On the long-range dependence properties of annual precipitation using a global network of instrumental measurements

Hristos Tyralis*, Panayiotis Dimitriadis, Demetris Koutsoyiannis, Patrick Enda O'Connell, Katerina Tzouka and Theano Iliopoulou

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

*Corresponding author, montchrister@gmail.com

Abstract: The long-range dependence (LRD) is considered an inherent property of geophysical processes, whose presence increases uncertainty. Here we examine the spatial behaviour of LRD in precipitation by regressing the Hurst parameter estimate of mean annual precipitation instrumental data which span from 1916-2015 and cover a big area of the earth's surface on location characteristics of the instrumental data stations. Furthermore, we apply the Mann-Kendall test under the LRD assumption (MKt-LRD) to reassess the significance of observed trends. To summarize the results, the LRD is spatially clustered, it seems to depend mostly on the location of the stations, while the predictive value of the regression model is good. Thus when investigating for LRD properties we recommend that the local characteristics should be considered. The application of the MKt-LRD suggests that no significant monotonic trend appears in global precipitation, excluding the climate type D (snow) regions in which positive significant trends appear.

Keywords: Hurst; long-range dependence; Mann-Kendall test; precipitation; random forests; trend analysis

1. Introduction

The long-range dependence (LRD), also known in hydrological science as the Hurst phenomenon, is a behaviour observed in geophysical processes in which wet years or dry years are clustered to respective long time periods (Koutsoyiannis, 2002). A common practice for evaluating the presence of the LRD is to model the geophysical time series with the Hurst-Kolmogorov process (HKp) and estimate its Hurst parameter

H (Koutsoyiannis, 2003; Tyralis and Koutsoyiannis, 2011) where high values of *H* indicate strong LRD.

The estimation of *H* is of great importance in engineering practice (Lins and Cohn, 2011). As indicated by Koutsoyiannis (2006), Koutsoyiannis and Montanari (2007) and Tyralis and Koutsoyiannis (2014) the uncertainty increases substantially when LRD is present. Furthermore, due to the increase in uncertainty, observed trends in data, even if they seem significant using classical statistical testing, can be insignificant under the LRD assumption as shown by Hamed (2008).

Most studies on the assessment of the magnitude of precipitation LRD using instrumental data are local (e.g. Liu et al., 2012; Munshi, 2015; Valle et al., 2013). However, some studies including Fatichi et al. (2012) and Iliopoulou et al. (2017) estimated the magnitude of the precipitation LRD from instrumental measurements in global spatial scale and argued for its weak existence although the evidence for its presence in annual precipitation records is inconclusive (O'Connell et al., 2015). Similar global studies based on dissimilar datasets include Kumar et al. (2013) who estimated the *H* parameter of Coupled Model Intercomparison Project (CMIP5) twentieth-century precipitation simulations, Sun et al. (2014) who used reanalysis datasets and Bunde et al. (2013) who used instrumental measurements, climate model simulations and precipitation reconstructions to infer the significance of LRD in precipitation.

The Mann-Kendall test is frequently used in hydrology to evaluate the significance of trends. However, the Mann-Kendall test under the LRD assumption (MKt-LRD) (Hamed, 2008), in which a possible presence of LRD is considered, has been less frequently adopted. A few local case studies, in which the authors applied the Mann-Kendall test considering the presence of LRD include the investigation of precipitation (Dinpashoh et al., 2014), stream flows (Ehsanzadeh and Adamowski, 2010; Khaliq et al., 2009; Kumar et al., 2009; Sagarika et al., 2014; Zamani et al., 2017) and both (Fathian et al., 2016).

The analysis of point precipitation at a global setting is an important topic in hydrology (for instance see de Lima et al., 2012). It can be supported by the analysis of precipitation instrumental data from stations that spatially cover the globe, which has become a common subject in the recent literature and is supported by the increasing availability and accessibility of global data sets (Bierkens, 2015) while it is an important constituent of global-scale hydrology whose emergence was highlighted by Eagleson

(1986; 1994). Such studies include the analysis of extremes (Alexander et al., 2006; Asadieh and Krakauer, 2015; Koutsoyiannis, 2004; Papalexiou and Koutsoyiannis, 2013), droughts (Nasrollahi et al., 2015), analysis of trends (van Wijngaarden and Syed, 2015), the temporal concentration of precipitation (Monjo and Martin-Vide, 2016) and reconstruction of past precipitation (Smith et al., 2012). Although the instrumental data need some processing to be used, they could be considered more reliable compared to climate simulations or reconstructions. However, the coverage of the earth's surface by rain gauges is not high, while it decreases considerably when the analysis demands a sufficient long time period to obtain more reliable results (New et al., 2001). In such cases, several alternative methods have been proposed including the use of satellite data (Kidd and Huffman, 2011).

The spatial analysis of precipitation based on instrumental measurements can be applied in local case studies, because the areas of interest are uniformly covered by the stations. This is the case, e.g. in Blanchet et al. (2009) who study the extreme statistics of snowfall, Villarini and Smith (2010) who investigate flood peak distributions, Li et al. (2011) who study precipitation trends and Dyrrdal et al. (2016) who analyse the extreme precipitation.

In this study, we estimate the H parameter of the mean annual precipitation from instrumental data from a large part of the earth. The database used in this study (Menne et al., 2012a,b) includes stations that cover the largest part of the inhabited earth surface. However, for statistical reasons we examine stations with data, which span the hundred-year period 1916-2015 and thus the coverage decreases considerably. However, we prefer to use this reduced dataset instead of reanalysis datasets, because the artificial nature of the latter can alter considerably the results, particularly when using reanalysis data from uncovered areas at early time periods.

The primary aim of our study is to investigate the relationship between H and locations features, which has been suggested for further research in Iliopoulou et al. (2017), while Fatichi et al. (2012) did not identify the presence of a particular geographical pattern. The results of Sun et al. (2014) and Markonis and Koutsoyiannis (2016, Figure S3) indicate that H varies considerably with the location of the stations; however they were obtained by reconstructions of past precipitation. Classical spatial statistical analysis cannot be applied, because the coverage of the earth's surface by the examined stations is low and strongly non-uniform. In such cases machine learning

methods are a useful alternative (e.g. Alobaidi et al., 2015; Leuenberger and Kanevski, 2015), as well as a complementary option (Kanevski and Demyanov, 2015). Therefore, to overcome the problem of non-uniform coverage an alternative approach is to regress the *H* parameter estimates on spatial characteristics of the stations, i.e. their coordinates. However, location characteristics of the stations such as their elevation and their Köppen-Geiger climate class (Kottek et al., 2006) may also be related to *H*. The ability to include predictor variables beyond the coordinates is another advantage of machine learning methods compared to classical statistical spatial analysis methods. To find all possible relationships, we apply both linear regression models and random forests algorithms (Breiman, 2001). The latter are classified as machine learning methods and they are particularly useful to model non-linear relationships between the dependent and the predictor variables, even when the latter are correlated. Furthermore, they have been applied in spatial analyses with better results compared to other machine learning methods (Cracknell and Reading, 2014).

A secondary aim of our study is to assess the significance of precipitation trends by applying the MKt-LRD test along with an exploratory analysis, in which we can present the relationship between the magnitude and significance of trends and the location characteristics. Morin (2011) also applied the MKt to the mean annual precipitation, but using a gridded dataset and pre-whitening to account for the presence of serial correlation. Additionally, Van Wijngaarden and Syed (2015) already examined the precipitation trends using nearly 1 000 stations for the time period 1700-2013. They assessed the significance of the trends using the statistical t-test at the 5% level and they concluded that "*some caution is warranted about claiming that large changes to global precipitation have occurred during the last 150 years*". Zhang et al. (2007) also examined the trends in global precipitation using instrumental measurements, and compared them to projections of climate models with the aim to find how anthropogenic forcing influences the precipitation.

In most studies, global precipitation and possible changes in the hydrological cycle are examined by extended use of General Circulation Models (for instance Allan et al., 2014; Allen and Ingram, 2002; Gu and Adler, 2015), or gridded datasets (for instance Hartmann et al., 2013; Gu and Adler, 2015; Morin, 2011). Climate models' outputs are not used here because they contain large errors in precipitation simulations (Trenberth, 2011). Gridded datasets are not used here because they are constructed using

observations of temporally varying spatial density, which may lead to inaccurate estimates of parameters used in climate modelling (Beguería et al., 2016; Sun et al., 2014), while they may also be based on climate models (reanalysis datasets). In particular, the present study examines large areas, thus temporal variations of the spatial density of the observations would be even higher. Instead, here we use a set of stations with constant spatial density during the study period. Furthermore, due to the long time period only instrumental measurements can provide sufficient data for the investigation of possible trends (New et al., 2001).

There is also a discussion on the relationship between the precipitation and temperature changes with the aim to find constraints of the maximum precipitation change conditionally as a function of temperature change with contradictory results. Allan et al. (2014) estimate an increase in precipitation approximately 2-3%/K, Lambert et al. (2008) estimated an increase approximately equal to 6%/K, while Wentz et al. (2007) estimated an increase equal to 7%/K.

The assessment of the significance of trends, under the assumption of LRD has been proposed for research by Morin (2011). The significance of trends has been also examined using instrumental measurements in Van Wijngaarden and Syed (2015) and using gridded datasets in Hartmann et al. (2013); however, they did not use the LRD assumption. There have been attempts to link the precipitation changes with the temperature; however, the statistical significance is a concept, which cannot be incorporated in such framework. A trend may be small but can be statistically significant and vice versa. Furthermore, Koutsoyiannis and Montanari (2014) noticed that geophysical processes can be modelled using non-stationary models if there is a well-defined deterministic relationship constructed by deduction. Therefore, changes in precipitation related to temperature, could be useful in such setting, i.e. if a deterministic function for the changes of the temperature could be also constructed by deduction. In our opinion, this is not possible (see also Koutsoyiannis, 2010). On the other hand, the assessment of the significance of trends can be incorporated in a stationary and completely stochastic framework, because it can test whether the particular stochastic framework is suitable to model the geophysical process of interest.

2. Data

We used daily precipitation data from the Global Historical Climatology Network (GHCN, Menne et al., 2012a,b). Time periods of precipitation records for each station differ. The length of the time series affects the bias and uncertainty related to the parameters estimation when the Maximum Likelihood Estimator (MLE) is used (Tyralis and Koutsoyiannis, 2011, see also Section 3.1). Therefore, we preferred to use the common time period 1916-2015, while we discarded data out of this period, even when the instrumental data were covering a longer time period. The code and the datasets used in the present study are available online (see Appendix A).

2.1 Station and data selection

The initial dataset included time series with missing or flagged (i.e. data of low quality for reasons explained in Menne et al., 2012a) values. We processed the dataset according to the following briefly described sequence of actions.

A. Flagged values were considered as missing values.

B. We used the values 0.34 and 0.83 to differentiate between the months. Months with a percentage of observations higher than 0.83 (i.e. with more than 25/30 or 26/31 daily observations) are considered good, while months with a percentage of recorded values less than 0.34 (i.e. equal or less than 10/30 and 10/31 daily observations) are considered of poor quality. The reason for the differentiation is that we first aggregate to the monthly time scale and then to the annual time scale. Thus even if all values in a month are missing we can fill the monthly value after the first aggregation as described in step C.

B1. Missing values within months with observed values more than 83% were filled using linear interpolation.

B2. All values within months with observed values less than 34% were considered as missing.

B3. For the rest of the months the missing values were filled using linear interpolation and then these months were considered as missing. The reason is explained in step D.

C. Missing months corresponding to steps B2 and B3 (the latter after the substitution with missing values) were filled using a seasonal Kalman filter, implemented in the R package zoo (Zeileis and Grothendieck, 2005).

D. Mean monthly values for months in which both steps B3 and C (i.e. months with missing values more than 34% and less than 83%) were applied, were calculated with the mean of monthly values of steps B3 and C.

E. From the mean monthly values we obtained the mean annual values.

F. Finally we discarded annual time series if one of the following constraints was satisfied:

F1. Two or more missing years.

F2. $\hat{H} \geq 0.95$, mean annual rainfall $\hat{\mu} \geq 3000$ mm, standard deviation of annual rainfall $\hat{\sigma} \geq 750$ mm, coefficient of variation of annual rainfall $\hat{c}_v \geq 0.8$. We set these constraints on the estimated parameters because a preliminary analysis showed that higher values were outliers.

F3. Four or more years with less than 60% of observed daily values.

The estimated parameters of the annual time series of step F2 are described in Section 3.1. The interested reader is referred to Part 3 of the Supplementary Information for more details regarding the use of selection algorithms, constraints for data inclusion and other details. In Part 4 of the Supplementary Information we present boxplots for every parameter of Step F2. Despite the removal of outliers based on the preliminary analysis in step F2, some outliers remained. This will not constitute a problem in our analysis, due to the robustness of the random forests (see Section 3.2) to outliers (Breiman 2001). One may argue that high values observed in step F2 may also be representative of the parameters' population, therefore they should not be removed (e.g. see the textbooks of Aggarwal, 2017; Barnett and Lewis, 1978; Hawkins, 1980). Since the aim of our study is to investigate the relationship between H and location features, and due to the low number of outliers found in step F2, their removal will not affect the analysis.

We present the locations of the subset of stations, which remained after the initial procedure, in Figure 1. 1 535 stations remained, most of which are located in Australia, Europe and North America. Data for each station include its geographic coordinates, i.e.

elevation, longitude and latitude. We calculated the Cartesian coordinates of stations under the assumption of a spherical earth using eqs (1)-(3) to model the proximity of stations, which appear to be algebraically distant when considering their longitudes.

$$x = R \cos(\text{lat}) \cos(\text{lon}) \quad (1)$$

$$y = R \cos(\text{lat}) \sin(\text{lon}) \quad (2)$$

$$z = R \sin(\text{lat}) \quad (3)$$

R denotes the radius of the earth. The x and y axes of the Cartesian coordinate system define a plane which includes all points with zero latitude, while the z axis is perpendicular to the plane. E.g. for given lat = 0°, stations with longitudes -180° and 180° are coincident. The coincidence can be reproduced by the transformations (1)-(3).

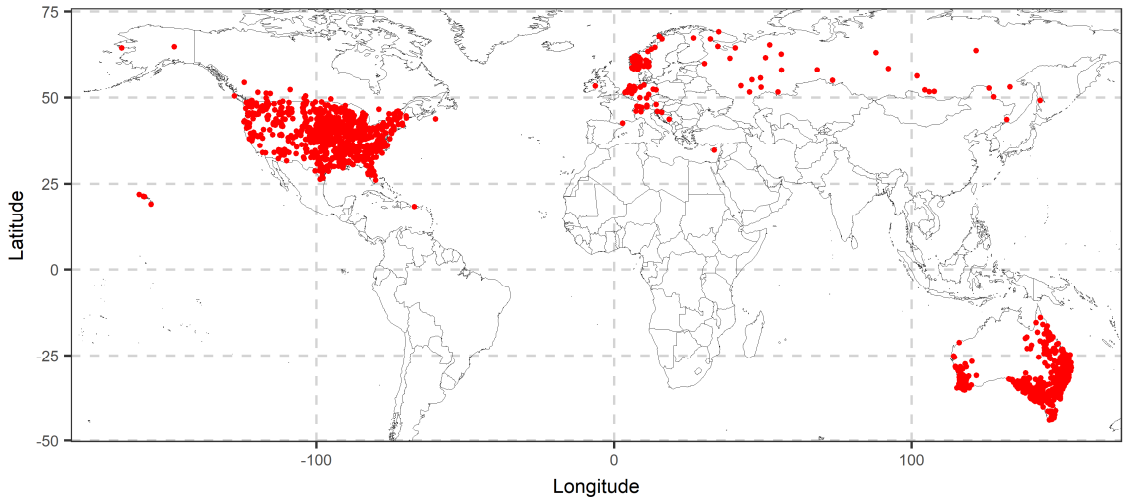

Figure 1. Map of locations for the 1 535 stations used in the analysis.

2.1.1 Grouping of stations to Köppen-Geiger climate classes

The stations are grouped based on the climate classification of Köppen-Geiger (Kottek et al., 2006). Table 1 presents the classes, whose combination gives the climatic types. We grouped the stations according to the climate type of the nearest point of the grid provided by Kottek et al. (2006). We calculated distances between stations and grid points using the Haversine ('half-versed-sine') formula as implemented in the R package geosphere (Hijmans, 2016a). Table 2 presents the 20 climate types of the stations. Twelve more climate types in Kottek et al. (2006) were not represented by the spatial distribution of the stations. The model calibration presented in Section 3.2 cannot be applied to the initial classification, because some climate types include a low number of stations. We regrouped the stations in the three groupings presented in Table 2. Grouping 1 included types with low number of stations together, considering their main

climate and precipitation type. Grouping 2 classified stations according to their main climate. Grouping 3 is similar to that of Ragulina and Reitan (2017), who regrouped the stations according to precipitation conditions.

Table 1. Köppen-Geiger climate classes (Adapted from Figure 1 in Kottek et al. 2006).

Main climate		Precipitation		Temperature	
A	equatorial	W	desert	h	hot arid
B	arid	S	steppe	k	cold arid
C	warm temperate	f	fully humid	a	hot summer
D	snow	s	summer dry	b	warm summer
E	polar	w	winter dry	c	cool summer
		m	monsoonal	d	extremely continental
				F	polar frost
				T	polar tundra

Table 2. Köppen-Geiger climate types of stations in Figure 1 and their regroupings.

Climate class	Number of stations	Grouping 1	Grouping 2	Grouping 3
Am	5	A	A	Am
As	4	A	A	As
Aw	9	A	A	Aw
BSh	65	BS	B	steppe
BSk	223	BS	B	steppe
BWh	21	BW	B	BWh
BWk	6	BW	B	without dry season
Cfa	419	Cfa	C	without dry season
Cfb	206	Cfb	C	without dry season
Csa	41	Ca	C	summer dry
Csb	125	Csb	C	summer dry
Cwa	5	Ca	C	winter dry
Dfa	181	Dfa	D	without dry season
Dfb	148	Dfb	D	without dry season
Dfc	52	Dfc	D	without dry season
Dsb	8	Dsw	D	summer dry
Dsc	1	Dsw	D	summer dry
Dwb	3	Dsw	D	winter dry
Dwc	4	Dsw	D	winter dry
ET	9	E	E	polar tundra

3. Methods

Here we present a minimum theoretical background of the methods, because they are established in the scientific literature.

3.1 Hurst-Kolmogorov process

We modelled the annual time series of Section 2.1 with the HKp. Let $\{\underline{x}_t\}$, t = 1,2,… be a HKp. The HKp is a three-parameter normal stationary stochastic process in discrete time. Its parameters μ, σ, H are defined by eqs (4)-(6) (Tyralis and Koutsoyiannis, 2011).

$$\mu := \mathrm{E}[\underline{x}_t] \tag{4}$$

$$\sigma := (\mathrm{Var}[\underline{x}_t])^{1/2} \tag{5}$$

$$\rho_k := \mathrm{Corr}[\underline{x}_t, x_{t+k}] = |k+1|^{2H} / 2 + |k-1|^{2H} / 2 - |k|^{2H},\ k = 0, 1,\ldots,\ H \in (0, 1) \tag{6}$$

The parameter μ is the mean of the stochastic process and the parameter σ is its standard deviation. The parameter H represents the magnitude of LRD, i.e. the tendency of wet or dry years to be clustered in long time periods (persistence), while the autocorrelation function ρ_k increases with H. High values of H denote strong long-term persistence, while when H < 0.5, the resulting stochastic process is antipersistent, but still stationary. H = 0.5 is equivalent to a stochastic process of independent variables. The implementation of the Maximum Likelihood Estimator in the R package HKprocess (Tyralis 2016) was applied for estimating μ, σ and H. Furthermore, we computed the maximum likelihood estimate of the coefficient of variation, defined as:

$$c_v = \sigma/\mu \tag{7}$$

The maximum likelihood estimate of c_v can be obtained from eq (7) after substitution of μ and σ with their maximum likelihood estimates due to the invariance properties of the MLE. From hereinafter μ, σ and H will denote the estimates of the respective parameters (we will not use hats as in step F.2 in Section 2.1).

The maximum likelihood estimator of the HKp parameters has excellent properties when compared to other estimators as shown in the simulation experiments in Tyralis and Koutsoyiannis (2011), while similar simulation experiments can be found in Taqqu et al. (1995), Jeong et al. 2007 and Rea et al. (2013). The latter three studies mostly implement estimators presented in Tyralis et al. (2011).

3.2 Model fitting and testing

We regressed H on combinations of other available variables related to local characteristics of the stations, i.e. their geographic coordinates, Cartesian coordinates, elevation, climate type, μ and σ. The use of geographic coordinates is more intuitive compared to Cartesian coordinates, thus we preferred to visualize the results using the

former coordinate system. The regression was applied using linear regression, the random forests algorithm (Biau and Scornet, 2016; Breiman, 2001) as implemented in the R package randomForest (Liaw and Wiener, 2002) and the random forests based on conditional inference trees (cforest algorithm, Strobl et al., 2007; 2008) as implemented in the R package party (Hothorn et al., 2017).

Properties of linear models are well known, however random forests are less used in hydrological sciences. Random forests can handle non-linear interactions and highly correlated variables and have high predictive power. Furthermore, random forest variable importance measures for variable selection purposes are available (Strobl et al., 2008). Therefore, despite being black boxes they can still provide information about the relationship between the dependent and the predictor variables. In this study, we used the permutation importance, which measures the mean increase of the prediction Mean Squared Error on the out-of-bag portion of the data after permuting each predictor variable in the trees of the trained model. More details can be found in the documentation of the importance function of the R package randomForest (Liaw and Wiener, 2002). Yet, the random forest importance variable measures are not reliable when the predictor variables vary in their scale of measurement or their number of categories (Strobl et al., 2007). In such cases, Strobl et al. (2007) propose the use of the cforest algorithm and its respective permutation importance measure, which we also used in our study.

The three algorithms are applied using the R package caret (Kuhn, 2008; Kuhn et al., 2017). We trained the three models on 80% of the sample, and we tested their performance on the remaining 20%, using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Pearson's *r* metrics. *H* was the dependent variable, while we used a combination of spatial and location variables as predictors.

The first testing of various combinations of predictor variables can be used initially to choose the most suitable for predicting *H* and reassess the selected combinations applying a computationally demanding 5-fold cross-validation. In the 5-fold cross-validation, the original sample is randomly divided into five equal sized subsamples. The model is fitted in four subsamples and tested in the remaining one, while the procedure is repeated five times. Consequently, the randomness of the partitioning of the dataset in the 5-fold cross-validation influences the results less compared to the simple cross-

validation. The 5-fold cross-validation was applied to four datasets sets, i.e. the full dataset and three subsets including stations in Australia, Europe and the USA respectively. In the 5-fold cross-validation, we compared the performance of random forests for predicting *H* with the simulations of a truncated normal distribution fitted to the sample of *H*s and with a naïve approach in which the predicted value is equal to the median of *H* of the fitting set. The maximum likelihood estimates of the parameters of the truncated normal distribution in each one of the 80% folds, were used for the simulation of the other 20%. The maximum likelihood estimates were obtained using the R package tmvtnorm (Wilhelm and Manjunath, 2015). The RMSE, Pearson's *r* and the slope of the regression line between the predicted and testing values metrics were used for the comparison.

For more details on the application of the algorithms and the use of tuning parameters on the case of random forests and cforest, through the R package caret the interested reader is referred to Parts 4-8 of the Supplementary Information. For more details regarding the use of metrics to assess the predictive performance of regression algorithms the reader is referred to Alexander et al. (2015) and Gramatica and Sangion (2016).

3.3 Mann-Kendall test under the long-range dependence assumption

The MKt-LRD consists of three consecutive hypothesis tests, namely **O** (Original MK test), **H** (Hurst Parameter test) and **M** (Hamed 2008). Let H_{0i} denote the null hypothesis of each test and let H_{1i} denote the alternative hypothesis, where i = **O**, **H**, **M** denotes the step of the MKt-LRD. The null hypotheses are as follows.

- $H_{0\mathbf{O}}$: No trend under the independence assumption.
- $H_{0\mathbf{H}}$: No significant LRD.
- $H_{0\mathbf{M}}$: No trend under LRD assumption.

The possible outcomes of the test are summarized by the following sequences.

- $\{H_{0\mathbf{O}}\}$: No significant trend.
- $\{H_{1\mathbf{O}}, H_{0\mathbf{H}}\}$: Significant trend exists.
- $\{H_{1\mathbf{O}}, H_{1\mathbf{H}}, H_{0\mathbf{M}}\}$: No significant trend.
- $\{H_{1\mathbf{O}}, H_{1\mathbf{H}}, H_{1\mathbf{M}}\}$: Significant trend exists.

We used the test implementation in the R package HKprocess (Tyralis, 2016) with a predefined significance level $\alpha = 0.05$ for all steps. For more details on the algorithm and its implementation using the R package HKprocess the interested reader is referred to Tegos et al. (2017). Furthermore, we estimated the trends of the annual time series with the fitting of a linear model. The estimated trends were set equal to the slope of the least squares line.

3.4 Global Moran's *I* test

The global Moran's *I* test uses the Moran's *I* statistic (Moran, 1950). The Moran's *I* statistic is a measure of spatial autocorrelation based on the location of variables and their observed values. In the global Moran's *I* test, the null hypothesis is that the observed spatial pattern is a realization of a random spatial process. The alternative hypotheses are that the spatial distribution of high values and/or low values are spatially clustered or they are spatially dispersed. The former holds for low *p*-values and positive *z*-score, while the latter for low *p*-values and negative *z*-score. More details on the global Moran's test can be found in Bivand et al. (2013b, pp. 275–284, 350–351). Here we applied the global Moran's *I* test by implementing the spdep R package (Bivand and Piras, 2015; Bivand et al., 2013a).

3.5 Kriging

Besides random forests, we furthermore used ordinary kriging to predict *H* spatially, for comparison and benchmarking reasons. Kriging is a stochastic method of spatial interpolation, which uses normal processes to model the observed values. For more details, the reader may refer to Bivand et al. (2013b, pp. 232, 233). Kriging is less computationally intensive compared to the random forests. We used the gstat R package to perform kriging (Pebesma, 2004; Gräler et al., 2016). In addition, we used exponential functions to model the autocovariance of the spatial model. Details on the autocovariance functions can be found in the documentation of the R package geoR (Ribeiro Jr and Diggle, 2016).

4. Methodology summary

Here we describe an outline of the method and the procedure of our analysis. Firstly, we selected stations with precipitation data in the time period 1916-2015, we filled the missing data, we computed the mean annual precipitation values and discarded some

stations, which did not satisfy the criteria set in Section 2.1. Then we grouped the stations in climate types (see Section 2.1.1). The record for each station includes its location (in geographic and Cartesian coordinates), its elevation, its climate type (three groupings) and mean annual precipitation time series.

We modelled the time series with HKp and we estimated the parameters μ, σ, H (Section 3.1). We regressed H on combinations of location parameters using linear regression, random forests and the cforest algorithm. The fitting of the algorithms was performed in the 80% of the 1 535 stations, while their performance was tested in the other 20%. We compared the predictions of H between the random forests, the simulation from a fitted truncated normal distribution and the naïve method in a 5-fold cross-validation using the RMSE, the Pearson's r and the slope of the regression line between the predicted and testing values. We applied the 5-fold cross validation to the entire dataset as well as three subsets, each one corresponding to a continent. Furthermore, we computed variable importance measures with the application of random forests and the cforest to the full dataset (Section 3.2). The combination of the validations and the use of variable importance measures can provide reliable information despite the shortcomings of each method when used individually.

We applied the global Moran's I test to the Hs to find possible spatial patterns with the aim to assess the results of the regression model. We furthermore compared the predictive performance of the random forests with ordinary kriging in which we used an exponential covariance function to model the spatial dependence of Hs. The comparison was performed using data from the contiguous part of the USA. To find a lower bound of predictive performance we modelled the spatial pattern of the Hs in the USA, using a Gaussian random field with an exponential covariance function, while we attempted to preserve the spatial dependence of the Hs. To this end, we simulated the spatial pattern 1000 times and we tested the predictive performance of ordinary kriging in predicting H in the 20% of the sample, when fitted in the 80% of the sample. Finally, we estimated the trend and its significance under the LRD assumption (Section 3.3) and we visualized the results coupled with location variables.

5. Long-range dependence analysis

In this Section we present the results of the analysis for the H parameter.

5.1 Overview of *H*

Figure 2 is the histogram of *H*s. The maximum likelihood estimated values are skewed to the right with skewness equal to 0.21, while the median value is equal to 0.56. For comparison reasons and as shown in a simulation study in Part 9 of the Supplementary Information, the median of the *H* estimates of 100 000 simulated time series of length equal to 100 and $H = 0.59$ is equal to 0.56. Furthermore, 95% of the simulated experiment *H*s are in the interval (0.413, 0.691), while 95% of the precipitation dataset *H*s are in the interval (0.402, 0.733). A truncated normal distribution with support (0,1) seems to be a reasonable model for *H*.

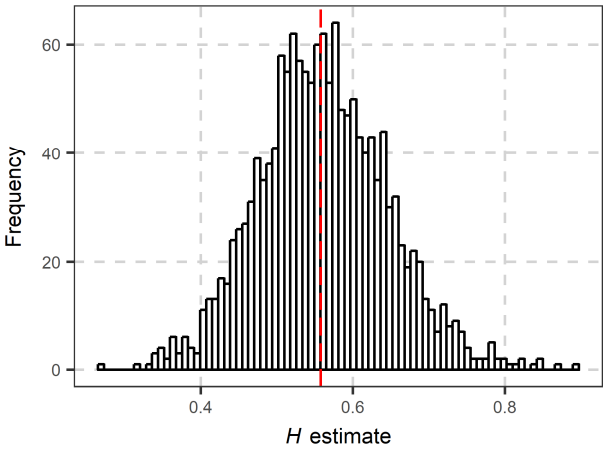

Figure 2. Histogram of *H* based on measurements from 1 535 stations. The median of the estimates is represented by the vertical red line and equals 0.56.

Figure 3 presents the correlations between some variables of interest. The longitude is omitted, while the inclusion of x and y coordinates as single variables would be meaningless. We observe a high correlation between μ and σ and between the absolute latitude and c_v. *H* is not highly correlated with any of the variables in Figure 3.

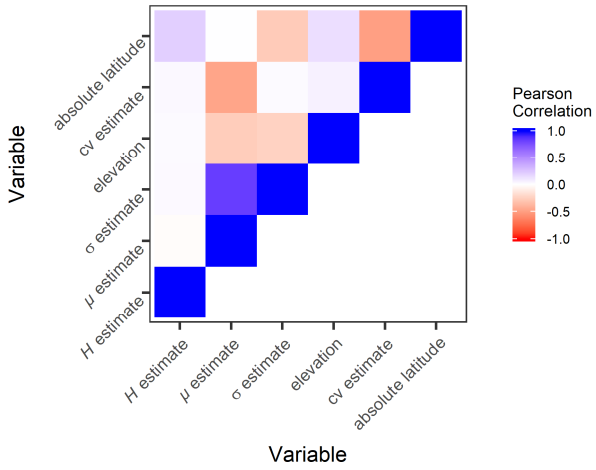

Figure 3. Correlations between numeric variables of each station based on the dataset of 1 535 stations.

To investigate the spatial properties of H we computed the spatial correlogram, presented in Figure 4 (see Part 9). The autocorrelation for distances equal to 250 km is approximately 0.1, while it increases to approximately 0.2 for distances equal to 100 km. The low autocorrelations for distances higher than 250 km may decrease the accuracy of predictions of H based on spatial characteristics. Furthermore, we calculated the Global Moran's I statistic (Moran, 1950) using the global dataset, as well as the respective datasets in the USA and Australia (Part 9 of the Supplementary Information). Applying the hypothesis tests based on the Global Moran's I statistic, we found p-values almost equal to 0 and positive z-scores implying that the spatial distribution of high or low values is spatially clustered, which rejects the hypothesis that the underlying spatial process is random.

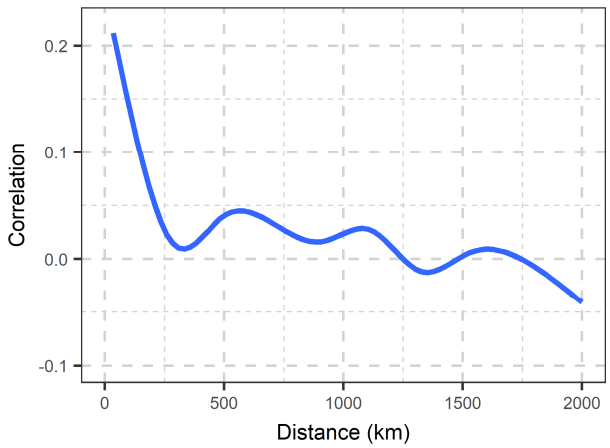

Figure 4. Smoothed spatial correlogram of the H values from the dataset of the 1 535 stations. The computations are based on Bjornstad (2016) and they are presented in Part 9 of the supplementary information.

5.2 Visualization of *H* coupled with the predictor variables

In this Section we visualize *H* coupled with the predictor variables. We present a full exploratory data analysis in the Supplementary Information, while here we present some important Figures for brevity. Figure 5 presents how *H* varies with the climate class of the station. Grouping 2 of Table 2 is used as the predictor variable. It seems that *H* does not significantly vary with grouping 2, while its values are near to the median value 0.56, computed in Section 5.1. On the other hand grouping 1 (see Table 2) in Figure 6 seems to be a better predictor, because of the higher variation of *H* between different climate classes.

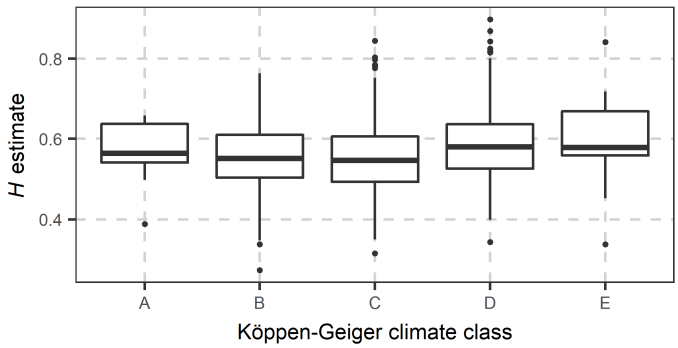

Figure 5. Boxplot of *H* based on the dataset of 1 535 stations conditional on the Köppen-Geiger climate class (grouping 2).

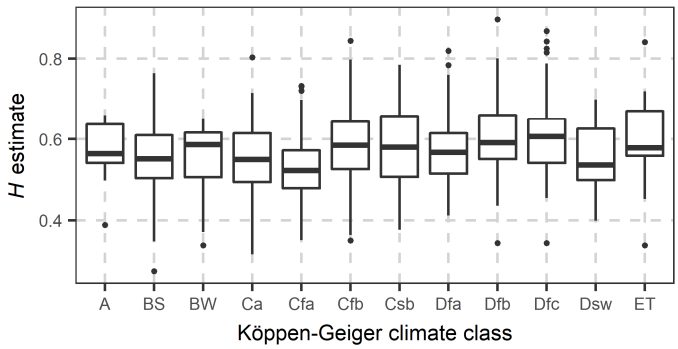

Figure 6. Boxplot of *H* from the dataset of 1 535 stations conditional on the Köppen-Geiger climate class (grouping 1).

In Figure 7, we observe the variation of *H* with the latitude. Higher *H* values are observed for positive latitude, however no trend prevails, while we do not observe any linear relationship between the two variables. Figure 7 also presents the relationship between *H* and longitude. Again, we do not observe any clear linear relationship between the two variables. Furthermore, *H* is not linearly related to the elevation of each station.

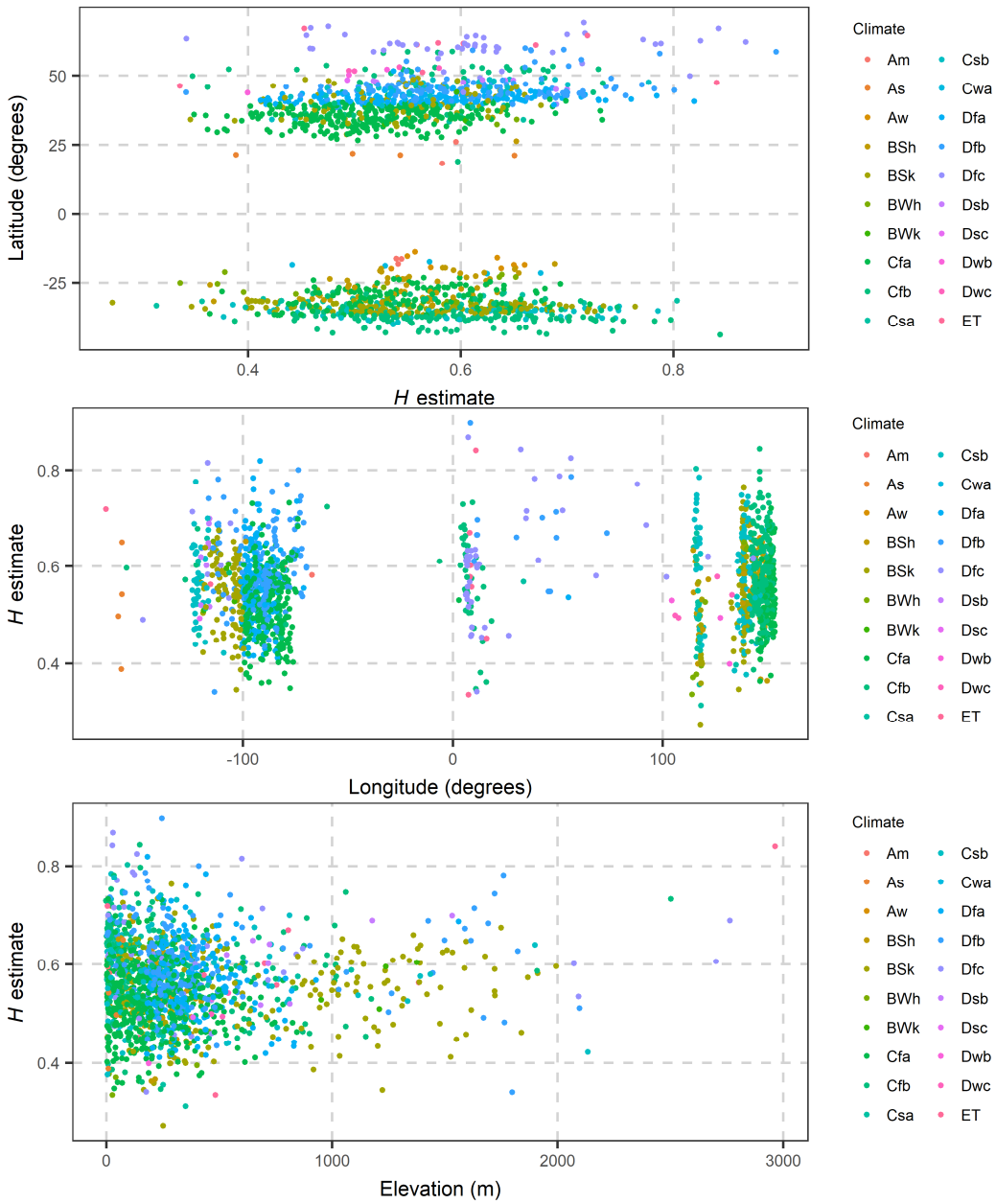

Figure 7. Scatterplot of H and the latitude (top), longitude (middle) and elevation (bottom) of each station. The legend presents the Köppen-Geiger climate class of each station.

5.3 Model fitting and testing

From the analysis in Section 5.2, it is apparent that a linear regression model between H and the location variables could be a benchmark, and be compared with the more complex random forests and the cforest algorithm. We examined combinations of predictor variables as shown in Table 3. Combinations 1-11 and 17-26 include the dependence of H on the location of the stations. Combination 12 examines its dependence on variables, which are features of the precipitation of the station, while combinations 13-16 and 27-30 examine both location and precipitation features. We

built the models of Table 3 using a stepwise regression method and in particular a forward selection approach, i.e. we started with no variables and we tested the addition of each variable using criteria such as the RMSE, the MAE, the MAPE and Pearson's *r*.

Table 3. Predictor variable combinations, examined in the fitting of models for the prediction of *H*. xyz are the Cartesian coordinates of each station. Grouping is defined in Table 2.

Combination	Predictors
1	elevation
2	grouping 1
3	grouping 2
4	grouping 3
5	x, y
6	x, y, z
7	x, y, z, grouping 1
8	x, y, z, elevation
9	x, y, z, elevation, grouping 1
10	x, y, z, elevation, grouping 2
11	x, y, z, elevation, grouping 3
12	μ, σ
13	x, y, z, elevation, grouping 1, μ
14	x, y, z, elevation, grouping 1, μ, σ
15	x, y, z, elevation, grouping 2, μ, σ
16	x, y, z, elevation, grouping 3, μ, σ
17	longitude
18	latitude
19	longitude, grouping 1
20	latitude, grouping 1
21	longitude, latitude
22	longitude, latitude, grouping 1
23	longitude, latitude, elevation
24	longitude, latitude, elevation, grouping 1
25	longitude, latitude, elevation, grouping 2
26	longitude, latitude, elevation, grouping 3
27	longitude, latitude, elevation, grouping 1, μ
28	longitude, latitude, elevation, grouping 1, μ, σ
29	longitude, latitude, elevation, grouping 2, μ, σ
30	longitude, latitude, elevation, grouping 3, μ, σ

We fitted the models on 80% of the data and we tested their performance in predicting *H* on the other 20%. In Table 4, we present the testing results of each model. Combinations 1 and 3-16 for the cforest algorithm were omitted due to high computational load combined with the fact that they would not behave considerably different compared to the respective application of random forests. Random forests and the cforest had good performance while the performance of linear models was poor, indicating a strong non-linear relationship between the predictor variables and *H*. The

term linear here refers to the relationship e.g. between H and the elevation or the latitude and not to spatial relationships. The cforest is more computationally intensive compared to the random forests. Firstly, we examined the dependence of H on the elevation and the climate (combinations 1-4). Grouping 1 (combination 2) was the best predictor with a similar performance for all methods. Then, we examined the dependence of H on the Cartesian coordinates combined with or without other variables (combinations 5-11, 13-16). The combination 5 (i.e. x and y coordinates) performed very good in random forests, while the inclusion of the z coordinate, the elevation and the climate type further improved the performance. Combination 11 which includes grouping 3 performed marginally better than combinations 9 and 10 which include groupings 1 and 2 respectively. Inclusion of μ and σ further improved the performance of the random forests (combinations 13-16). Secondly, we performed a similar investigation using the geographic coordinates instead of the Cartesian coordinates (combinations 17-30). The longitude and latitude (combinations 17, 18) are not good predictors. When we combine each one of them with grouping 1 (combinations 19, 20) the results are worse or similar with using grouping 1 as a single predictor. The combination 21 (i.e. longitude and latitude) performed well, while the inclusion of grouping 1 (combination 22) weakened the regression model. On the other hand, the inclusion of the elevation (combination 23) improved marginally the performance of the model. Climate type (combinations 24-26) worsened the performance, while inclusion of μ and σ (combinations 27-30) further improved the performance of the random forests. It is noteworthy that some results seem incoherent. E.g. in the case of Cartesian coordinates, climate improves the random forests results (combinations 8-11), while for the geographic coordinates (combinations 23-26), it is the opposite. This may be explained by the slight deviations induced by the inclusion of climate. In this case, the 5-fold cross-validation presented in the following is a valid method to obtain a more reliable inference.

Table 4. Model errors in the test set for predicting *H* for each method and metric. Comb is the combination of predictor variables as presented in Table 3. RMSE is the Root Mean Squared Error, MAE is the Mean Absolute Error, MAPE is the Mean Absolute Percentage Error and *r* is the Pearson's *r*.

Comb	Linear model				Random forests				cforest			
	RMSE	MAE	MAPE	*r*	RMSE	MAE	MAPE	*r*	RMSE	MAE	MAPE	*r*
1	0.086	0.068	0.124	0.01	0.096	0.075	0.137	0.02				
2	0.084	0.068	0.124	0.24	0.084	0.068	0.124	0.24	0.084	0.068	0.124	0.24
3	0.086	0.068	0.124	0.09	0.086	0.068	0.124	0.09				
4	0.088	0.069	0.126	-0.03	0.087	0.069	0.126	−0.03				
5	0.086	0.068	0.125	0.06	0.080	0.063	0.114	0.42				
6	0.086	0.068	0.123	0.11	0.079	0.061	0.110	0.44				
7	0.084	0.068	0.123	0.26	0.079	0.061	0.111	0.43				
8	0.086	0.068	0.123	0.11	0.077	0.059	0.107	0.47				
9	0.084	0.068	0.123	0.26	0.077	0.060	0.109	0.45				
10	0.085	0.067	0.122	0.17	0.077	0.059	0.108	0.46				
11	0.086	0.068	0.124	0.13	0.076	0.059	0.106	0.48				
12	0.086	0.068	0.124	0.07	0.091	0.071	0.130	0.09				
13	0.083	0.067	0.123	0.27	0.076	0.059	0.108	0.47				
14	0.082	0.067	0.123	0.31	0.073	0.058	0.106	0.53				
15	0.085	0.067	0.122	0.21	0.073	0.058	0.105	0.54				
16	0.086	0.068	0.124	0.14	0.073	0.057	0.104	0.54				
17	0.086	0.068	0.124	0.05	0.098	0.077	0.141	0.14	0.084	0.067	0.121	0.28
18	0.086	0.069	0.124	0.00	0.097	0.078	0.142	0.09	0.087	0.069	0.127	0.19
19	0.084	0.068	0.125	0.24	0.092	0.072	0.132	0.25	0.081	0.064	0.117	0.36
20	0.084	0.068	0.125	0.24	0.092	0.072	0.131	0.19	0.082	0.064	0.117	0.34
21	0.086	0.068	0.123	0.12	0.080	0.063	0.113	0.42	0.078	0.061	0.110	0.43
22	0.084	0.068	0.124	0.25	0.081	0.063	0.115	0.39	0.080	0.062	0.114	0.39
23	0.086	0.067	0.123	0.13	0.078	0.060	0.110	0.45	0.078	0.061	0.110	0.43
24	0.084	0.068	0.124	0.25	0.079	0.061	0.112	0.42	0.080	0.062	0.114	0.39
25	0.085	0.067	0.122	0.17	0.078	0.061	0.111	0.43	0.080	0.062	0.113	0.38
26	0.086	0.067	0.123	0.14	0.078	0.060	0.110	0.44	0.079	0.061	0.112	0.40
27	0.083	0.068	0.124	0.26	0.078	0.061	0.111	0.44	0.079	0.062	0.112	0.41
28	0.082	0.067	0.122	0.32	0.075	0.059	0.109	0.50	0.077	0.061	0.111	0.46
29	0.085	0.067	0.122	0.20	0.075	0.059	0.108	0.50	0.077	0.060	0.110	0.46
30	0.086	0.068	0.123	0.15	0.075	0.059	0.107	0.50	0.077	0.060	0.109	0.47

In Figure 8, we present the predicted *H* from the application of the trained random forests for the combination 24 to the test set. Pearson's *r* indicates a good prediction, while the range of predicted *H*s is smaller than the range of *H*s in the test set. We observe the same behaviour for the cforest algorithm in Figure 8 albeit Pearson's *r* is somewhat lower. When Pearson's *r* is used to assess the predictive performance of regression models in the test set it is not a measure of correlation, while r^2 does not explain variability.

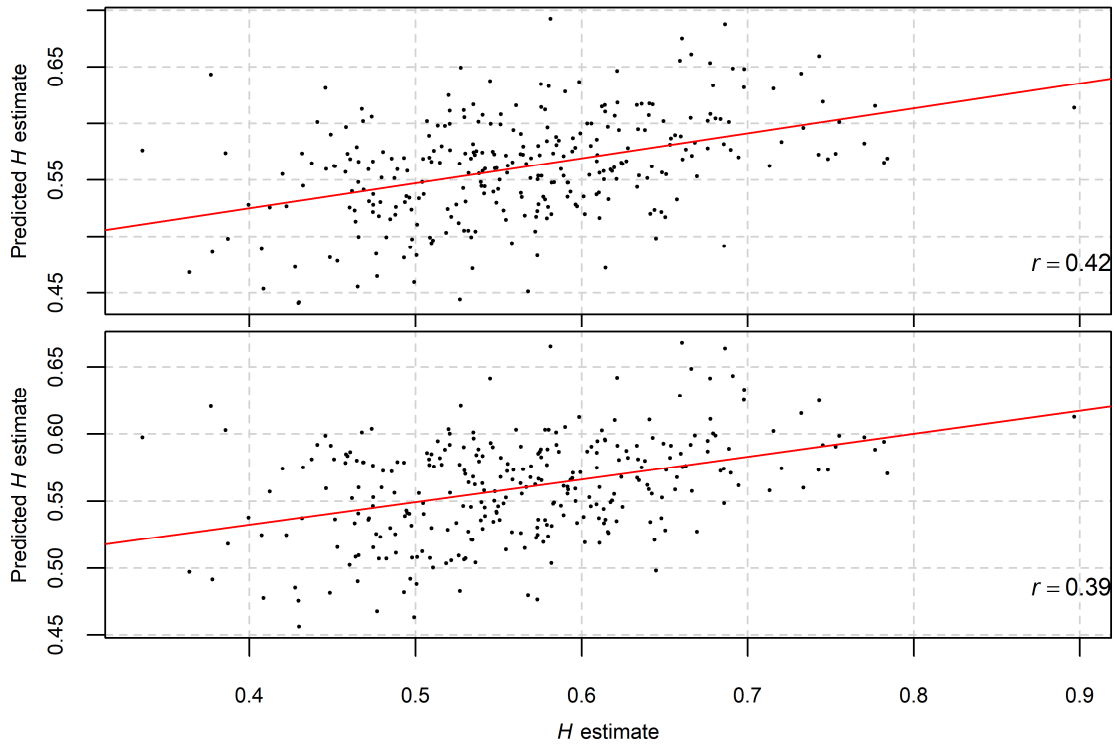

Figure 8. *H* of the test set in the x-axis and predicted *H* for the test set in the y-axis using the random forests (top) and the cforest algorithm (bottom) for the combination 24 of predictor variables defined in Table 3. *r* denotes Pearson's *r*.

In Table 5, we present the results of a 5-fold cross-validation for the prediction of *H*. We compare the random forests in the combinations 2, 9, 17-22, 24, 27, 28 of predictor variables with the truncated normal distribution and the naïve approach. The combinations 2, 9, 17-22, 24 can be used for the prediction of *H* at ungauged locations. The combinations 27 and 28, which include μ and σ could be useful to predict *H* at an ungauged location, if we could make assumptions about the μ and σ parameters based on experts' experience. This is also possible for cases with few years of observed data, since the uncertainty in estimating μ and to a lesser extend σ is less sensitive to the sample size. Besides μ and σ are also representative of the climate at the given location, thus they are equally useful at the stage of the analysis.

Regarding the overall view, the RMSE of the random forests is lower than that of the truncated normal distribution and the naïve approach in most cases. However, it is notable, albeit expected, that the Pearson's *r* and the slope are approximately 0 for the truncated normal distribution. This highlights the importance of the higher predicting performance of the random forests in terms of Pearson's *r* and the slope. While in all cases, the performance of the random forests is not perfect, i.e. Pearson's *r* is in the neighbourhood of 0.5 and the slope in the neighbourhood of 0.3, the improvement over

the benchmark approaches is somewhat significant. Indeed the RMSE is 10-15% lower. Furthermore, we note that the variation of RMSE, Pearson's *r* and the slope values is low for all 11 random forests cases, meaning that the algorithm is stable with respect to the choice of the fitting sample.

There is a rather weak relationship between H and grouping 1 (combination 2), while there is a rather moderate relationship between H and the longitude and latitude predictors (combination 21). The inclusion of grouping 1 to the longitude and latitude predictors (combination 22) did not improve the model compared to combination 21. However, the inclusion of grouping 1 and the elevation (combination 24) as predictor variables improved marginally the predictive performance of the fitted model. A possible explanation is that all information about H is included in the geographic location of the stations. Knowing the climate class of the stations does not add any information to that obtained by their locations. However the inclusion of μ and σ (combination 28) which may represent the climate in the location better than the climate class, further improved noticeably the performance of the prediction. On the other hand the better performance of the combination 28 compared to combination 27 is possibly owed to that H and σ are not orthogonal (Tyralis and Koutsoyiannis, 2011) and therefore the properties of the estimates of the former depend on the latter.

Table 5. 5-fold cross-validation for predicting *H* using the random forests, the truncated normal distribution and the naïve method. Comb is the combination of predictor variables as presented in Table 3. Val denotes the number of the cross-validation. Three metrics were used, i.e. RMSE which is the Root Mean Squared Error, *r* which is the Pearson's *r* and the slope of the regression line between the predicted and the observed values. The last column is equal to the mean value of the metrics.

Method	Comb	Metric	Val 1	Val 2	Val 3	Val 4	Val 5	Mean
Random forests	2	RMSE	0.079	0.079	0.080	0.086	0.083	0.082
	2	*r*	0.35	0.28	0.28	0.24	0.30	0.29
	2	slope	0.11	0.10	0.09	0.08	0.10	0.09
	9	RMSE	0.074	0.072	0.074	0.081	0.076	0.075
	9	*r*	0.49	0.49	0.49	0.42	0.49	0.48
	9	slope	0.29	0.28	0.28	0.24	0.27	0.27
	17	RMSE	0.092	0.091	0.089	0.089	0.087	0.090
	17	*r*	0.22	0.19	0.22	0.32	0.31	0.25
	17	slope	0.15	0.13	0.14	0.20	0.20	0.17
	18	RMSE	0.095	0.094	0.090	0.102	0.096	0.095
	18	*r*	0.15	0.10	0.19	0.07	0.12	0.12
	18	slope	0.10	0.06	0.11	0.04	0.07	0.08
	19	RMSE	0.082	0.084	0.083	0.089	0.080	0.084
	19	*r*	0.40	0.33	0.35	0.33	0.44	0.37
	19	slope	0.29	0.24	0.24	0.22	0.28	0.25
	20	RMSE	0.092	0.089	0.087	0.096	0.088	0.090
	20	*r*	0.21	0.20	0.24	0.17	0.28	0.22
	20	slope	0.14	0.13	0.15	0.10	0.17	0.14
	21	RMSE	0.078	0.075	0.074	0.081	0.077	0.077
	21	*r*	0.45	0.44	0.49	0.43	0.48	0.46
	21	slope	0.30	0.27	0.31	0.26	0.30	0.29
	22	RMSE	0.078	0.073	0.075	0.082	0.076	0.077
	22	*r*	0.44	0.47	0.48	0.42	0.49	0.46
	22	slope	0.29	0.28	0.29	0.25	0.30	0.28
	24	RMSE	0.074	0.072	0.073	0.082	0.077	0.076
	24	*r*	0.49	0.48	0.50	0.41	0.48	0.47
	24	slope	0.28	0.27	0.28	0.23	0.26	0.26
	27	RMSE	0.074	0.072	0.072	0.080	0.076	0.075
	27	*r*	0.49	0.49	0.52	0.43	0.49	0.48
	27	slope	0.28	0.27	0.28	0.22	0.26	0.26
	28	RMSE	0.071	0.068	0.070	0.079	0.075	0.072
	28	*r*	0.55	0.56	0.56	0.45	0.51	0.53
	28	slope	0.29	0.31	0.29	0.23	0.27	0.28
Truncated normal		RMSE	0.085	0.082	0.085	0.089	0.087	0.086
		r	−0.04	−0.01	−0.04	0.01	0.01	−0.01
		slope	0.00	0.00	0.00	0.00	0.00	0.00
Naïve		RMSE	0.085	0.082	0.084	0.089	0.087	0.085

We repeated the 5-fold cross validation for stations in the USA, Australia and Europe (Parts 6-8 of the supplementary information) and we present the results for the USA and Australia for reasons of brevity in Table 6. While the overall comparison pattern remains the same with Table 5, there is a remarkable difference between the RMSEs of the random forests. They are lower in the USA and higher in Australia. While the RMSE of the truncated normal distribution is equal in the entire dataset and Australia, the

inclusion of the predictor variables has a lesser effect in the case of Australia. For instance, the RMSE of combination 28 is lower by 0.11 (= 0.087 − 0.076) in Australia, while 0.14 (= 0.086 − 0.072) for the entire dataset. We observe a similar behaviour in the USA (0.12 = 0.078− 0.066), albeit the differences are less significant. We attribute this behaviour to the combined information offered by entire dataset, which can be effectively exploited by the random forests.

Table 6. 5-fold cross-validation for predicting *H* in the USA (left) and Australia (right) using the random forests, the truncated normal distribution and the naïve method. For further explanations, see Table 5.

			USA						Australia					
Method	Comb	Metric	Val 1	Val 2	Val 3	Val 4	Val 5	Mean	Val1	Val 2	Val 3	Val 4	Val 5	Mean
Random forests	2	RMSE	0.068	0.073	0.070	0.074	0.078	0.073	0.089	0.085	0.080	0.093	0.081	0.086
	2	*r*	0.35	0.30	0.37	0.46	0.36	0.37	0.07	0.13	0.22	0.24	0.32	0.20
	2	slope	0.15	0.13	0.15	0.16	0.13	0.14	0.02	0.03	0.06	0.05	0.07	0.05
	9	RMSE	0.066	0.067	0.071	0.068	0.071	0.069	0.078	0.079	0.073	0.084	0.076	0.078
	9	*r*	0.45	0.49	0.40	0.58	0.51	0.49	0.46	0.40	0.47	0.47	0.48	0.46
	9	slope	0.28	0.29	0.25	0.32	0.27	0.28	0.23	0.23	0.29	0.25	0.29	0.26
	17	RMSE	0.083	0.086	0.081	0.089	0.088	0.086	0.094	0.092	0.086	0.103	0.086	0.092
	17	*r*	0.12	0.22	0.19	0.18	0.14	0.17	0.22	0.23	0.31	0.15	0.31	0.25
	17	slope	0.08	0.16	0.13	0.11	0.08	0.11	0.15	0.15	0.25	0.09	0.20	0.17
	18	RMSE	0.082	0.086	0.080	0.083	0.086	0.084	0.099	0.095	0.092	0.103	0.097	0.097
	18	*r*	0.13	0.18	0.30	0.30	0.22	0.22	0.06	0.15	0.07	0.16	0.08	0.10
	18	slope	0.09	0.13	0.23	0.18	0.13	0.15	0.03	0.10	0.04	0.10	0.05	0.06
	19	RMSE	0.077	0.080	0.077	0.080	0.087	0.080	0.086	0.087	0.087	0.092	0.080	0.086
	19	*r*	0.28	0.32	0.34	0.36	0.23	0.31	0.35	0.32	0.27	0.37	0.43	0.35
	19	slope	0.20	0.25	0.26	0.22	0.15	0.22	0.22	0.22	0.20	0.23	0.30	0.24
	20	RMSE	0.080	0.082	0.077	0.082	0.083	0.081	0.094	0.094	0.084	0.098	0.087	0.091
	20	*r*	0.18	0.28	0.33	0.32	0.26	0.27	0.20	0.15	0.23	0.20	0.26	0.21
	20	slope	0.12	0.21	0.24	0.20	0.13	0.18	0.12	0.09	0.13	0.10	0.15	0.12
	21	RMSE	0.067	0.067	0.068	0.068	0.073	0.069	0.080	0.084	0.074	0.087	0.076	0.080
	21	*r*	0.45	0.50	0.47	0.57	0.48	0.50	0.44	0.37	0.46	0.44	0.50	0.44
	21	slope	0.32	0.33	0.33	0.32	0.25	0.31	0.25	0.26	0.31	0.25	0.32	0.28
	22	RMSE	0.066	0.068	0.070	0.068	0.072	0.069	0.080	0.083	0.076	0.088	0.073	0.080
	22	*r*	0.45	0.49	0.44	0.57	0.50	0.49	0.43	0.35	0.42	0.42	0.53	0.43
	22	slope	0.30	0.31	0.29	0.32	0.26	0.30	0.24	0.23	0.27	0.24	0.33	0.26
	24	RMSE	0.066	0.068	0.072	0.069	0.072	0.069	0.080	0.082	0.076	0.086	0.080	0.081
	24	*r*	0.45	0.48	0.38	0.56	0.50	0.47	0.42	0.34	0.40	0.45	0.42	0.41
	24	slope	0.29	0.29	0.24	0.30	0.25	0.27	0.21	0.19	0.24	0.24	0.26	0.23
	27	RMSE	0.066	0.067	0.070	0.068	0.072	0.068	0.079	0.079	0.077	0.084	0.078	0.079
	27	*r*	0.45	0.49	0.41	0.58	0.49	0.48	0.45	0.38	0.39	0.48	0.45	0.43
	27	slope	0.28	0.29	0.24	0.31	0.24	0.27	0.23	0.19	0.23	0.25	0.27	0.23
	28	RMSE	0.064	0.067	0.066	0.066	0.068	0.066	0.074	0.077	0.076	0.081	0.073	0.076
	28	*r*	0.48	0.50	0.48	0.62	0.56	0.53	0.54	0.45	0.41	0.52	0.54	0.49
	28	slope	0.30	0.29	0.27	0.32	0.29	0.29	0.26	0.24	0.25	0.26	0.30	0.26
Truncated normal		RMSE	0.073	0.076	0.076	0.083	0.083	0.078	0.088	0.086	0.081	0.095	0.085	0.087
		r	−0.02	0.05	−0.11	0.06	−0.02	−0.01	-0.01	-0.05	0.08	0.07	0.05	0.03
		slope	0.00	0.00	−0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
Naïve		RMSE	0.072	0.076	0.074	0.083	0.082	0.078	0.088	0.085	0.082	0.096	0.085	0.087

In Figure 9, we present the variable importance for the combination 24 of predictor variables because it includes all predictor variables excluding μ and σ. The location parameters combined are the most important for predicting *H*, followed by the elevation and the climate classification. On the other hand, the cforest algorithm differs in that it estimates higher importance of the climate classification as presented in Figure 9

(bottom). This is possibly owed to the better performance of the cforest algorithm when estimating categorical variables' importance.

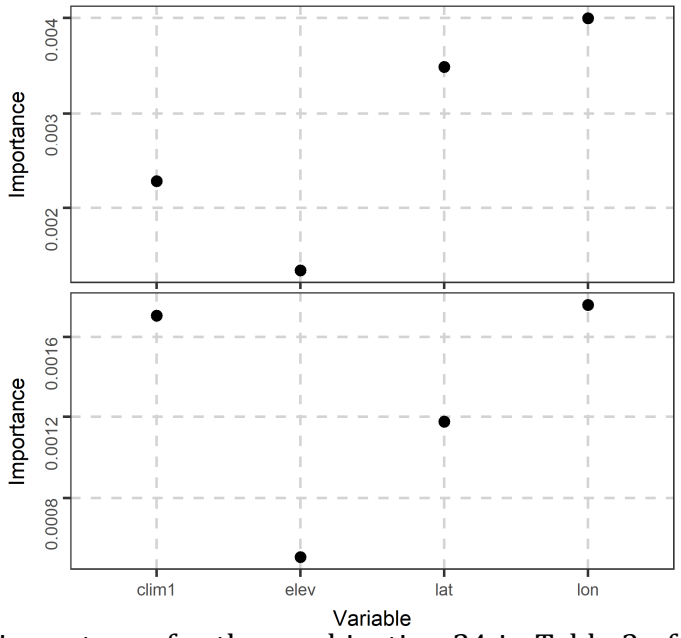

Figure 9. Variable importance for the combination 24 in Table 3 of predictor variables when random forests (top) and the cforest algorithm (bottom) is applied in the dataset of the 1 535 stations. The variable importance of a particular variable is the percentage of increase in mean square error observed in out-of-bag (OOB) prediction when this variable is randomly permuted.

To find a lower bound for the errors, we simulated the observed spatial pattern in the USA using Gaussian random fields as analysed in Sections 3.5 and 4 and we predicted H using ordinary kriging. The results are presented in Part 12 of the supplementary information. We found a lower bound for the RMSE equal to 0.059, which is an improvement in predictive performance equal to 24%. Furthermore, we predicted H in the USA using kriging. We performed a 5-fold cross validation, presented in Part 12 of the supplementary information. The RMSE was equal to 0.069 while r = 0.48 and the slope was equal to 0.26 indicating a slighter better performance of the random forests.

To present an application of the prediction model, in Figure 10 we show the variation of H in the USA and Australia obtained from the prediction of the random forests using the geographic coordinates as predictor variables. The spatial clustering appearing in Figure 10 is in accordance with the results of the hypothesis test based on the global Moran's I statistic presented in Section 5.1. In Figure 11, we present the errors of the random forests models when they are fitted in the 80% of the sample using the geographic coordinates as predictor variables and predict the remaining 20% of the

sample, in a 5-fold cross validation. The pattern of the errors seems to suggest a random spatial underlying process for the errors.

Summarizing the results of Section 5.3, one may interpret that the random forests could simply perform better than the naïve methods in cases in which the underlying process is random, implying that the better performance of the random forests is not the consequence of specific spatial patterns. To assess such arguments, we reassigned uniformly and randomly to the stations the *H*s and obtained a completely random pattern (Part 11 of the supplementary information). The 5-fold cross validation proved that the naïve methods had a better predictive performance compared to the random forests in the new dataset. In fact, the predictive performance of the random forests decreased considerably, compared to the real dataset, while the performance of the naïve methods, did not change, as expected.

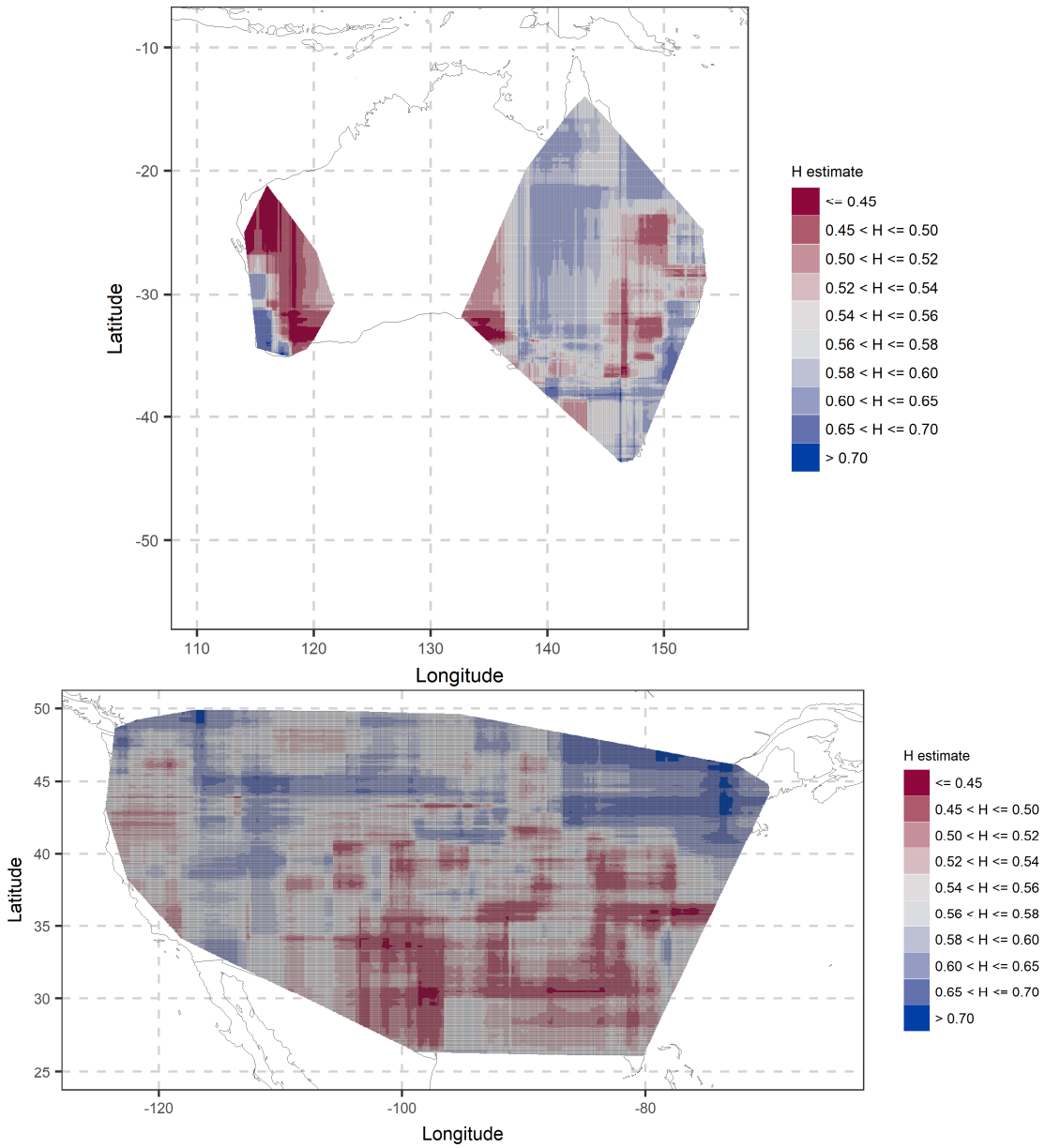

Figure 10. Heatmap of the Hurst parameter in Australia (top) and the USA (bottom) based on the prediction of the random forests when using the longitude and the latitude as predictor variables and the dataset of the 1 535 stations for fitting.

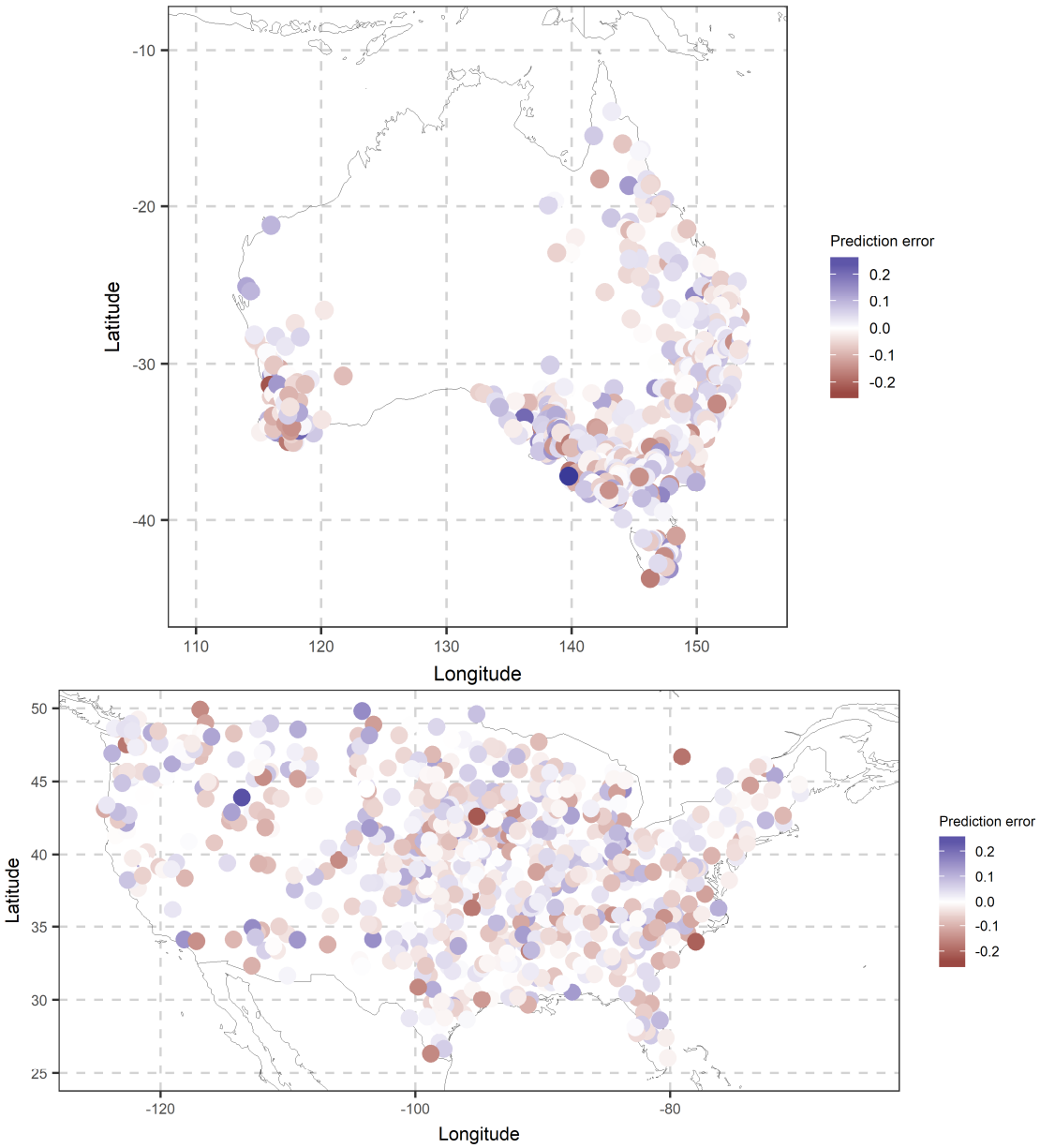

Figure 11. Errors of the prediction of random forests using the geographic coordinates as predictor variables in a 5-fold cross-validation.

6. Trend analysis

In this Section we present the analysis on the significance of the observed trends under the LRD assumption.

6.1 Overview of trend estimates

Figure 12 is the histogram of estimated trends from the dataset of the 1 535 stations for the time period 1916-2015. The median value is equal to 0.36 mm/year, i.e. in the last 100 years we observed an increase in the annual precipitation of 36 mm. For

comparison with the mean precipitation values, we note that the median annual precipitation for the 1535 stations is equal to 718 mm.

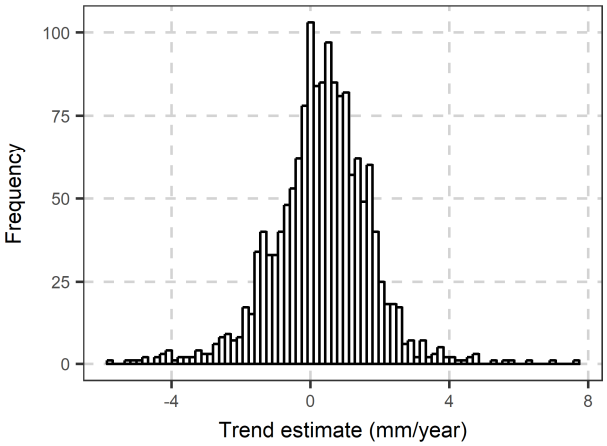

Figure 12. Histogram of trends based on the dataset of the 1 535 stations.

6.2 Visualization of trend estimates coupled with the location variables

In this Section we visualize the estimated trends as well as their significance coupled with location parameters. The full exploratory analysis is presented in the Supplementary Information, while here we present some important observations. In Figure 13, we present how the precipitation trend varies with the climate type. In all five types of grouping 2 the estimated trend is positive, while we observe a larger variation for climate type "A".

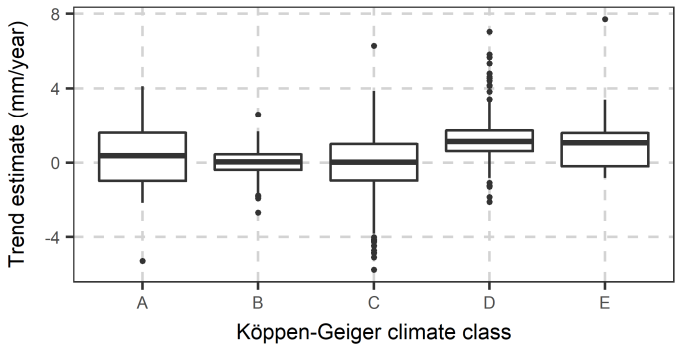

Figure 13. Boxplot of trend estimates based on the dataset of the 1 535 stations conditional on the Köppen-Geiger climate class (grouping 2).

Figure 14 presents the variation of trends conditional on grouping 3. It seems that non-significant differences are observed between different climate types.

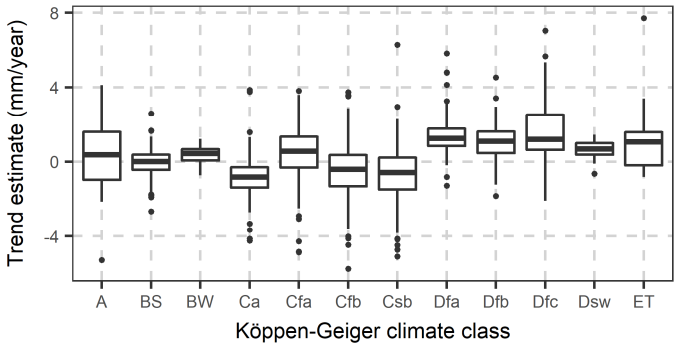

Figure 14. Boxplot of trend estimates from the dataset of the 1 535 stations conditional on the Köppen-Geiger climate class (grouping 1).

Notably, as shown in Figure 15, the mean annual precipitation seems to have been slightly increased in the Northern hemisphere and slightly decreased in the Southern hemisphere. This slight increase in the Northern hemisphere confirms the findings of van Wijngaarden and Syed (2015).

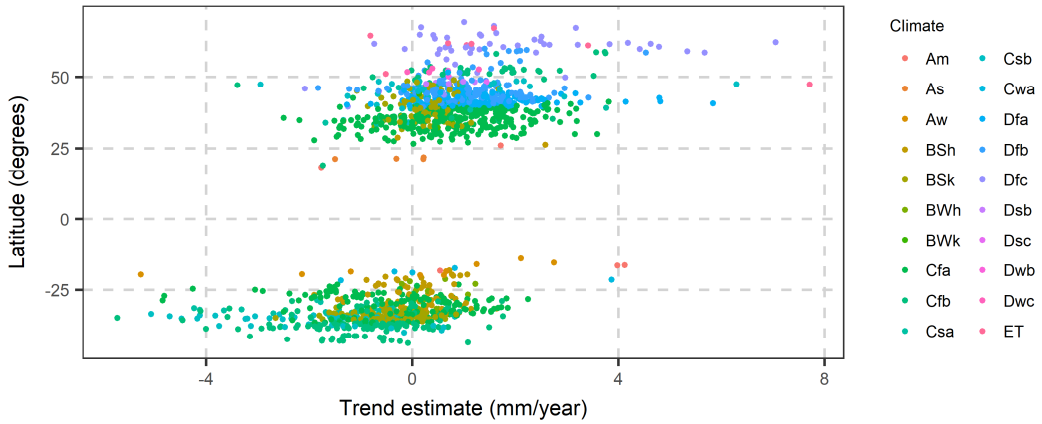

Figure 15. Scatterplot of trend estimate and the latitude of each station. The legend presents the Köppen-Geiger climate class of each station.

Figure 16 depicts the monotonicity and significance of trends for each main climate type, after application of the MKt-LRD with a predefined significance level $\alpha = 0.05$ for all steps to the mean annual precipitation time series. The absolute number of stations with main climate type D and positive significant trend is considerably higher compared to the number of stations with significant negative trend. However, the main climate types B and C are characterized by mostly significant negative trends. We cannot infer on stations with main climate types A and E because of the low number of stations. The observed patterns are also shown in a different form in Figure 16. We observe insignificant trends in approximately 50% of the stations, for main climate types A, B, C and D. However, the percentage of stations with positive significant trends is higher than

the percentage of negative significant trends for main climate type D (snow) and E (polar), while the opposite is true for main climate types A, B and C (all other climates).

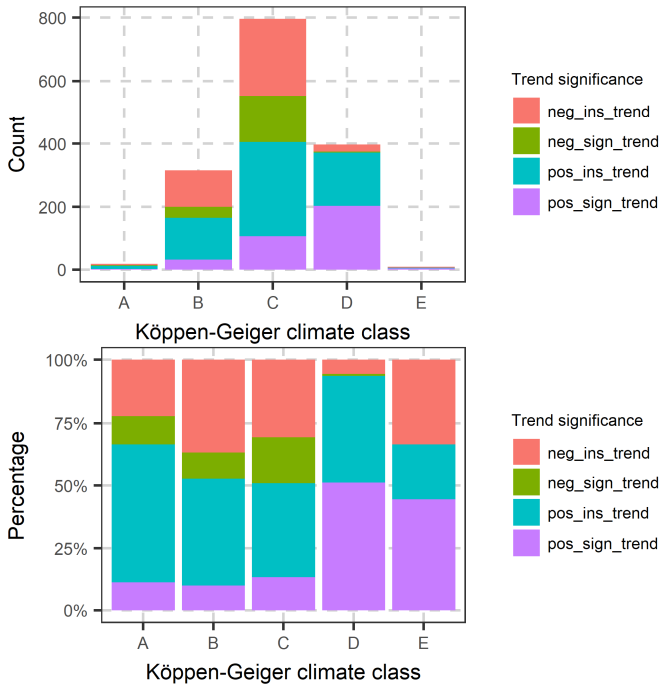

Figure 16. Number of stations with their trend significance (top) and percentages of stations for each type of trend significance (bottom) in each Köppen-Geiger climate class (grouping 2 of Table 2). Significance was estimated applying the MKt-LRD to the mean annual precipitation time series. The legend presents the sign of the trend (pos for positive and neg for negative) and its significance (sign for significant and ins for insignificant).

7. Summary, discussion and conclusions

We examined the long-range dependence properties of mean annual precipitation of 1 535 stations for the time period 1916-2015 and we tested the trends under the assumption of long-range dependence. Based on the maximum likelihood estimates of Hurst parameter H, which is a measure of long-range dependence, we found that the median value of H is equal to 0.56. This result is consistent with those of Fatichi et al. (2012) and Iliopoulou et al. (2017) regarding the LRD properties of the mean annual precipitation from instrumental measurements, which cover large part of the earth's land surface. Fatichi et al. (2012) estimated a median value H = 0.597 using an estimator based on the periodogram (Taqqu et al., 1995). Iliopoulou et al. (2017) estimated a mean H = 0.58.

In Section 5.3, we showed that the patterns of LRD are spatially clustered, while Fatichi et al. (2012) did not identify any geographical pattern. A spatially clustered

pattern was produced by predicting *H* using random forests. Furthermore, we showed that the location of the station and the climate type are the most important predictor variables of *H*, followed by the elevation of the station. The order of importance of the three former variables depends on the algorithm. The cforest algorithm estimates that the climate type is the most important, while due to its simultaneous handling of continuous and categorical variables can be considered more reliable than the random forests in estimating the variable importance. The combinations 6 and 21 of predictor variables, which include, respectively, the Cartesian coordinates and the geographic coordinates of the stations performs well in terms of the error metrics, but most importantly, their predictions had good correlation with the tested values. This correlation cannot be achieved with fitting a distribution to the set of the *H* values therefore the truncated normal distribution should be used with caution when modelling *H* and only as a prior that needs updating in a Bayesian setting conditional on the observed precipitation of the location. The inclusion of the climate type and the elevation (combinations 9, 24) improved further, albeit little, the performance of the random forests. However, this marginal improvement means that the information obtained from the geographic location of the station already includes the information of the climate type. Overall, the improvement from the truncated normal distribution to the inclusion of the geographic coordinates is 10%, which is not negligible. Considering that the median value ± 2 RMSE forms an approximately 95% confidence interval (Hunter and Goodchild, 1995), the improved confidence interval is narrower by 0.04. This improvement may sound negligible but it is not, as pointed out by Koutsoyiannis and Montanari (2007).

One could claim that the ability of the used algorithms to explain the spatial patterns of *H* is low. The low predictive performance could be due to the uncertainty in estimating *H*, or due to the weak spatial autocorrelation of *H*s. Here, we proved that for the given spatial autocorrelation, the best performance for a regular spatial pattern could not be more than 24%. The improvement in the real dataset was equal to 12%, which is sufficient in our opinion. Despite the uncertainty in the estimation of *H*, we identified spatial patterns, indicating that the influence of the errors in the estimation procedure was remedied by the large size of the sample. We proved that naïve methods are better than random forests when predicting in cases of complete randomness, therefore the better performance of the random forests here is due to the spatial

distribution of *H*s, and not due to their ability to predict better than naïve methods when complete randomness is present.

The overall result is that the random forest algorithm can predict well the LRD of the mean annual precipitation, when the location characteristics are used as predictor variables, while their performance is considerably better compared to the predictive ability of the simple distribution of *H*, particularly in terms of the correlation between the predicted and the estimated values. Therefore, the random forests can be used to predict *H* in locations without data or insufficient quantity of data and can serve as a substitute of spatial interpolation methods. Compared to spatial algorithms the random forests excel in combining information from distant locations through the common latitude, climate type and elevation variables, even if the spatial coverage is limited and non-uniform. Ordinary kriging has similar performance with the random forests when using the geographic coordinates as predictor variables, but it can't use other explanatory variables which proved to further slightly improve the predictive performance. The "Hurst_df.RData", which is the outcome of Part 4 of the Supplementary Information can be used by the interested reader to fit a model and predict *H* for other applications.

Regarding the presence of trends in the mean annual precipitation for the time period 1916-2015, it seems that the magnitude and sign of trends depend on the latitude and climate type of the station. The median of estimated trends was equal to 0.36 mm/year; however, it varies with the climate types in grouping 3 and the latitude. The MKt-LRD indicates that positive significant trends have been observed for the main climate type D (snow). In the other climate types the percentage of stations with positive significant trends was approximately equal to that of negative significant trends, while 50% of all stations do not exhibit significant trends at all.

A limitation of our study is that the random forests algorithm can predict values only if given values of the predictor variables are within the range of the fitting set. Thus, the limited availability of data prohibits the generalization of the method to regions and Köppen-Geiger climate classes which are not represented by the dataset. However, the random forests algorithm could provide information about the full conditional distribution of *H* (e.g. see Coulston et al., 2016; Meinshausen, 2006). These probabilistic predictions could be more appropriate for determining an initial prior distribution for *H* in a Bayesian setting compared e.g. to the uniform distribution in Tyralis et al. (2014) or

to a fitted distribution in a sample of estimated *H* values which is independent of the location. The random forests algorithm provides additional means to examine the effect of interaction between the predictor variables and *H*, which could give some insights on the natural explanation of the long-range dependence in precipitation. The latter issue is of high importance in hydrological science. To this end, non-linear transformations of the variables could be tested in addition to the exploratory data analysis presented here. Furthermore, the same fitting and testing procedure can be applied to the estimated trends and their estimated significances, to generalize the preliminary results of the trend analysis.

Appendix A Statistical software and supplementary information

The analyses and visualizations were performed in R Programming Language (R Core Team 2017). We used the contributed R packages caret (Kuhn 2008, Kuhn et al. 2017), devtools (Wickham and Chang, 2017), fBasics (Rmetrics Core Team et al., 2014), FGN (McLeod and Veenstra, 2014), gdata (Warnes et al., 2017), geoR (Ribeiro Jr and Diggle, 2016), geosphere (Hijmans, 2016a), ggplot2 (Wickham, 2016), gstat (Pebesma, 2004, Gräler et al., 2016), HKprocess (Tyralis, 2016), Hmisc (Harrell Jr et al., 2017), hydroTSM (Zambrano-Bigiarini, 2017), knitr (Xie, 2014; 2015; 2017), lubridate (Grolemund and Wickham, 2011), magrittr (Bache and Wickham, 2014), maps (Brownrigg et al., 2017), Matrix (Bates and Maechler, 2017), ncf (Bjornstad, 2016), party (Hothorn et al., 2017), plyr (Wickham, 2011), randomForest (Liaw and Wiener, 2002), raster (Hijmans, 2016b), readr (Wickham et al., 2017), reshape2 (Wickham, 2007), scales (Wickham, 2017), sp (Bivand et al., 2013b; Pebesma and Bivand, 2005), spdep (Bivand and Piras, 2015; Bivand et al., 2013a), tmvtnorm (Wilhelm and Manjunath, 2015), truncnorm (Trautmann et al., 2014), xts (Ryan and Ulrich, 2017), zoo (Zeileis and Grothendieck, 2005).

The code used for analysing the dataset is available online as supplementary information online at Tyralis (2017). The supplementary information also contains the 12 .html outcomes of the code, named Part 1, …, 12, the data and information about the data (in a readme.txt file in the main folder). The interested reader can use it to reproduce our analysis.

Acknowledgements: We thank Dr. Yiannis Markonis and Miss Georgia Papacharalampous for comments on an earlier version of this paper. Two anonymous reviewers' comments helped us improve the original version of the paper.

Funding information: The authors received no funding for this research, which was performed for scientific curiosity.

References

[1] Aggarwal, C.C., 2017. Outlier Analysis. Springer International Publishing. https://doi.org/10.1007/978-3-319-47578-3.

[2] Alexander, D.L.J., Tropsha, A., Winkler, D.A., 2015. Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. J. Chem. Inf. Model. 55 (7), 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206.

[3] Alexander, L.V., Zhang, X., Peterson, T.C., Caesar, J., Gleason, B., Klein Tank, A.M.G., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Rupa Kumar, K., Revadekar, J., Griffiths, G., Vincent, L., Stephenson, D.B., Burn, J., Aguilar, E., Brunet, M., Taylor, M., New, M., Zhai, P., Rusticucci, M., Vazquez-Aguirre, J.L., 2006. Global observed changes in daily climate extremes of temperature and precipitation. J. Geophys. Res. 111 (D05109). https://doi.org/10.1029/2005JD006290.

[4] Allan, R.P., Liu, C., Zahn, M., Lavers, D.A., Koukouvagias, E., Bodas-Salcedo, A., 2014. Physically Consistent Responses of the Global Atmospheric Hydrological Cycle in Models and Observations In: Bengtsson, L., Bonnet, R.M., Calisto, M., Destouni, G., Gurney, R., Johannessen, J., Kerr, Y., Lahoz, W.A., Rast, M. (Eds.), The Earth's Hydrological Cycle. Springer Netherlands, pp. 533–552. https://doi.org/10.1007/978-94-017-8789-5_4.

[5] Allen, M.R., Ingram, W.J., 2002. Constraints on future changes in climate and the hydrologic cycle. Nature 419, 224–232. https://doi.org/10.1038/nature01092.

[6] Alobaidi, M.H., Marpu, P.R., Ouarda, T.B.M.J., Chebana, F., 2015. Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework. Adv. Water Resour. 84, 103–111. https://doi.org/10.1016/j.advwatres.2015.07.019.

[7] Asadieh, B., Krakauer, N.Y., 2015. Global trends in extreme precipitation: climate models versus observations. Hydrol. Earth Syst. Sc. 19 (2), 877–891. https://doi.org/10.5194/hess-19-877-2015.

[8] Bache, S.M., Wickham, H., 2014. magrittr: A Forward-Pipe Operator for R. R package version 1.5. Available from: https://CRAN.R-project.org/package=magrittr.

[9] Barnett, V., Lewis, T., 1978. Outliers in statistical data. John Wiley & Sons.

[10] Bates, D., Maechler, M., 2017. Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-11. Available from: https://CRAN.R-project.org/package=Matrix.

[11] Beguería, S., Vicente-Serrano, S.M., Tomás-Burguera, M., Maneta, M., 2016. Bias in the variance of gridded data sets leads to misleading conclusions about changes in climate variability. Int. J. Climatol. 36 (9), 3413–3422. https://doi.org/10.1002/joc.4561.

[12] Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25 (2), 197–227. https://doi.org/10.1007/s11749-016-0481-7.

[13] Bierkens, M.F.P., 2015. Global hydrology 2015: State, trends, and directions. Water Resour. Res. 51 (7), 4923–4947. https://doi.org/10.1002/2015WR017173.

[14] Bivand, R.S., Piras, G., 2015. Comparing Implementations of Estimation Methods for Spatial Econometrics. J. Stat. Softw. 63 (18). https://doi.org/10.18637/jss.v063.i18.

[15] Bivand, R.S., Hauke, J., Kossowski, T., 2013a. Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods. Geogr. Anal. 45 (2), 150–179. https://doi.org/10.1111/gean.12008.

[16] Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., 2013b. Applied Spatial Data Analysis with R. Springer-Verlag New York. https://doi.org/10.1007/978-1-4614-7618-4.

[17] Bjornstad, O.N., 2016. ncf: Spatial Nonparametric Covariance Functions. R package version 1.1-7. Available from: https://CRAN.R-project.org/package=ncf.

[18] Blanchet, J., Marty, C., Lehning, M., 2009. Extreme value statistics of snowfall in the Swiss Alpine region. Water Resour. Res. 45 (W05424). https://doi.org/10.1029/2009WR007916.

[19] Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

[20] Brownrigg, R., Minka, T.P., Deckmyn, A., 2017. maps: Draw Geographical Maps. R package version 3.2.0. Available from: https://CRAN.R-project.org/package=maps.

[21] Bunde, A., Büntgen, U., Ludescher, J., Luterbacher, J., von Storch, H., 2013. Is there memory in precipitation?. Nat. Clim. Change 3, 174–175. https://doi.org/10.1038/nclimate1830.

[22] Coulston, J.W., Blinn, C.E., Thomas, V.A., Wynne, R.H., 2016. Approximating Prediction Uncertainty for Random Forest Regression Models. Photogramm. Eng. Rem. S. 82 (3), 189–197. https://doi.org/10.14358/PERS.82.3.189.

[23] Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci. 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008.

[24] Dinpashoh, Y., Mirabbasi, R., Jhajharia, D., Abianeh, H.Z., Mostafaeipour, A., 2014. Effect of Short-Term and Long-Term Persistence on Identification of Temporal Trends. J. Hydrol. Eng. 19 (3), 617–625. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000819.

[25] Dyrrdal, A.V., Skaugen, T., Stordal, F., Førland, E.J., 2016. Estimating extreme areal precipitation in Norway from a gridded dataset. Hydrolog. Sci. J. 61 (3), 483–494. https://doi.org/10.1080/02626667.2014.947289.

[26] Eagleson, P.S., 1986. The emergence of global-scale hydrology. Water Resour. Res. 22 (9S), 6S–14S. https://doi.org/10.1029/WR022i09Sp0006S.

[27] Eagleson, P.S., 1994. The evolution of modern hydrology (from watershed to continent in 30 years). Adv. Water Resour. 17 (1–2), 3–18. https://doi.org/10.1016/0309-1708(94)90019-1.
[28] Ehsanzadeh, E., Adamowski, K., 2010. Trends in timing of low stream flows in Canada: impact of autocorrelation and long-term persistence. Hydrol. Process. 24 (8), 970–980. https://doi.org/10.1002/hyp.7533.
[29] Fathian, F., Dehghan, Z., Bazrkar, M.H., Eslamian, S., 2016. Trends in hydrological and climatic variables affected by four variations of the Mann-Kendall approach in Urmia Lake basin, Iran. Hydrolog. Sci. J. 61 (5), 892–904. https://doi.org/10.1080/02626667.2014.932911.
[30] Fatichi, S., Ivanov, V.Y., Caporali, E., 2012. Investigating Interannual Variability of Precipitation at the Global Scale: Is There a Connection with Seasonality?. J. Climate 25, 5512–5523. https://doi.org/10.1175/JCLI-D-11-00356.1.
[31] Gräler, B., Pebesma, E.J., Heuvelink, G., 2016. Spatio-Temporal Interpolation using gstat. The R Journal 8 (1), 204–218.
[32] Gramatica, P., Sangion, A., 2016. A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology. J. Chem. Inf. Model. 56 (6), 1127–1131. https://doi.org/10.1021/acs.jcim.6b00088.
[33] Grolemund, G., Wickham, H., 2011. Dates and Times Made Easy with lubridate. J. Stat. Softw. 40 (3). https://doi.org/10.18637/jss.v040.i03.
[34] Gu, G., Adler, R.F., 2015. Spatial Patterns of Global Precipitation Change and Variability during 1901–2010. J. Climate 28, 4431–4453. https://doi.org/10.1175/JCLI-D-14-00201.1.
[35] Hamed, K.H., 2008. Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis. J. Hydrol. 349 (3–4), 350–363. https://doi.org/10.1016/j.jhydrol.2007.11.009.
[36] Harrell Jr, F.E., with contributions from Dupont C and many others, 2017. Hmisc: Harrell Miscellaneous. R package version 4.0-3. Available from: https://CRAN.R-project.org/package=Hmisc.
[37] Hartmann, D.L., Klein Tank, A.M.G., Rusticucci, M., Alexander, L.V., Brönnimann, S., Charabi, Y., Dentener, F.J., Dlugokencky, E.J., Easterling, D.R., Kaplan, A., Soden, B.J., Thorne, P.W., Wild, M., Zhai, P.M., 2013. Observations: Atmosphere and Surface. In: Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
[38] Hawkins, D.M., 1980. Identification of Outliers. Springer Netherlands. https://doi.org/10.1007/978-94-015-3994-4.
[39] Hijmans, R.J., 2016a. geosphere: Spherical Trigonometry. R package version 1.5-5. Available from: https://CRAN.R-project.org/package=geosphere.
[40] Hijmans, R.J., 2016b. raster: Geographic Data Analysis and Modeling. R package version 2.5-8. Available from: https://CRAN.R-project.org/package=raster.
[41] Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., 2017. party: A Laboratory for Recursive Partytioning. R package version 1.2-3. Available from: https://CRAN.R-project.org/package=party.

[42] Hunter, G.J., Goodchild, M.F., 1995. Dealing with Error in Spatial Databases: A simple Case Study. Photogramm. Eng. Rem. S. 61 (5), 529–537.
[43] Iliopoulou, T., Papalexiou, S.M., Markonis, Y., Koutsoyiannis, D., 2017. Revisiting long-range dependence in annual precipitation. J. Hydrol.. https://doi.org/10.1016/j.jhydrol.2016.04.015.
[44] Jeong, H.D.J., Lee, J.S.R., McNickle, D., Pawlikowski, K., 2007. Comparison of various estimators in simulated FGN. Simul. Model Pract. Th. 15 (9), 1173–1191. https://doi.org/10.1016/j.simpat.2007.08.004.
[45] Kanevski, M., Demyanov, V., 2015. Statistical learning in geoscience modelling: Novel algorithms and challenging case studies. Comput. Geosci. 85 (Part B), 1–2. https://doi.org/10.1016/j.cageo.2015.10.007.
[46] Khaliq, M.N., Ouarda, T.B.M.J., Gachon, P., 2009. Identification of temporal trends in annual and seasonal low flows occurring in Canadian rivers: The effect of short- and long-term persistence. J. Hydrol. 369 (1–2), 183–197. https://doi.org/10.1016/j.jhydrol.2009.02.045.
[47] Kidd, C., Huffman, G., 2011. Global precipitation measurement. Meteorol. Appl. 18, 334–353. https://doi.org/10.1002/met.284.
[48] Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rybel, F., 2006. World Map of the Köppen-Geiger climate classification updated. Meteorol. Z. 15 (3), 259–263. https://doi.org/10.1127/0941-2948/2006/0130.
[49] Koutsoyiannis, D., 2002. The Hurst phenomenon and fractional Gaussian noise made easy. Hydrolog. Sci. J. 47 (4), 573–595. https://doi.org/10.1080/02626660209492961.
[50] Koutsoyiannis, D., 2003. Climate change, the Hurst phenomenon, and hydrological statistics. Hydrolog. Sci. J. 48 (1), 3–24. https://doi.org/10.1623/hysj.48.1.3.43481.
[51] Koutsoyiannis, D., 2004. Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records. Hydrolog. Sci. J. 49 (4), 591–610. https://doi.org/10.1623/hysj.49.4.591.54424.
[52] Koutsoyiannis, D., 2006. Nonstationarity versus scaling in hydrology. J. Hydrol. 324 (1–4), 239–254. https://doi.org/10.1016/j.jhydrol.2005.09.022.
[53] Koutsoyiannis, D., Montanari, A., 2007. Statistical analysis of hydroclimatic time series: Uncertainty and insights. Water Resour. Res. 43 (W05429). https://doi.org/10.1029/2006WR005592.
[54] Koutsoyiannis, D., 2010. HESS Opinions "A random walk on water". Hydrol. Earth Syst. Sc. 14, 585–601. https://doi.org/10.5194/hess-14-585-2010.
[55] Koutsoyiannis, D., Montanari, A., 2014. Negligent killing of scientific concepts: the stationarity case. Hydrolog. Sci. J. 60 (7–8), 1174–1183. https://doi.org/10.1080/02626667.2014.959959.
[56] Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28 (5), 1–26. https://doi.org/10.18637/jss.v028.i05.
[57] Kuhn, M., Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2017. caret: Classification and Regression Training. R package version 6.0-76. Available from: https://CRAN.R-project.org/package=caret.
[58] Kumar, S., Merwade, V., Kam, J., Thurner, K., 2009. Streamflow trends in Indiana: Effects of long term persistence, precipitation and subsurface drains. J. Hydrol. 374 (1–2), 171–183. https://doi.org/10.1016/j.jhydrol.2009.06.012.

[59] Kumar, S., Merwade, V., Kinter III, J.L., Niyogi, D., 2013. Evaluation of Temperature and Precipitation Trends and Long-Term Persistence in CMIP5 Twentieth-Century Climate Simulations. J. Climate 26, 4168–4185. https://doi.org/10.1175/JCLI-D-12-00259.1.

[60] Lambert, F.H., Stine, A.R., Krakauer, N.Y., Chiang, J.C.H., 2008. How Much Will Precipitation Increase With Global Warming?. Eos, Transactions, AGU 89 (21), 193–194. https://doi.org/10.1029/2008EO210001.

[61] Leuenberger, M., Kanevski, M., 2015. Extreme Learning Machines for spatial environmental data. Comput. Geosci. 85 (Part B), 64–73. https://doi.org/10.1016/j.cageo.2015.06.020.

[62] Li, Q., Chen, Y., Shen, Y., Li, X., Xu, J., 2011. Spatial and temporal trends of climate change in Xinjiang, China. J. Geogr. Sci. 21 (6), 1007–1018. https://doi.org/10.1007/s11442-011-0896-8.

[63] Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

[64] de Lima, M.I.P., Krajewski, W.F., de Lima, J.L.M.P., 2012. Space–time precipitation from urban scale to global change. Adv. Water Resour. 45:1. https://doi.org/10.1016/j.advwatres.2012.06.001.

[65] Lins, H.F., Cohn, T.A., 2011. Stationarity: Wanted Dead or Alive?. J. Am. Water Resour. As. 47 (3), 475–480. https://doi.org/10.1111/j.1752-1688.2011.00542.x.

[66] Liu, L., Xu, Z.X., Huang, J.X., 2012. Spatio-temporal variation and abrupt changes for major climate variables in the Taihu Basin, China. Stoch. Env. Res. Risk A. 26 (6), 777–791. https://doi.org/10.1007/s00477-011-0547-8.

[67] Markonis, Y., Koutsoyiannis, D., 2016. Scale-dependence of persistence in precipitation records. Nat. Clim. Change 6, 399–401. https://doi.org/10.1038/nclimate2894.

[68] McLeod, A.I., Veenstra, J., 2014. FGN: Fractional Gaussian Noise and power law decay time series model fitting. R package version 2.0-12. Available from: https://CRAN.R-project.org/package=FGN.

[69] Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

[70] Menne, M.J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R.S., Gleason, B.E., Houston, T.G., 2012a. Global Historical Climatology Network - Daily (GHCN-Daily), Version 3.22. NOAA National Climatic Data Center. Available from: https://doi.org/10.7289/V5D21VHZ [access date:2016-09-02].

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

[72] Moran, P.A.P., 1950. Notes on Continuous Stochastic Phenomena. Biometrika 37 (1–2), 17–23. https://doi.org/10.2307/2332142.

[73] Morin, E., 2011. To know what we cannot know: Global mapping of minimal detectable absolute trends in annual precipitation. Water Resour. Res. 47 (W07505). https://doi.org/10.1029/2010WR009798.

[74] Munshi, J., 2015. The Hurst Exponent of Precipitation. Available at Social Science Research Network. https://doi.org/10.2139/ssrn.2695753.

[75] Monjo, R., Martin-Vide, J., 2016. Daily precipitation concentration around the world according to several indices. Int. J. Climatol. 36 (11), 3828–3838. https://doi.org/10.1002/joc.4596.

[76] Nasrollahi, N., AghaKouchak, A., Cheng, L., Damberg, L., Phillips, T.J., Miao, C., Hsu, K., Sorooshian, S., 2015. How well do CMIP5 climate simulations replicate historical trends and patterns of meteorological droughts?. Water Resour. Res. 51, 2847–2864. https://doi.org/10.1002/2014WR016318.

[77] New, M., Todd, M., Hulme, M., Jones, P., 2001. Precipitation measurements and trends in the twentieth century. Int. J. Climatol. 21 (15), 1889–1922. https://doi.org/10.1002/joc.680.

[78] O'Connell, P.E., Koutsoyiannis, D., Lins, H.F., Markonis, Y., Montanari, A., Cohn, T.A., 2015. The scientific legacy of Harold Edwin Hurst (1880–1978). Hydrolog. Sci. J. 61 (9), 1571–1590. https://doi.org/10.1080/02626667.2015.1125998.

[79] Papalexiou, S.M., Koutsoyiannis, D., 2013. Battle of extreme value distributions: A global survey on extreme daily rainfall. Water Resour. Res. 49, 187–201. https://doi.org/10.1029/2012WR012557.

[80] Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012.

[81] Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R News 5 (2), 9–13.

[82] R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

[83] Ragulina, G., Reitan, T., 2017. Generalized Extreme Value's shape parameter and its nature for extreme precipitation using long-time series and Bayesian approach. Hydrolog. Sci. J. 62 (6), 863–879. https://doi.org/10.1080/02626667.2016.1260134.

[84] Rea, W., Oxley, L., Reale, M., Brown, J., 2013. Not all estimators are born equal: The empirical properties of some estimators of long memory. Math. Comput. Simulat. 93, 29–42. https://doi.org/10.1016/j.matcom.2012.08.005.

[85] Ribeiro Jr, P.J., Diggle, P.J., 2016. geoR: Analysis of Geostatistical Data. R package version 1.7-5.2. Available from: https://CRAN.R-project.org/package=geoR.

[86] Rmetrics Core Team, Wuertz, D., Setz, T., Chalabi, Y., 2014. fBasics: Rmetrics - Markets and Basic Statistics. R package version 3011.87. Available from: https://CRAN.R-project.org/package=fBasics.

[87] Ryan, J.A., Ulrich, J.M., 2017. xts: eXtensible Time Series. R package version 0.10-0. Available from: https://CRAN.R-project.org/package=xts.

[88] Sagarika, S., Kalra, A., Ahmad, S., 2014. Evaluating the effect of persistence on long-term trends and analyzing step changes in streamflows of the continental United States. J. Hydrol. 517, 36–53. https://doi.org/10.1016/j.jhydrol.2014.05.002.

[89] Smith, T.M., Arkin, P.A., Ren, L., Shen, S.S.P., 2012. Improved Reconstruction of Global Precipitation since 1900. J. Atmos. Ocean Tech. 29, 1505–1517. https://doi.org/10.1175/JTECH-D-12-00001.1.

[90] Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(25). https://doi.org/10.1186/1471-2105-8-25.

[91] Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional Variable Importance for Random Forests. BMC Bioinformatics 9 (307). https://doi.org/10.1186/1471-2105-9-307.

[92] Sun, Q., Kong, D., Miao, C., Duan, Q., Yang, T., Ye, A., Di, Z., Gong, W., 2014. Variations in global temperature and precipitation for the period of 1948 to 2010. Environ. Monit. Assess. 186 (9), 5663–5679. https://doi.org/10.1007/s10661-014-3811-9.

[93] Taqqu, M.S., Teverovsky, V., Willinger, W., 1995. Estimators for long-range dependence: an empirical study. Fractals 3 (4), 785–798. https://doi.org/10.1142/S0218348X95000692.

[94] Tegos, A., Tyralis, H., Koutsoyiannis, D., Hamed, K.H., 2017. An R function for the estimation of trend significance under the scaling hypothesis- application in PET parametric annual time series. Open Water Journal 4 (1):66–71.

[95] Trautmann, H., Steuer, D., Mersmann, O., Bornkamp, B., 2014. truncnorm: Truncated normal distribution. R package version 1.0-7. Available from: https://CRAN.R-project.org/package=truncnorm.

[96] Trenberth, K.E., 2011. Changes in precipitation with climate change. Clim. Res. 47 (1–2), 123–138. https://doi.org/10.3354/cr00953.

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

[98] Tyralis, H., 2017. Supplementary information for the paper "On the long-range dependence properties of annual precipitation using a global network of instrumental measurements". figshare. https://doi.org/10.6084/m9.figshare.4892447.v1.

[99] Tyralis, H., Koutsoyiannis, D., 2011. Simultaneous estimation of the parameters of the Hurst-Kolmogorov stochastic process. Stoch. Env. Res. Risk A. 25 (1), 21–33. https://doi.org/10.1007/s00477-010-0408-x.

[100] Tyralis, H., Koutsoyiannis, D., 2014. A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. Clim. Dynam. 42 (11–12), 2867–2883. https://doi.org/10.1007/s00382-013-1804-y.

[101] Valle, M.A.V., García, G.M., Cohen, I.S., Oleschko, L.K., Corral, J.A.R., Korvin, G., 2013. Spatial Variability of the Hurst Exponent for the Daily Scale Rainfall Series in the State of Zacatecas, Mexico. J. Appl. Meteorol. Clim. 52, 2771–2780. https://doi.org/10.1175/JAMC-D-13-0136.1.

[102] Villarini, G., Smith, J.A., 2010. Flood peak distributions for the eastern United States. Water Resour. Res. 46 (W06504). https://doi.org/10.1029/2009WR008395.

[103] Warnes, G.R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., Rogers, J. and others, 2017. gdata: Various R Programming Tools for Data Manipulation. R package version 2.18.0. Available from: https://CRAN.R-project.org/package=gdata.

[104] Wentz, F.J., Ricciardulli, L., Hilburn, K., Mears, C., 2007. How Much More Rain Will Global Warming Bring?. Science 317 (5835), 233–235. https://doi.org/10.1126/science.1140746.

[105] van Wijngaarden, W.A., Syed, A., 2015. Changes in annual precipitation over the Earth's land mass excluding Antarctica from the 18th century to 2013. J. Hydrol. 531 (Part 3), 1020–1027. https://doi.org/10.1016/j.jhydrol.2015.11.006.

[106] Wickham, H., 2007. Reshaping Data with the reshape Package. J. Stat. Softw. 21 (12). https://doi.org/10.18637/jss.v021.i12.

[107] Wickham, H., 2011. The Split-Apply-Combine Strategy for Data Analysis. J. Stat. Softw. 40 (1). https://doi.org/10.18637/jss.v040.i01.
[108] Wickham, H., 2016. ggplot2. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4.
[109] Wickham, H., 2017. scales: Scale Functions for Visualization. R package version 0.5.0. Available from: https://CRAN.R-project.org/package=scales.
[110] Wickham, H., Chang, W., 2017. devtools: Tools to Make Developing R Packages Easier. R package version 1.13.3. Available from: https://CRAN.R-project.org/package=devtools.
[111] Wickham, H., Hester, J., Francois, R., Jylänki, J., Jørgensen, M., 2017. readr: Read Rectangular Text Data. R package version 1.1.1. Available from: https://CRAN.R-project.org/package=readr.
[112] Wilhelm, S., Manjunath, B.G., 2015. tmvtnorm: Truncated Multivariate Normal and Student t Distribution. R package version 1.4-10. Available from: https://CRAN.R-project.org/package=tmvtnorm.
[113] Xie, Y., 2014. knitr: A Comprehensive Tool for Reproducible Research in R. In: Stodden, V., Leisch, F., Peng, R.D. (Eds.), Implementing Reproducible Computational Research. Chapman and Hall/CRC.
[114] Xie, Y., 2015. Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC.
[115] Xie, Y., 2017. knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.17. Available from: https://CRAN.R-project.org/package=knitr.
[116] Zamani, R., Mirabbasi, R., Abdollahi, S., Jhajharia, D., 2017. Streamflow trend analysis by considering autocorrelation structure, long-term persistence, and Hurst coefficient in a semi-arid region of Iran. Theor. Appl. Climatol. 129 (1–2), 33–45. https://doi.org/10.1007/s00704-016-1747-4.
[117] Zambrano-Bigiarini, M., 2017. hydroTSM: Time Series Management, Analysis and Interpolation for Hydrological Modelling. R package version 0.5-1. Available from: https://doi.org/10.5281/zenodo.839864.
[118] Zeileis, A., Grothendieck, G., 2005. zoo: S3 Infrastructure for Regular and Irregular Time Series. J. Stat. Softw. 14 (6), 1–27. https://doi.org/10.18637/jss.v014.i06.
[119] Zhang, X., Zwiers, F.W., Hegerl, G.C., Lambert, F.H., Gillett, N.P., Solomon, S., Stott, P.A., Nozawa, T., 2007. Detection of human influence on twentieth-century precipitation trends. Nature 448, 461–465. https://doi.org/10.1038/nature06025.