robust fitting techniques are also explored, namely median quantile regression (Koenker and Hallock, 2001) and the Theil-Sen slope estimation (Sen, 1968; Theil, 1992), but they did not yield better predictions, and hence, the least-squares approach, which is also more rigorous in theoretical terms (e.g. Papoulis, 1990), was retained. For details of the application and discussion on the results, the reader is referred to the analysis presented in Appendix III.

The second model considered is the mean model, including only one parameter, the mean of the present climatic period, extrapolated to the unobserved periods:

$$\mu(t) = a \tag{3}$$

According to the followed calibration scheme, fitted to block-moving (local) 30 years or to all the known (global) period, the trend model is termed local trend (L-Trend) and global trend (G-Trend), respectively, and likewise, the mean model, is termed local mean (L-Mean) and global mean (G-Mean). In the local models, the period [$i$ – 59, $i$ – 30] is used for calibration and the

[$i-29, i$] for validation, while in the global models, the period [1, $i$ –30] is used for calibration and the [$i$ – 29, $i$] period for validation as in the former scheme. We note that these two seemingly simplistic predictive models, i.e. the linear model fitted with least-squares and local average, can be found in a variety of theoretical results in statistical sciences, for instance use of (temporally) local data constitutes a central concept in the *k*-nearest neighbours technique, as discussed in Hastie et al. (2005), as well as in local regression as discussed in Chandler and Scott (2011).

### 3.4 Selected indices of rainfall extremes and quality control

We examine four statistical indices of rainfall: annual maxima (AM), annual totals (AT), annual wet-day average rainfall (WDAV) and probability dry (PD) also computed at the annual scale. As wet, we consider any day with rainfall surpassing the threshold of 1 mm, while values below this threshold are counted 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 low percentages of missing values (Table A1), which in most cases are clustered in the beginning of the records. A few records have consecutive missing periods which might imply a change of instrumentation or relocation of the gauge. To avoid possible artefacts in trend estimation in static validation (in backward validation) that may arise from such cases, we analyze periods containing less than 5% of consecutive missing values of the yearly indices. For

the dynamic calibration and validation scheme, we fit the models only if there exist at least 27 valid indices in each of the 30-year periods of calibration and validation.

### 3.5 Predictability of climatic changes under natural variability

In order to understand the predictive performance of the considered models under typical conditions of natural variability, we run similar experiments with synthetic timeseries reproducing increasing degrees of persistence. We recall that persistence, also known as Hurst-Kolmogorov dynamics, is associated with enhanced natural variability at all scales (Koutsoyiannis, 2003), which in turn implies increased unpredictability at large time horizons, with some potential for predictability at short time steps due to the presence of temporal clustering (Dimitriadis et al., 2016). This provides a scientifically relevant comparison to the empirical data as rainfall series are known to exhibit mild to moderate degree of persistence (e.g. Iliopoulou et al., 2018b; Iliopoulou and Koutsoyiannis, 2019). Moreover, segments of persistent series resemble trends and can easily be misinterpreted as such (Cohn and Lins, 2005).

Therefore, we examine the comparative predictive performance of the four models for persistent processes, where long-term changes are the rule (Serinaldi and Kilsby, 2018), and the effect of available record length on the quality of the model predictions. The latter becomes relevant in the global-moving scheme, in which the calibration period varies in length.

## 4. Results

### 4.1 Models' performance in static validation

Results from the performance of the local mean and local trend models on the last 30 years of each station, as well as on the years preceding the 30-year calibration, are shown in Figure 3 for

296 all studied indices. The local mean model performs on average better than the local trend model
297 for all indices in capturing their most recent changes of extremes, while the performance of the
298 local trend deteriorates considerably with respect to hindcasting the past. Interestingly, the larger
299 discrepancies of the trends —both in future and past validation periods, are encountered in the
300 annual maxima, followed by probability dry. In most of the opposite cases, of trends showing a
301 better performance, the fitted slope is very mild, thus hardly differing from the local mean. A
302 visual examination of the plots of the 60 long-term stations, provided in the Appendix figures
303 (A4-A7), suggests a positive answer to the opening question, providing empirical evidence that
304 climatic trends fluctuate and in fact, abruptly reverse.

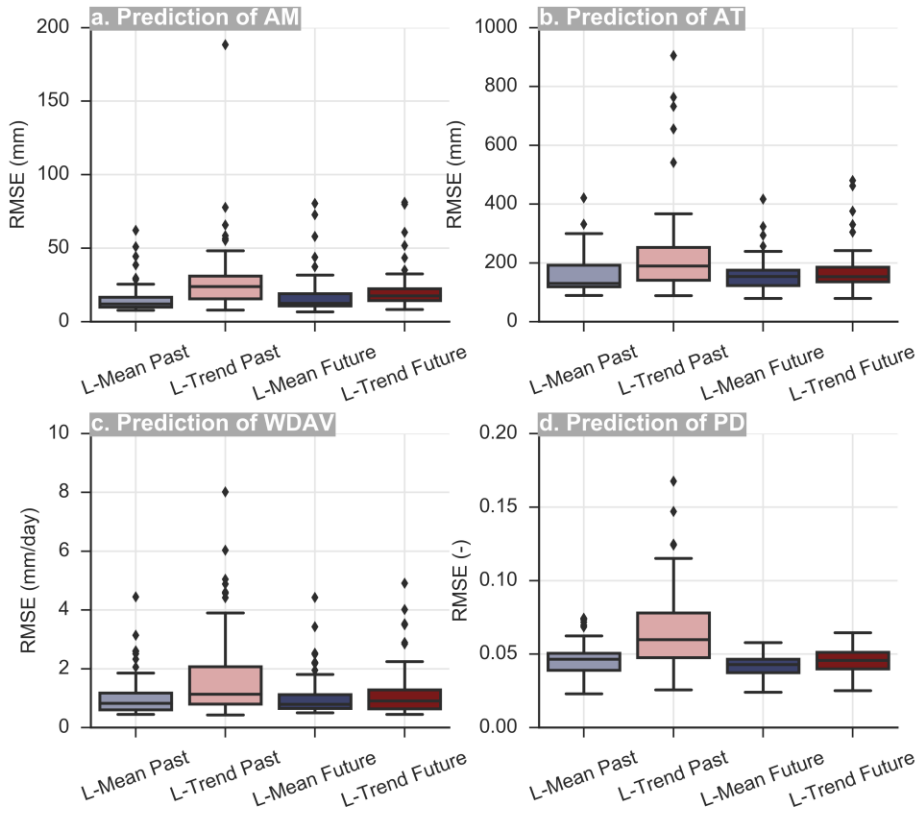



**Figure 3**. Boxplots of the RMSE distribution from the static validation application to all stations, for the local mean (L-Mean) and local trend (L-Trend) models, for all rainfall indices. The band inside the box reports the median of the distribution, the lower and upper ends of the box represent 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.

## 4.2 Moving-window validation of predictive performance

In this section, we explore the predictive qualities of the models by delving into the statistical analysis of the whole record, considering the models from the global-moving calibration as well, namely, the global trend and the global mean.

*4.2.1 An examination of one of the longest records*

As an illustration of the application of the methodology, we first explore the longest uninterrupted station of our dataset, i.e. the Prague station in Czech Republic (211 years), shown in Figure 4. The error propagation pattern of the models is reflective of their performance. For the majority of time, the mean models are at the lower front of the errors, with the local mean model showing slightly superior performance. The local linear trend model results in higher errors and its predictions may quickly deteriorate, taking longer to converge to the mean models in areas of lower errors (Fig. 4). This is attributed to the fact that the trend model projects sensitive features of the calibration period, i.e. extreme observations or 'trendy' behaviour, which do not have a high chance to survive the end of the calibration sample. The more parsimonious structure of the mean model encapsulates minimal but robust knowledge of the process behaviour, which is more likely to characterize its future evolution as well. In the absence of an underlying global trend and as the sample grows larger, the global trend model converges to the predictions of the mean models, but its performance remains slightly inferior even towards the end of the record.

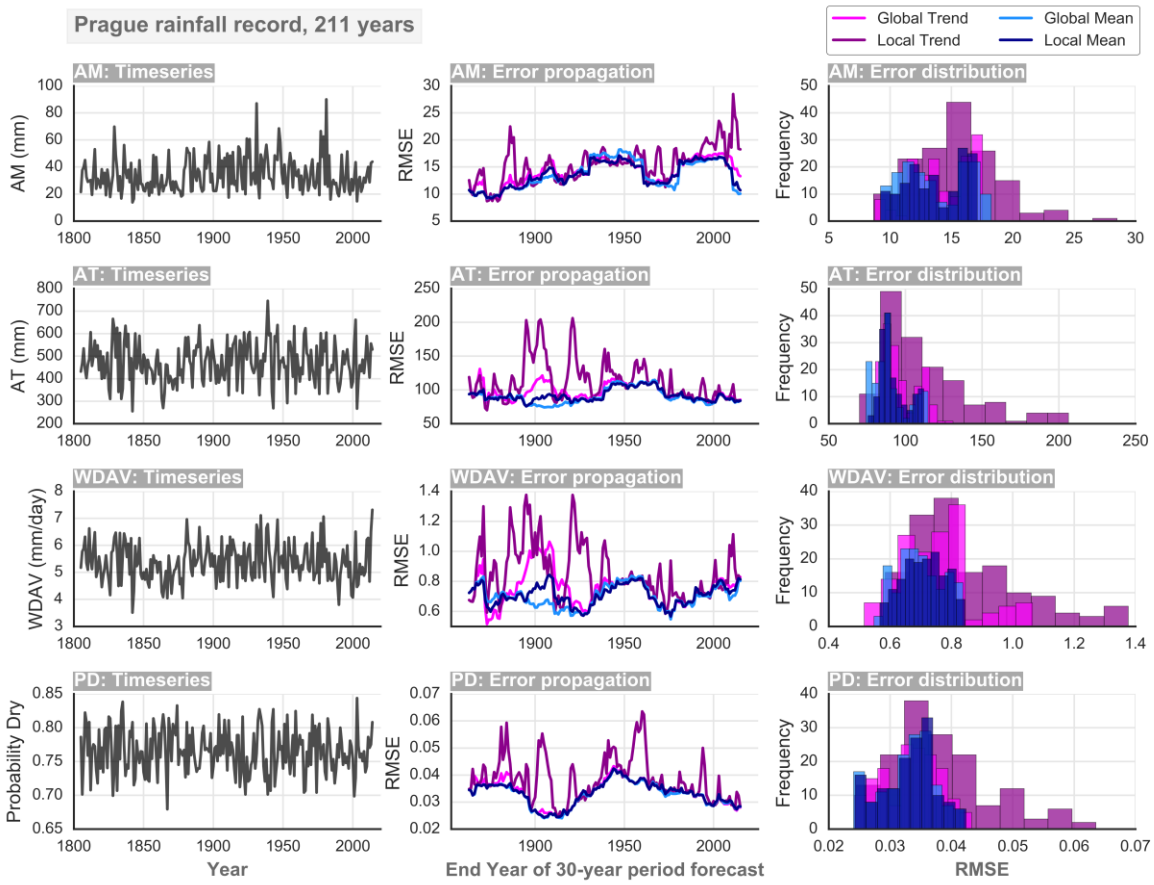



**Figure 4.** Case study of the rainfall station in Prague. Timeseries of annual maxima, annual totals, annual wet-day average and annual probability dry, error propagation and distribution of the prediction RMSE for the four prediction models, global and local trend, and global and local mean.

### *4.2.2 Application to all records*

Figures 5-8 show all 60 stations' empirical distributions of the predictive RMSE of each of model and rainfall index. For most stations the local mean and global mean models have the lower probabilities of exceeding high errors, contrary to the local trend model whose error distribution is clearly shifted to the left, in the higher error area. The distribution of the predictive RMSE of global trend model is located in between the two, showing in general a better behaviour than the local trend.

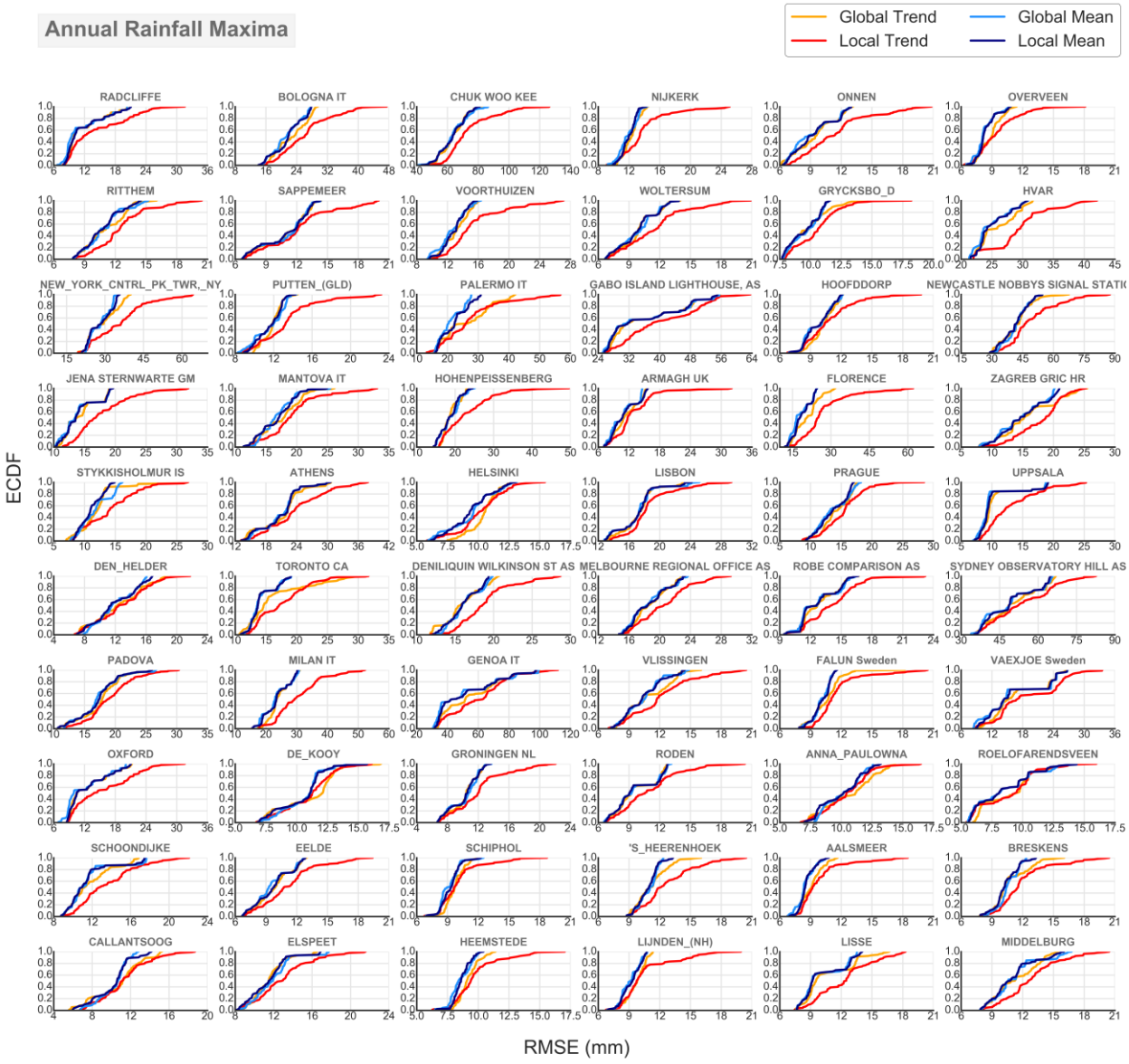
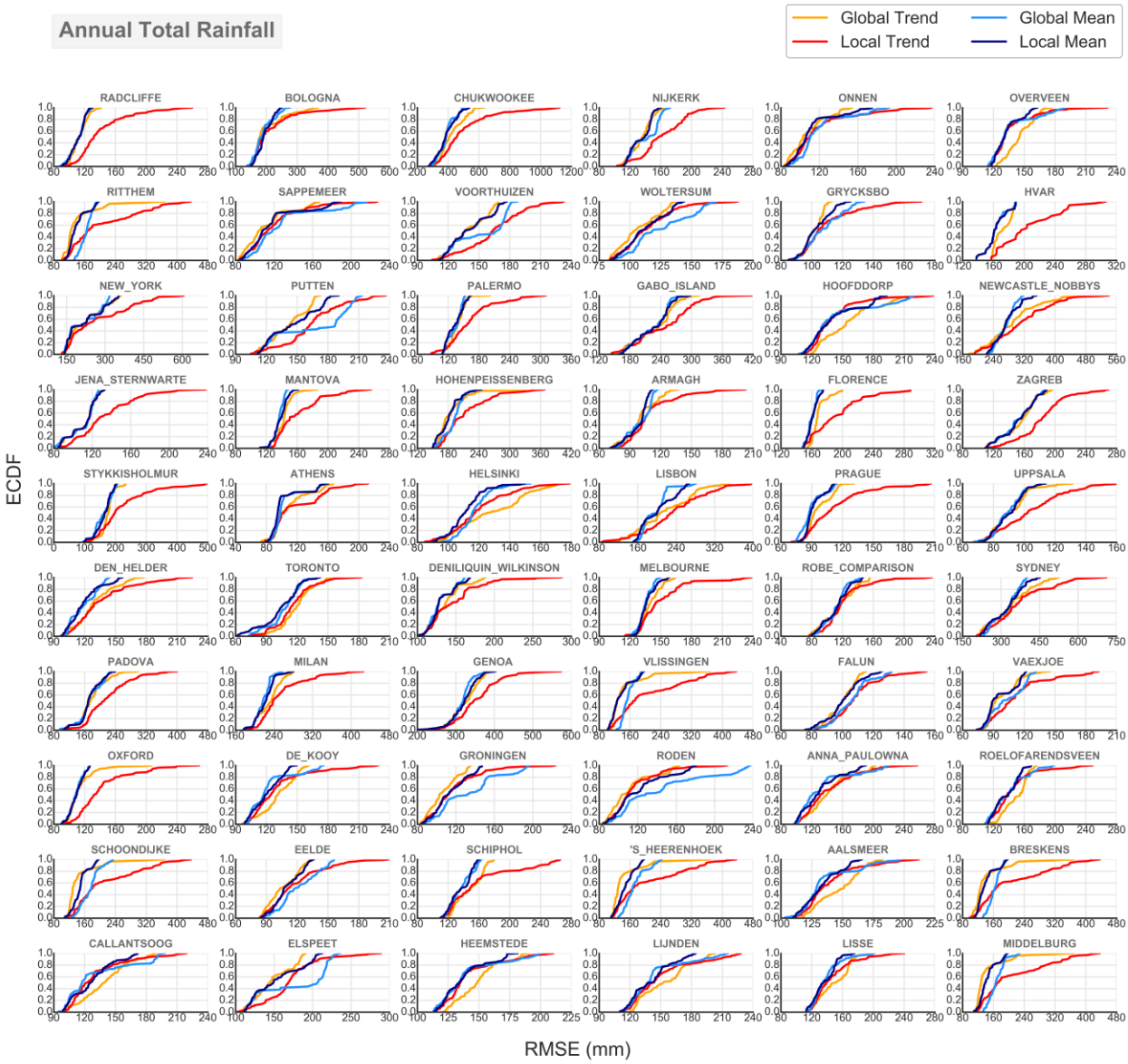


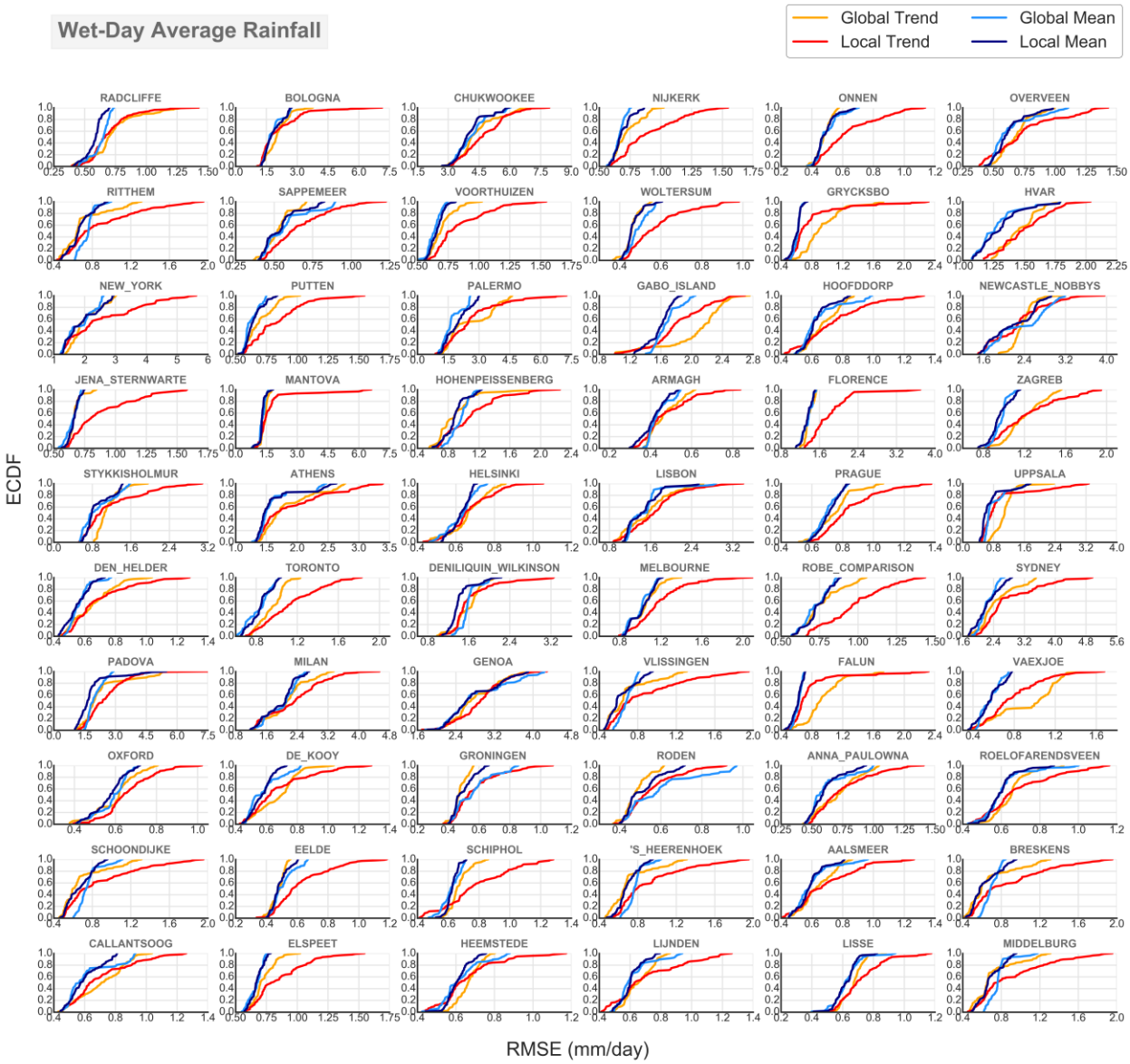

**Figure 5**. Empirical cumulative distribution function (ECDF) for the prediction RMSE of annual maxima for the local trend, the global trend, the global mean and the local mean model for the 60 stations.




350

351 **Figure 6**. Empirical cumulative distribution function (ECDF) for the prediction RMSE of annual
352 totals for the local trend, the global trend, the global mean and the local mean model for the 60
353 stations.

354




355

356 **Figure 7**. Empirical cumulative distribution function (ECDF) for the prediction RMSE of wet-
357 day average rainfall for the local trend, the global trend, the global mean and the local mean
358 model for the 60 stations.

359

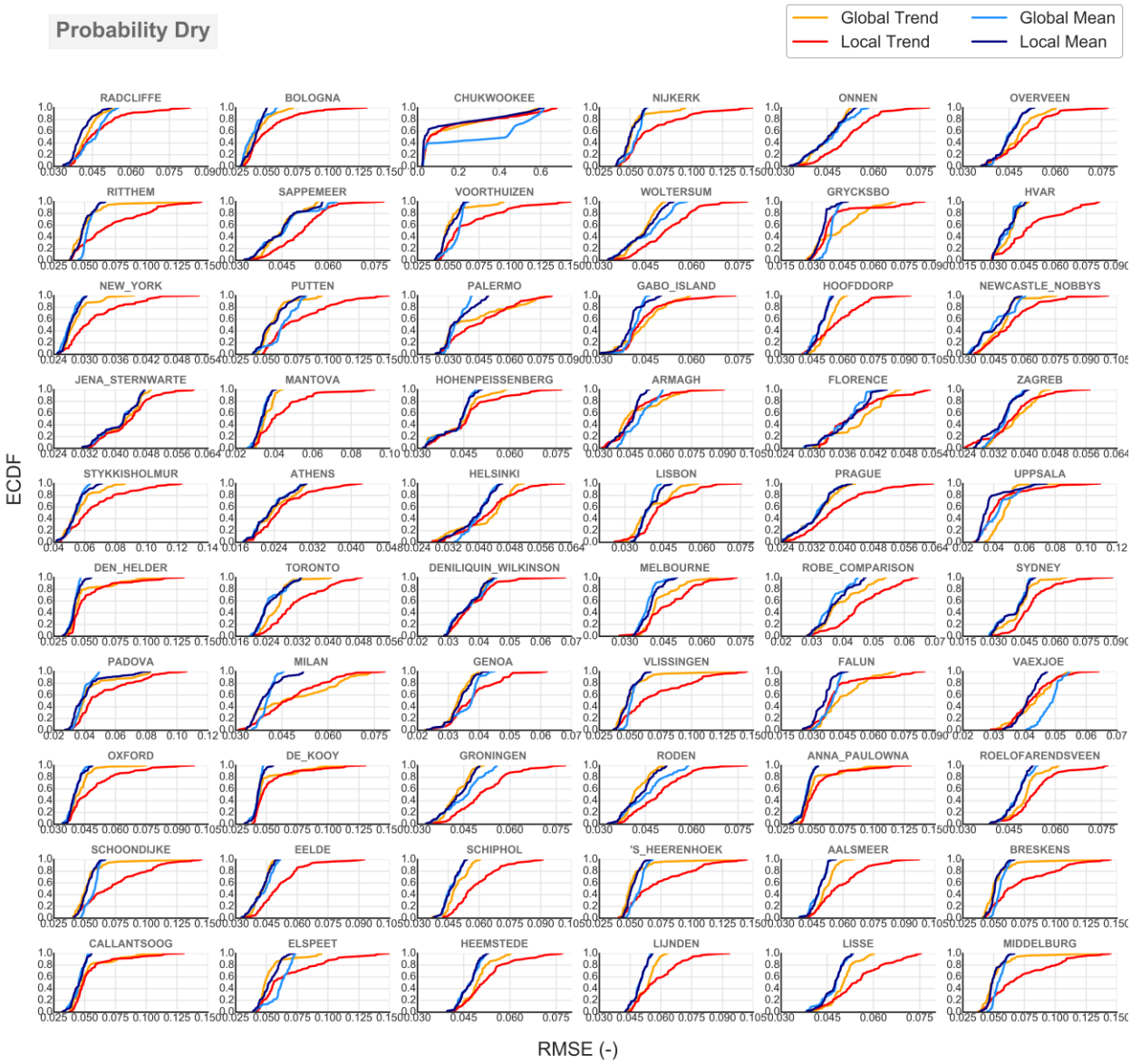
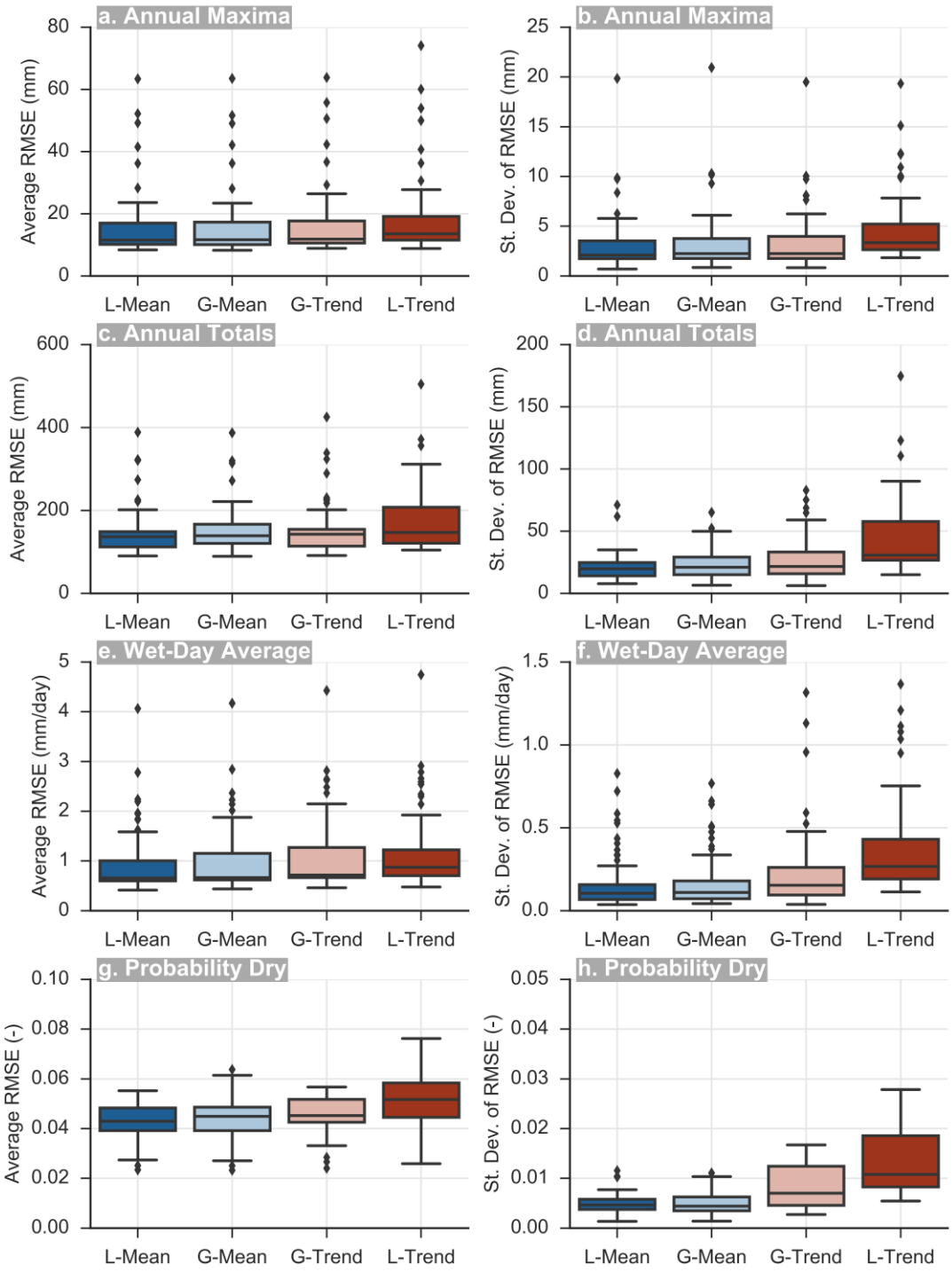



**Figure 8**. Empirical cumulative distribution function (ECDF) for the prediction RMSE of probability dry for the local trend, the global trend, the global mean and the local mean model for the 60 stations.

A summary of the distributional properties of the prediction RMSE of Fig. 5-8, is provided in Fig. 9, in terms of the average and the standard deviation of the RMSE distribution of each station. Accordingly, the models' performance can be ranked from best to worse as follows: (1) local mean, (2) global mean, (3) global trend and (4) local trend. The local mean model marginally outperforms the global mean with respect to the average RMSE, yet in terms of the

standard deviation of the RMSE distribution, it is evident that the local mean model prevails. In this case, the linear trend model shows markedly inferior performance. The sample distribution of the average RMSE and the standard deviation of RMSE from each station's distribution is shown in Fig.9, with the average values of the latter also summarized in Table 1.




373

374 **Figure 9**. Boxplots of the average RMSE and standard deviation of RMSE as estimated for each
375 station from moving window application of the local (L-) mean, global (G-) mean and local (L-)
376 and global (G-) trend for all the indices. The band inside the box reports the median of the
377 distribution, the lower and upper ends of the box represent the 1st and 3rd quartiles, respectively,
378 and the whiskers extend to the most extreme value within 1.5 IQR (interquartile range) from the
379 box ends; outliers are plotted as points.

**Table 1** Averages of the average RMSE and the standard deviation of RMSE of the four models (local (L-) mean, global (G-) mean, local (L-) trend and global (G-) trend) from all stations and for all four indices, as shown in Figure 9.

| | **Annual Maxima** | | | | **Annual Totals** | | | |
|---|---|---|---|---|---|---|---|---|
| | L-mean | G-mean | G-trend | L-trend | L-mean | G-mean | G-trend | L-trend |
| Average RMSE | 16.00 | 16.05 | 16.73 | 18.76 | 149.07 | 154.18 | 154.77 | 174.7 |
| St. Dev. RMSE | 3.04 | 3.13 | 3.37 | 4.74 | 21.52 | 23.02 | 27.4 | 45.45 |
| | **Wet-Day Average** | | | | **Probability Dry** | | | |
| | L-mean | G-mean | G-trend | L-trend | L-mean | G-mean | G-trend | L-trend |
| Average RMSE | 0.98 | 1.01 | 1.11 | 1.2 | 0.04 | 0.05 | 0.05 | 0.05 |
| St. Dev. RMSE | 0.18 | 0.18 | 0.27 | 0.39 | 0.01 | 0.01 | 0.01 | 0.02 |

## 4.3 Models' performance under natural variability

### *4.3.1 An experiment with synthetic series*

Following the rationale outlined in Section 3.5, the goal of this experiment is to test the predictive models in conditions of enhanced structured uncertainty, characterized by changes at all scales and 'trend-like' behaviour for small periods. As the latter are distinctive features of persistent processes (Koutsoyiannis, 2002), we produce five long-term timeseries from a standard normal distribution with length $N$ = 10 000 that reproduce HK dynamics, using the SMA algorithm (Koutsoyiannis, 2000; Dimitriadis and Koutsoyiannis, 2018). The series are generated with increasing degree of persistence, quantified through the Hurst parameter $H$, from mild persistence $H$ = 0.6 to very strong $H$ = 0.99. In order to explore the impact of record length we also examine smaller segments of the same timeseries of lengths $N$ = 100 and $N$ = 1000.

Because smaller segments are impacted by larger estimation uncertainty, we plot the average ECDF of non-overlapping segments extracted from the original timeseries of length $N$ = 10 000. Therefore, the $N$ = 100 plots correspond to the average of 100 timeseries of length 100, derived from the 10 000 series. Likewise, the $N$ = 1000 series are the average of 10 timeseries of length 1000. The plots of the ECDF distribution (Fig.10) of the prediction RMSE for the four prediction models are produced employing the dynamic validation schemes applied for the real-world stations.

The contrasting performance of the two local models is observed here as well; local features are better exploited by the mean rather than the trend model, irrespectively of the record size. The latter becomes important when the global models are considered. In the absence of a global underlying trend, the increased variability encountered in small calibration samples ($N$ = 100) leads the global trend model to bad predictions. When the trend model is calibrated from larger series, the trend component is smoothed out, and therefore, the prediction performance approaches the one from the mean models. Regarding the competition between global and local mean, it appears that it is a function of both the record length and degree of persistence. For large record lengths and $H > 0.7$, the local mean model prevails, while for small record lengths and medium persistence, the two are comparable. In persistent process, where clustering arises, local information is likely to be more relevant for prediction, yet for long-term prediction as is the case here, ‘local’ may need to extend a few steps back in the past, which for small record lengths could be within the reach of the calibration period employed for the global mean model. Obviously though, results from the global model become less relevant when the sample is large and therefore global information extends too far in the past. 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).

We note that the behavior observed in the $N$=100 plots is qualitatively consistent with the one observed from the rainfall records. Moreover, 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 slight 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 in terms of the average RMSE are indistinguishable (Fig. 9); still the variance of the errors being smaller for the latter.

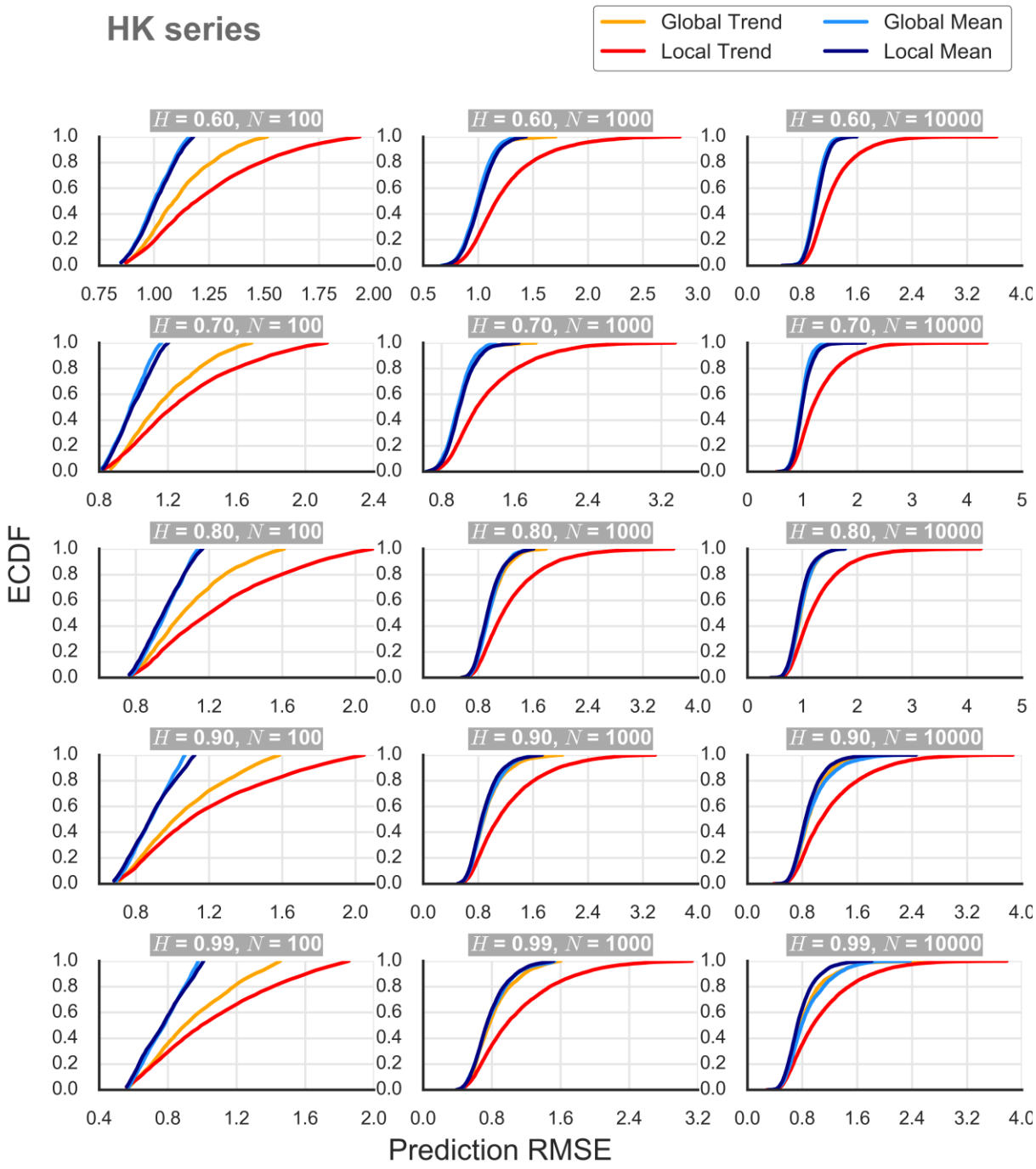



**Figure 10**. Empirical cumulative distribution function (ECDF) for the prediction RMSE of the HK timeseries resulting from application of the local trend, the global trend, the global mean and the local mean model, for segments of the original timeseries with increasing sample size, *N* =100, 1000, 10 000 (original). The ECDF for the first two lengths are the averages as computed from 100 and 10 non-overlapping segments of the 10 000 values.

### *4.3.2 A discussion on parsimony and predictive accuracy*

In the above controlled experiment, where the generating mechanism of the data is known, it is evident that among the four 'false' models, the local mean yields the most accurate predictions in terms of RMSE, using in-sample data more efficiently by means of its single parameter. The increase in predictive accuracy and statistical efficiency is tightly associated with the notion of parsimony, which is a dual criterion measuring the model's fit to the data as well its simplicity (Gauch, 2003). In these terms, the local mean model is deemed to be a parsimonious model, since it fits the out-of-sample data either better or at least equally well to the more complicated trend model.

The reason behind the sometimes interchangeable use of the words parsimony and simplicity, is a certain tendency of simple models to make reliable predictions, which among other approaches, is also incorporated as a concept in Bayesian analysis assigning higher prior probabilities to simpler models, and a posteriori favouring the simpler model (Berger and Bernardo, 1992; Berger and Pericchi, 1996; Gauch, 2003 and references therein). Another demonstration of the relation between predictive accuracy and simplicity is the possibly better predictive performance in terms of mean square error of simpler, yet misspecified models, compared to the ones derived from the correctly structured model (Hocking, 1976); for instance, Wu et al. (2007) provided a set of conditions for which this holds true in the case of linear models. Therefore, theoretical arguments are in favour of simpler predictive models, all the more so in the case of natural processes characterized by great degree of variability, for which our understanding is limited. A comprehensive discussion on the connection of simplicity to wider epistemological and philosophical principles is provided in Gauch (2003).

## 5. Summary and conclusions

A considerable deal of contemporary research in hydrology revolves around the study of temporal changes of extremes, with the application of trend analyses being on the rise during the past two decades (as illustrated in Appendix I). While the explanatory analysis of trends has dominated the relevant studies, assessment of the predictive value of trend models has not been equally assessed, despite the apparent significance of such a task for risk planning. This research reframes the problem of trend evaluation, as a model selection problem oriented towards identifying the model with the best predictive qualities in deterministic terms, which is neither equivalent to the 'true' model nor to the model better at explaining the in-sample data.

For this purpose, we introduce a systematic framework for evaluating projections of trends by means of comparing the prediction RMSE to the one obtained from simple mean models. We perform a variation of cross-validation, also known as walk-forward analysis, devising two distinct calibration and validation schemes (Fig. 2). In block-moving calibration we fit the linear trend and mean models to 30 years of data (local trend and local mean) and we validate the results based on the outcome of their predictions for the next 30 years, repeating the procedure using sliding windows, till the end of the record is met. In global-moving calibration, we fit the models to all the known period (global trend and global mean), assuming that in the beginning, one knows only the first 30 years, and progressively the calibration sample grows larger. In this case too, we evaluate the outcome of the predictions of the models for the next 30 years, therefore the projections of the four models can be compared in terms of the statistics of their empirical distribution of errors.

The models compete in predicting the out-of-sample behaviour of four rainfall indices: annual maxima, annual totals, annual wet-day average rainfall and probability dry at the annual

scale, as estimated from a unique dataset comprising the 60 longest rainfall records surpassing 150 years of daily data. Results show that models rank from best to worst as follows: local mean, global mean, global trend and local trend. A separate examination of the latest 30-year period is also performed in order to track the predictive performance of recent trends. This analysis confirmed the above rank of the models as well. Results from both analyses show that future rainfall variability is better predicted by mean models, since local trend models identify features of the process that are unlikely to survive the end of the calibration sample, either being extreme observations, or 'trend-like' behaviour. These features are smoothed out in longer segments, which is the reason behind the better performance of global trends. Robust regression techniques were also employed for the calibration of local trends but perhaps not surprisingly, did not improve the out-of-sample predictions (see discussion in Appendix III).

In an attempt to reproduce the observed behaviour, we generate long-term timeseries exhibiting long-term persistence or HK dynamics (Koutsoyiannis, 2011; O'Connell et al., 2016; Dimitriadis, 2017), and carry out the same analysis. Persistent processes show enhanced variability and a user unfamiliar with their properties may misinterpret segments of their timeseries as trends, which perhaps explains why trend claims have been that common lately. Results from the synthetic records show qualitative similarities with the ones from empirical rainfall records, known to have persistence, depending on the scale and studied index (Koutsoyiannis, 2006; Markonis and Koutsoyiannis, 2016; Iliopoulou et al., 2018b; Iliopoulou and Koutsoyiannis, 2019). The local and global mean outperform the local trend model for all degrees of persistence and sample sizes, while for small samples ($N$=100) the performance of the global trend model is notably inferior too. Local and global mean models hardly show differences for medium degrees of persistence, but the local mean prevails for strong persistence.

From a systematic investigation of long-term rainfall records, corroborated by simulation results, we have verified that local trends have poor out-of-sample performance, being outperformed in their predictions by simpler models, as the local mean. This empirical finding suggests that the large inherent variability present in the rainfall process makes the practice of extrapolating local features in the long-term future dubious, especially when the complexity of the latter increases. This in turn questions the theoretical and practical relevance of projections of rainfall trends and the grounds of the related abundant publications.

## Acknowledgments

We thank the anonymous reviewers and Associate Editor for providing constructive comments, which resulted in substantial improvements. 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. Academic word occurrence code developed by Strobel (2018), available at http://doi.org/10.5281/zenodo.1218409.

## 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.
Berger, J.O., Bernardo, J.M., 1992. On the development of the reference prior method. Bayesian statistics 4, 35–60.
Berger, J.O., Pericchi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. Journal of the American Statistical Association 91, 109–122.
Biasutti, M., 2019. Rainfall trends in the African Sahel: Characteristics, processes, and causes. Wiley Interdisciplinary Reviews: Climate Change e591.
Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16, 199–231.
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.
Cairncross, A., 1969. Economic forecasting. The Economic Journal 79, 797–812.
Chandler, R., Scott, M., 2011. Statistical methods for trend detection and analysis in the environmental sciences. John Wiley & Sons.
Cohn, T.A., Lins, H.F., 2005. Nature's style: Naturally trendy. Geophysical Research Letters 32.
Conover, W.J., Conover, W.J., 1980. Practical nonparametric statistics.
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.
Dimitriadis, P., Koutsoyiannis, D., 2018. Stochastic synthesis approximating any process dependence and distribution. Stoch Environ Res Risk Assess 32, 1493–1515. https://doi.org/10.1007/s00477-018-1540-2
Dimitriadis, P., Koutsoyiannis, D., Tzouka, K., 2016. Predictability in dice motion: how does it differ from hydro-meteorological processes? Hydrological Sciences Journal 61, 1611–1622.
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.
Fatichi, S., Barbosa, S.M., Caporali, E., Silva, M.E., 2009. Deterministic versus stochastic trends: Detection and challenges. Journal of Geophysical Research: Atmospheres 114.
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.
Gauch Jr, H.G., Gauch, H.G., Gauch Jr, H.G., 2003. Scientific method in practice. Cambridge University Press.
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
Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27, 83–85.

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.
Hinkley, D.V., 1970. Inference about the change-point in a sequence of random variables.
Hocking, R.R., 1976. A Biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics 32, 1–49.
Houghton, J.T., Jenkins, G.J., Ephraums, J.J., 1991. Climate change.
Iliopoulou, T., Koutsoyiannis, D., 2019. Revealing hidden persistence in maximum rainfall records. Hydrological Sciences Journal 1–17.
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.
Inoue, A., Kilian, L., 2006. On the selection of forecasting models. Journal of Econometrics 130, 273–306.
Inoue, A., Kilian, L., 2005. In-sample or out-of-sample tests of predictability: Which one should we use? Econometric Reviews 23, 371–402.
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.
Kirkpatrick II, C.D., Dahlquist, J.A., 2010. Technical analysis: the complete resource for financial market technicians. FT press.
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.
Koenker, R., Hallock, K.F., 2001. Quantile regression. Journal of economic perspectives 15, 143–156.
Koutsoyiannis, D., 2020. Stochastics of hydroclimatic extremes: a cool look at risk (in preparation).
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., 2003. Climate change, the Hurst phenomenon, and hydrological statistics. Hydrological Sciences Journal 48, 3–24.
Koutsoyiannis, D., 2002. The Hurst phenomenon and fractional Gaussian noise made easy. Hydrological Sciences Journal 47, 573–595.
Koutsoyiannis, D., 2000. A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. Water Resources Research 36, 1519–1533.

626 Koutsoyiannis, D., Montanari, A., 2015a. Negligent killing of scientific concepts: the stationarity
627 case. Hydrological Sciences Journal 60, 1174–1183.
628 Koutsoyiannis, D., Montanari, A., 2015b. Negligent killing of scientific concepts: the stationarity
629 case. Hydrological Sciences Journal 60, 1174–1183.
630 Koutsoyiannis, D., Montanari, A., 2007. Statistical analysis of hydroclimatic time series:
631 Uncertainty and insights. Water resources research 43.
632 Kumar, V., Jain, S.K., Singh, Y., 2010. Analysis of long-term rainfall trends in India.
633 Hydrological Sciences Journal–Journal des Sciences Hydrologiques 55, 484–496.
634 Kutiel, H., Trigo, R.M., 2014. The rainfall regime in Lisbon in the last 150 years. Theoretical
635 and applied climatology 118, 387–403.
636 Laio, F., Di Baldassarre, G., Montanari, A., 2009. Model selection techniques for the frequency
637 analysis of hydrological extremes. Water Resources Research 45.
638 Lorenc, A.C., 1986. Analysis methods for numerical weather prediction. Quarterly Journal of the
639 Royal Meteorological Society 112, 1177–1194.
640 Marani, M., Zanetti, S., 2015. Long-term oscillations in rainfall extremes in a 268 year daily time
641 series. Water Resources Research 51, 639–647.
642 Markonis, Y., Koutsoyiannis, D., 2016. Scale-dependence of persistence in precipitation records.
643 Nature Climate Change 6, 399–401.
644 McCarl, B.A., Villavicencio, X., Wu, X., 2008. Climate change and future analysis: is
645 stationarity dying? American Journal of Agricultural Economics 90, 1241–1247.
646 McKitrick, R., Christy, J., 2019. Assessing Changes in US Regional Precipitation on Multiple
647 Time Scales. Journal of Hydrology 124074.
648 Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G., 2012. An Overview of the
649 Global Historical Climatology Network-Daily Database. J. Atmos. Oceanic Technol. 29,
650 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1
651 Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier,
652 D.P., Stouffer, R.J., 2008. Stationarity is dead: Whither water management? Science 319,
653 573–574.
654 Milly, P.C., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier,
655 D.P., Stouffer, R.J., Dettinger, M.D., Krysanova, V., 2015. On critiques of "Stationarity
656 is dead: Whither water management?" Water Resources Research 51, 7785–7789.
657 Mitchell, W.C., 1930. Business cycles: the problems and its setting Business cycles: The
658 problem and its setting. National Bureau of Economic Research, New York.
659 Modarres, R., da Silva, V. de P.R., 2007. Rainfall trends in arid and semi-arid regions of Iran.
660 Journal of arid environments 70, 344–355.
661 Montanari, A., Koutsoyiannis, D., 2014. Modeling and mitigating natural hazards: Stationarity is
662 immortal! Water Resources Research 50, 9748–9756.
663 Moss, R.H., Edmonds, J.A., Hibbard, K.A., Manning, M.R., Rose, S.K., Van Vuuren, D.P.,
664 Carter, T.R., Emori, S., Kainuma, M., Kram, T., 2010. The next generation of scenarios
665 for climate change research and assessment. Nature 463, 747.
666 Ntegeka, V., Willems, P., 2008. Trends and multidecadal oscillations in rainfall extremes, based
667 on a more than 100-year time series of 10 min rainfall intensities at Uccle, Belgium.
668 Water Resources Research 44.
669 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.

Papoulis, A., 1990. Probability & statistics. Prentice-Hall Englewood Cliffs.

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

Persons, W.M., 1922. Measuring and Forecasting General Business Conditions. American institute of finance.

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., n.d. 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.

Santer, B.D., Wigley, T.M.L., Boyle, J.S., Gaffen, D.J., Hnilo, J.J., Nychka, D., Parker, D.E., Taylor, K.E., 2000. Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. Journal of Geophysical Research: Atmospheres 105, 7337–7356.

Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. Journal of the American statistical association 63, 1379–1389.

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.

Sharma, P.N., Shmueli, G., Sarstedt, M., Danks, N., Ray, S., 2019. Prediction-oriented model selection in partial least squares path modeling. Decision Sciences.

Shibata, R., 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. The annals of statistics 147–164.

Shmueli, G., 2010. To explain or to predict? Statistical science 25, 289–310.

Simonoff, J.S., 2012. Smoothing methods in statistics. Springer Science & Business Media.

Slutsky, E.E., 1927. Slozhenie sluchainykh prichin, kak istochnik tsiklicheskikh protsessov. Voprosy kon''yunktury 3, 34–64. (English edition: Slutzky, E., 1937. The summation of random causes as the source of cyclic processes. Econometrica: Journal of the Econometric Society, 105-146).

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.

Stein, R.M., 2002. Benchmarking default prediction models: Pitfalls and remedies in model validation. Moody's KMV, New York 20305.

Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society: Series B (Methodological) 39, 44–47.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) 36, 111–133.

Strobel, V (2018, April 14). Pold87/academic-keyword-occurrence: First release (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.1218409

Theil, H., 1992. A rank-invariant method of linear and polynomial regression analysis, in: Henri Theil's Contributions to Economics and Econometrics. Springer, pp. 345–381.

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.

Wei, C.-Z., 1992. On predictive least squares principles. The Annals of Statistics 20, 1–42.

Wu, S., Harris, T.J., McAuley, K.B., 2007. The use of simplified or misspecified models: Linear case. The Canadian Journal of Chemical Engineering 85, 386–398.

Yarkoni, T., Westfall, J., 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science 12, 1100–1122.

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

## I. A brief quantitative literature review

The aim of this literature review is to evaluate the academic interest in trends of rainfall variables by means of a quantitative analysis of research papers appearing in Google Scholar. We base this analysis on the quantification of the occurrence of associated words in Google Scholar using Python code developed by Strobel (2018), omitting results related to citations and patents. This analysis was performed on 21/10/2019 and in order to refer to full calendar years it contains results published till the end of 2018.

In Fig. A1, we show the temporal evolution of the ratio of appearance of the word 'trends' in items also containing the complete list of words ['precipitation', 'hydrology', 'extremes']. Results have been randomly varying from the beginning till the mid 20$^{th}$ century, when there were less than 100 results per year fulfilling the criteria of containing the list in the denominator of the ratio. It can be seen though that approximately from the 1960 and later on there has been an increasing trend of publications containing the word 'trends,' reaching 89% in 2018. Obviously, results belonging to a different context than the one assumed might have been calculated as well but we assume their effect to be analogous both in the nominator and the denominator of the ratio, thus not significantly affecting the conclusion.

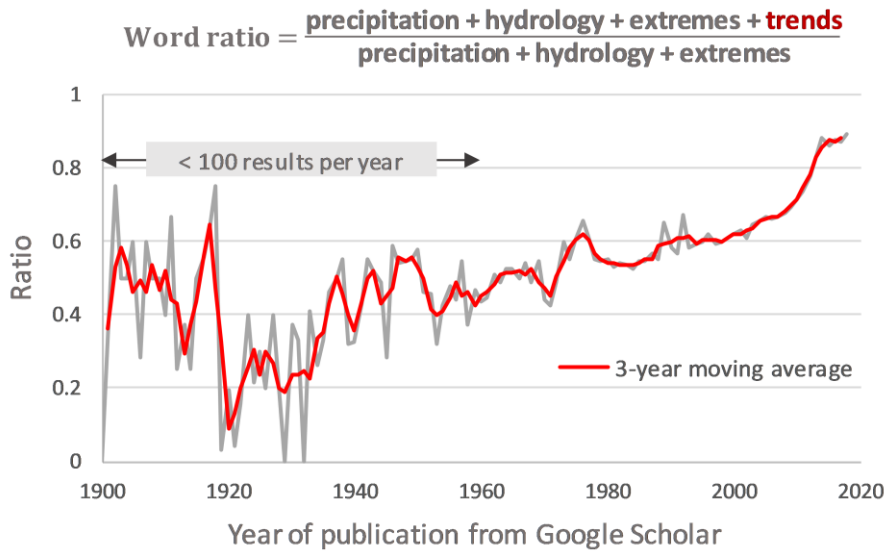



**Figure A1**. Temporal evolution along with three-year moving average of the ratio of the occurrence of the word 'trends' in Scholar items containing the words 'precipitation', 'hydrology' and 'extremes'.

To further refine our search to more technical papers explicitly referring to rainfall trends we define the following search items. Word combination A is the full list ['precipitation|rainfall trends', 'precipitation|rainfall data|records'], where the symbol | refers to 'or', and word combinations inside '' should be found together, i.e. one possible combination is the list ['precipitation trends', 'rainfall data']. Word combination B is an extension of word combination A that also includes the word 'projections', while word combination C is an extension of word combination A also including the word sequence 'linear trend|trends|model|regression'. The absolute numbers of the results are shown in Fig. A2a, while in Fig.A2b we show their relative ratio. Expectedly, the total number of studies containing rainfall trends are rising, however this is not surprising in terms of absolute numbers, considering the increasing availability of papers in Scholar over the years. However, the use of the word 'projections' appears to be increasing in relative terms as well. The rate of word use in relation to the linear trend (C) has slightly

increased too over the years, stabilizing over the past 5-year period to approximately half of the related publications (Fig.A2b).

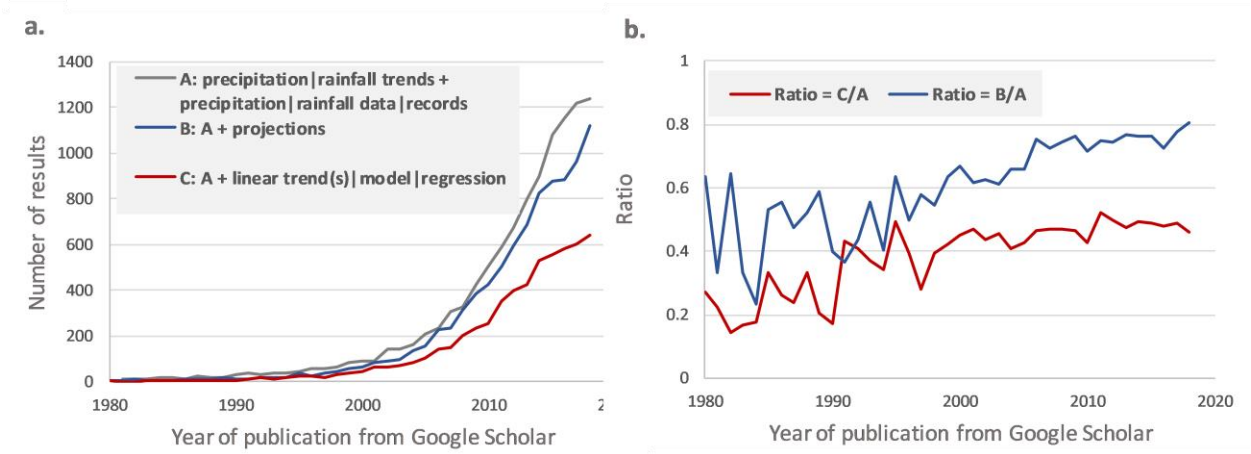
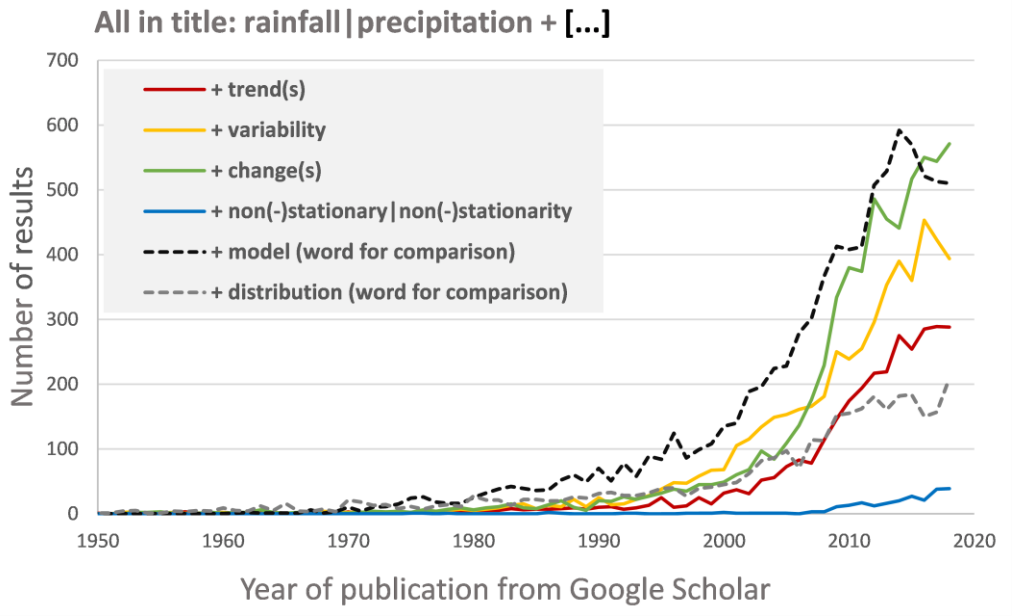



**Figure A2**. (a) Temporal evolution of the occurrence of the word combinations A, B and C and their relative ratio (b).

As a final refinement, we consider words appearing only in the title of papers, which should limit the results to strictly related papers. Results are shown in Fig. A3. The standard term that is contained in every result is 'rainfall|precipitation' followed by the appearance, anywhere in the title, of the single terms, trends|trend, variability, change|changes, and non-stationary|non-stationarity|nonstationary|nonstationarity. Note that we consider also plural terms where applicable, as well as possible differences in spelling, while this time, we do not require words to be found in a specific order as in the previous in-text search (for instance, it could be "**trends** in **rainfall**..." or "**rainfall trends** in the.."). We do not compute ratios over the items containing in their title the words 'rainfall|precipitation' because these terms alone are too generic, and can be found in a variety of studies, a significant part of which are only loosely related to hydrology (e.g. physics, chemistry, radar technologies etc.). Instead, to provide a more relevant reference point for comparison, we use two words semantically 'uncharged' with the trend concept, which

804 are however widely used in combination with the standard terms, namely the words 'model' and
805 'distribution' (e.g. "a **rainfall model**…" or "the **distribution** of the … **precipitation**").

806 Apparently, the conceptually more inclusive terms 'changes' and 'variability' are ranking
807 first in the related search terms, with the explicit use of the word 'trend(s)' ranking third,
808 yielding consistently over the last ten years above 200 results per year (288 in 2018, as per
809 results appearing on Google Scholar on 21/10/2019). Terms related to non-stationarity are slowly
810 rising over the past ten years (39 in-title results in 2018), while being close to zero before 2000.
811 It is interesting to note the evolution of the use of terms explicitly associated with the temporal
812 properties of rainfall compared to the terms more related to marginal properties ('distribution'),
813 or being more of a general use, perhaps implying both properties ('model'). The mere use of the
814 word 'trend(s)' has exceeded the use of an all-times classic word for rainfall, i.e. distribution,
815 which clearly shows a certain shift in academic interest. Likewise, the ever higher-scoring word
816 'model' has been outnumbered in the past three years by the word 'change(s)'.




817

818 **Figure A3**. Temporal evolution of the occurrence of the word combinations in titles of Scholar
819 items.

In conjunction, these results suggest that over the last two decades, there has been a rising scientific interest in the temporal properties of rainfall and their future evolution, with ‘trends’ taking up a considerable share of this emerging focus.

## II. Rainfall records properties and long-term variability

Table 1 summarizes the properties of the long-term rainfall stations. In Fig A4-A7, we illustrate the static validation scheme showing results from the projections of the local trend and the local mean model for all rainfall indices.

**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 |

836

837

838

839


**Figure A4**. Local trend vs the local mean in projecting annual maxima for the 60 longest rainfall stations.


**Figure A5**. Local trend vs the local mean in projecting annual totals for the 60 longest rainfall stations.

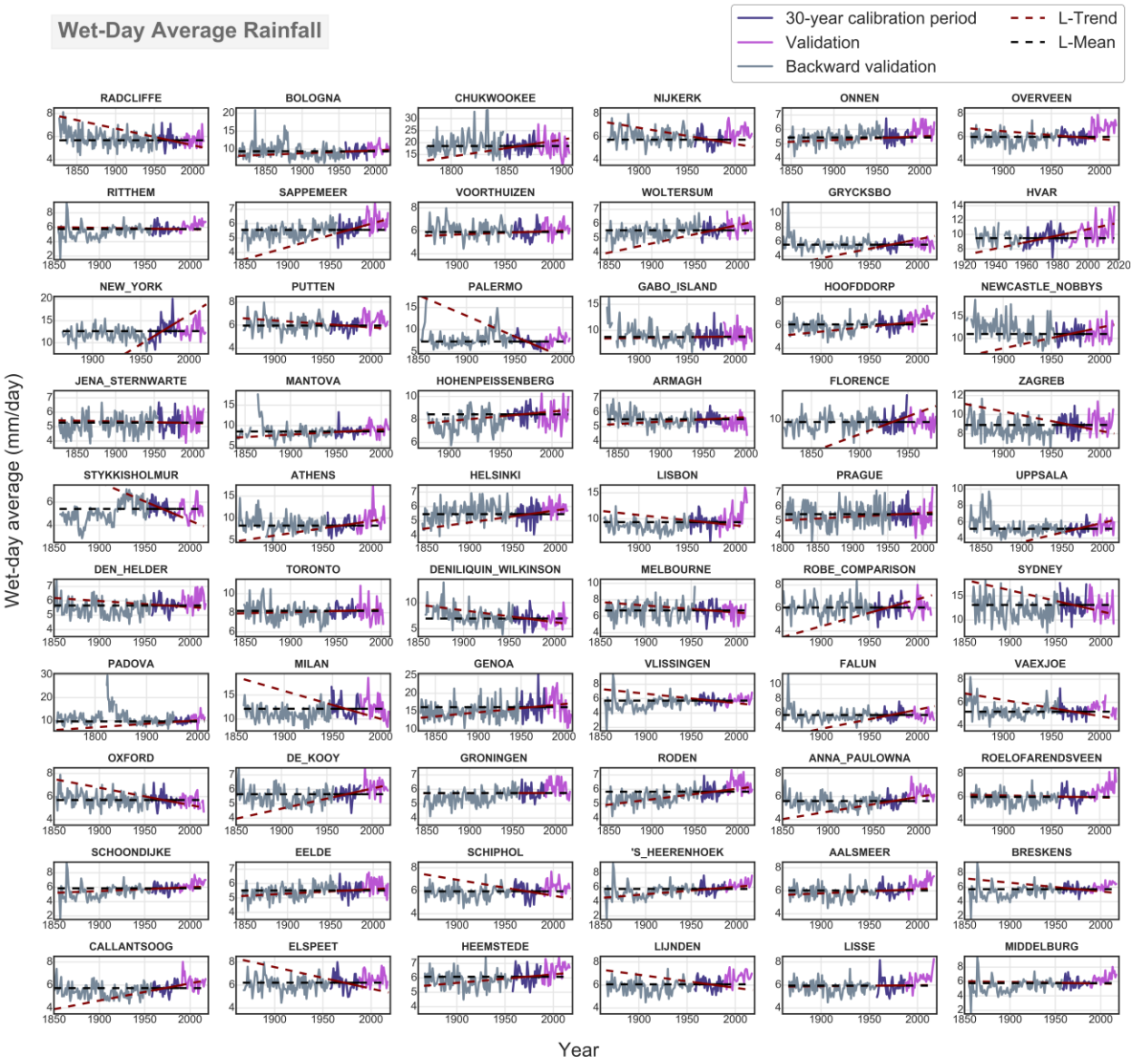



**Figure A6**. Local trend vs the local mean in projecting wet-day average rainfall for the 60 longest rainfall stations.

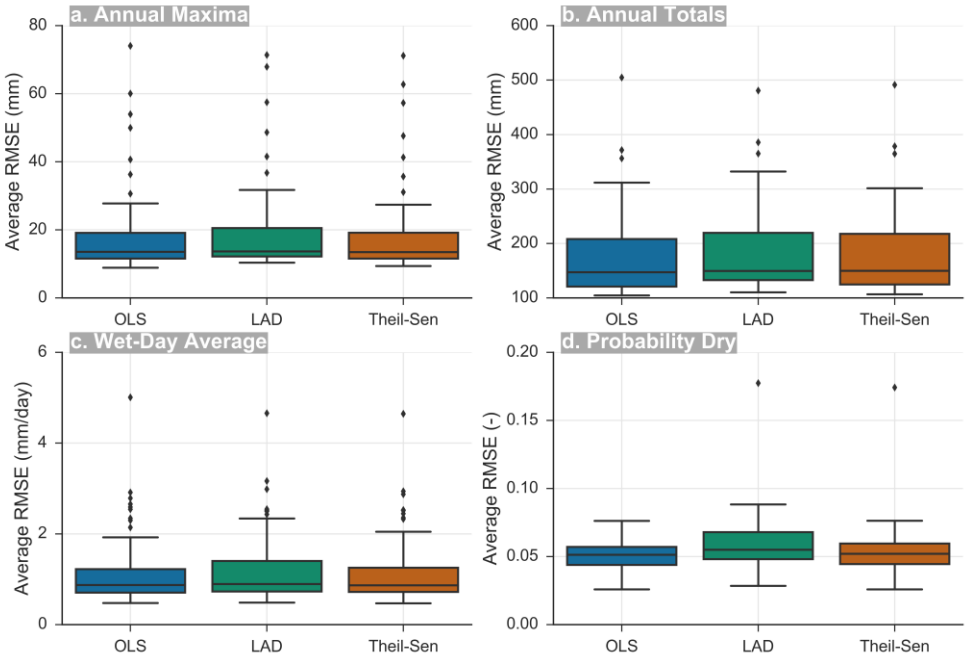

**Figure A7**. Local trend vs the local mean in projecting probability dry for the 60 longest rainfall stations.

## III. Fitting algorithms: Least-squares vs robust regression

We explore the effect of the linear trend definition and fitting algorithm on the results of the local trends, as trends in small segments are expected to be more sensitive to the choice of the fitting algorithm (Santer et al., 2000). The first algorithm is the widely used ordinary least-square estimation (OLS), which fits Equation 1 to the data, by minimizing the sum of the squares of the differences between the observed and those predicted by the linear model. Secondly, two alternative trend calibration approaches are explored that place less weight on influential observations (outliers) and thus belong to the range of 'robust regression' techniques. The first is least absolute deviations (LAD), which estimates the regression coefficients by minimising the sum of absolute deviations of the predicted from the observed values, and can be shown to be a special case of quantile regression, fitting the trend line to the median of the observations, rather than the mean (Chandler and Scott, 2011). The second is the non-parametric method of Theil-Sen slope estimation (Sen, 1968; Theil, 1992), which estimates the slope $b$ of the linear model as the median of the pairwise slopes of all sample points. Among the different approaches that exist for the intercept coefficient, we follow Conover (1980) and estimated the intercept as $a = y_{0.5} - bx_{0.5}$, where $y_{0.5}$ and $x_{0.5}$ the sample medians.

Results from the comparison of the prediction RMSE from these three algorithms are shown in Figure A8. Evidently, the ordinary least square estimator performs better than the LAD estimation, while its results are very close to the Theil-Sen estimation. Therefore, the OLS estimator is retained for the main analysis due to its better performance compared to LAD, non-ambiguity in definition compared to the Theil-Sen estimator, and well-studied mathematical properties (Papoulis, 1990). As a final note, we underline that the notion of 'robustness' of statistical regression has arisen as a positive trait for systems with known and expected behaviour, where extreme values are considered either 'outliers' or erroneous measurements,

897 which “contaminate” the record. Yet for natural systems, producing extremes as part of a large
898 and inherent variability, and exhibiting irregular ‘trends’ difficult or perhaps impossible to
899 attribute to causal mechanisms, we deem that there might be no theoretical reason behind the
900 expected superiority of robust statistics, which is in fact empirically shown in this experiment.




901

902 **Figure A8**. Boxplots of the average RMSE as estimated for each station from moving window
903 application of the local trend using Least Squares regression (LS), least absolute deviation
904 regression (LAD) and the Theil-Sen regression. The band inside the box reports the median of
905 the distribution, the lower and upper ends of the box represent the 1st and 3rd quartiles,
906 respectively, and the whiskers extend to the most extreme value within 1.5 IQR (interquartile
907 range) from the box ends; outliers are plotted as points.