Peer-reviewed version available at *Journal of Hydrology* **2019**, *574*, 628-645; doi:10.1016/j.jhydrol.2019.04.070

How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset

Hristos Tyralis[1], Georgia Papacharalampous[2], and Sarintip Tantanee[3]

[1]Air Force Support Command, Hellenic Air Force, Elefsina Air Base, 192 00 Elefsina, Greece (https://orcid.org/0000-0002-8932-4997)

[2]Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Iroon Polytechniou 5, 157 80 Zografou, Greece (https://orcid.org/0000-0001-5446-954X)

[3]Department of Civil Engineering, Engineering Faculty, Naresuan University, Nakhonsawan – Phitsanulok Rd., 650 00 Phitsanulok, Thailand

Corresponding author: Hristos Tyralis (montchrister@gmail.com)

This is a pre-print of an article published in Journal of Hydrology (2019). The final authenticated version is available online at: https://doi.org/10.1016/j.jhydrol.2019.04.070.

Please cite this article as:

Tyralis H, Papacharalampous G, Tantanee S (2019) How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset. Journal of Hydrology 574:628–645. doi:10.1016/j.jhydrol.2019.04.070

Abstract: The finding of important explanatory variables for the location parameter and the scale parameter of the generalized extreme value (GEV) distribution, when the latter is used for the modelling of annual streamflow maxima, is known to have reduced the uncertainties in inferences, as estimated through regional flood frequency analysis frameworks. However, important explanatory variables have not been found for the GEV shape parameter, despite its critical significance, which stems from the fact that it determines the behaviour of the upper tail of the distribution. Here we examine the nature of the shape parameter by revealing its relationships with basin attributes. We use a dataset that comprises information about daily streamflow and forcing, climatic indices, topographic, land cover, soil and geological characteristics of 591 basins with minimal human influence in the contiguous United States. We propose a framework that uses random forests and linear models to find (a) important predictor variables of the shape parameter and (b) an interpretable model with high predictive performance. The process

of study comprises of assessing the predictive performance of the models, selecting a parsimonious predicting model and interpreting the results in an ad-hoc manner. The findings suggest that the shape parameter mostly depends on climatic indices, while the selected prediction model results in more than 20% higher accuracy in terms of RMSE compared to a naïve approach. The implications are important, since incorporating the regression model into regional flood frequency analysis frameworks can considerably reduce the predictive uncertainties.

Keywords: CAMELS; flood frequency; hydrological signatures; extreme value theory; random forests; spatial modelling

1. Introduction

1.1 Flood frequency analysis and hydrological signatures

Floods are one of the most important natural hazards (see e.g. Odry and Arnaud 2017), with a large part of the hydrological literature being devoted to their study (see e.g. Parkes and Demeritt 2016). Flood frequency analysis (FFA) is a statistical approach aiming at determining the magnitude of floods for a predefined return period (Thorarinsdottir et al. 2018). The simplest approach in FFA is to model data at a single site (at-site FFA or local modelling, Thorarinsdottir et al. 2018). However, when at-site data are limited, the models' results can be very uncertain. To obtain accurate results, information from adjacent or similar sites can be exploited. This approach is termed regional flood frequency analysis (RFFA, Thorarinsdottir et al. 2018). Transfer of information from one catchment to the other can be achieved by purely data-based or by rainfall-runoff models. A more detailed classification of the RFFA models can be found in Odry and Arnaud (2017).

The RFFA methodologies are related to the initiative for Predictions in Ungauged Basins (PUB) of the International Association of Hydrological Sciences (IAHS) (Hrachowitz et al. 2013) in the sense that information from gauged basins can be used to decrease the uncertainties of predictions in sparsely gauged basins or estimate uncertainties in ungauged basins (e.g. Bourgin et al. 2015). The investigation of this practice for the case of floods is particularly recommended by Stedinger and Griffis (2008). Furthermore, they are related to the notion of hydrological signatures. The latter are defined as "*index values derived from observed or modelled series of hydrological data such as rainfall, flow or soil moisture*" (McMillan et al. 2017; see also the discussion in

Gupta et al. (2008), and Wagener and Montanari (2011)). From a statistician's point of view the hydrological signatures are values of a statistic; therefore, they summarize the information provided by the data.

Hydrological signatures can be used for hydrological model calibration (Shafii and Tolson 2015). From a statistical point of view, this approach is similar to the data-based approaches mentioned earlier. Hydrological signatures may depend on local climatic conditions, as well as on attributes related to the local topography, land cover, soil and geology. Attempts have been made to find such relationships using regression and/or classification methods (Viglione et al. 2013b; Singh et al. 2014; Beck et al. 2015; Addor et al. 2018). Frameworks have also been developed for computing the uncertainty in the estimation of hydrological signatures (Westerberg and McMillan 2015; Westerberg et al. 2016).

1.2 Frameworks with separate parameter estimation

A common approach in the class of at-site data-based models of FFA is to model the annual (or seasonal) discharge block maxima (peak discharges) with the generalized extreme value (GEV) distribution. This approach is supported by empirical evidence (Vogel and Wilson 1996), albeit other distributions have also been considered in the literature (Vogel and Wilson 1996; Griffis and Stedinger 2007). The modelling choice of the GEV distribution is justified by limiting theorems and constitutes a common ground for hydrologists (Coles 2001; Reiss and Thomas 2007, pp. 337, 338). The cumulative distribution function of the GEV distribution is given by the following equation (Coles 2001, pp.47, 48; Dey et al. 2016, see also Stedinger et al. 1993; Hosking and Wallis 1997; Koutsoyiannis 2004 for equivalent expressions of the GEV).

$$F(x|\boldsymbol{\theta}) := \exp(-(1 + k((x-\mu)/\sigma))_+^{-1/k}),\ \boldsymbol{\theta} = (\mu, \sigma, k),\ \sigma > 0 \qquad (1)$$

Here μ is the location parameter, σ is the scale parameter and k is the shape parameter. The shape parameter determines the behaviour (or shape) of the tail of the distribution. In particular, higher values of k result in heavier tails. Bayesian frameworks for flood frequency modelling based on eq. (1) are available when streamflow data of a basin are given (Northrop and Attalides 2016). While such models quantify the probability of extreme events rigorously, the estimated posterior regions, confidence interval or predictive intervals are wide (see e.g. Xu et al. 2010).

Data-based RFFA models can be used to decrease the uncertainties related to the above quantities. Here we are interested in using data-based RFFA models which separately model the parameters of the GEV distribution as functions of the basin attributes. Regression-based models and, in particular, models using parameter regression techniques (see e.g. Ahn and Palmer 2016). The parameter regression techniques can be viewed as subcases of the Generalized Additive Models for Location Scale and Shape (GAMLSS, Rigby and Stasinopoulos 2005), albeit the software connected with the latter method is restricted to certain types of implemented regression techniques (linear and non-linear). In this category of models, the parameters are modelled separately as functions of the attributes of the gauged basins by mostly (but not exclusively) using linear models. The information is transferred to the ungauged basins through the prediction made by the fitted regression model. This category of models is similar to another category of models, in which quantiles of the GEV distribution (i.e. flood magnitudes for a given return period) are directly computed by regression models. This last category of models has been extensively investigated and includes linear (see e.g. Stedinger and Tasker 1985) and non-linear models. Examples of this type of non-linear models are quantile regression (see e.g. Haddad et al. 2012, Ouali et al. 2016), generalized additive models (GAM; see e.g. Ouali et al. 2017, Rahman et al. 2018) and artificial neural networks (ANN; see e.g. Ouali et al. 2017). Such models can be applied directly to the dataset or after partitioning the dataset into homogenous regions (see the literature reviews in Gaume et al. 2010; Merz and Blöschl 2005; Requena et al. 2017).

Separate modelling of the parameters of the GEV is also required by Bayesian models (see e.g. Lima and Lall 2010; Yan and Moradkhani 2015, 2016; Wu et al. 2018), while comprehensive relevant frameworks have been proposed by Northrop (2004), Viglione et al. (2013a) and Lima et al. (2016). In this category of models, the parameters are separately modelled as linear functions of basin attributes and the linear models are inserted in the final model. Posterior distributions of the parameters given the available data, as well as predictive intervals for the variable of interest, are then computed.

1.3 Relationship with basin's attributes

To assist in the design of the methods presented in Section 1.2, as well as to understand how extreme events depend on the attributes of the basins studied, investigation of large discharge datasets from Central Europe, UK and USA focused on the estimates of the $\boldsymbol{\theta}$

parameters and their empirical relationship with attributes of the basins (Northrop 2004; Villarini and Smith 2010; Smith et al. 2011; Villarini et al. 2011a, b, 2012).

The most frequently met μ and σ parameterizations include their relation with the area of the basin. Lima et al. (2016) justify this parameterization based on theoretical and empirical prescriptions, and subsequently cite the relevant studies of Gupta and Waymire (1990), Gupta et al. (1994, 2007), Gupta and Dawdy (1995), Morrison and Smith (2002), Northrop (2004), Lima and Lall (2010), Villarini and Smith (2010) and Villarini et al. (2011b). Parameterization of the coefficient of variation cv := σ/μ, which depends on μ and σ, is also a frequent subject in the literature (see e.g. Blöschl and Sivapalan 1997; Vogel and Sankarasubramanian 2000; Morrison and Smith 2001; Kuzuha et al. 2009; Veneziano and Langousis 2010).

However, the k parameter in Lima et al. (2016) is modelled by a normal distribution with common mean across all sites; thus, it is independent on attributes of the basin. This choice is based on the studies of Gupta and Waymire (1990), Burlando and Rosso (1996) and Morrison and Smith (2002). On the other hand, He et al. (2015) conclude that it is worthwhile considering the effect of other catchment attributes than the area of the basin (such as meteorological and topological factors) in the estimation of the shape parameter. Moreover, Gvoždíková and Müller (2017) suggest the investigation of the relationship of major floods with extreme precipitation. Other studies also find unclear (slightly significant) relationships between the k parameter and other basin attributes (see Northrop 2004, Villarini and Smith 2010, and Villarini et al. 2011a, b; see also the discussion in Section 4).

The parameters of the GEV distribution fitted to the annual block maxima of streamflow are certainly related to the distribution of the daily streamflow, which could be considered its parent distribution. Attempts have been made to estimate a common type of distribution for the statistical modelling of daily streamflow (Blum et al. 2017). Nonetheless and despite the excellent fit of the proposed distribution, theoretical issues related to the dependence and the seasonality in the daily streamflow have not been treated.

1.4 Aim of the present study

The aim of the present study is (a) to present a framework that can be used to reveal relationships among the shape parameter of the GEV distribution when fitted to annual

discharge block maxima, and characteristics of the respective basin (in particular topographic characteristics, climatic indices, land cover characteristics, soil characteristics and geological characteristics), as well as (b) to better predict the shape parameter conditional on the basin's attributes. Obtained relationships, when incorporated in the regression-based or Bayesian frameworks presented in Section 1.3, can support the understanding of the mechanism behind the generation of floods and decrease the uncertainties of flood design. Concerning the discovered relationships, the findings of the present study are also original in comparison to previous studies in which the *k* parameter was examined.

To reveal such relationships we expand the methodology implemented by Tyralis et al. (2018) and Addor et al. (2018). In both these studies, random forests were used due to their excellent predictive performance and their ability to find important predictor variables (Biau and Scornet 2016). These two studies are also similar in their approach to finding spatial relationships, as they both use the Moran's *I* test for this purpose. The use of the Moran's *I* test is avoided here, because its common implementation requires the use of Euclidean distances, while the spatial behaviour of rivers can be examined in a framework based on river networks. In such frameworks, other types of distances are calculated. Random forests are a machine learning algorithm (Breiman 2001) of increasing interest in geosciences (e.g. Tyralis and Papacharalampous 2017; Papacharalampous and Tyralis 2018; Papacharalampous et al. 2018a, b). Tyralis et al. (2018) implemented this methodology with the aim to find important characteristics of precipitation and Addor et el. (2018) implemented this methodology to find such relationships for hydrological signatures.

While random forests are a flexible algorithm with high predictive power, it is less interpretable than linear regression models, since there is a trade-off between flexibility (and predictive power) of machine learning models and their interpretability (James et al. 2013, p.25). Thus, we enhance the implemented methodology by finding linear regression models with comparable predictive performance to random forests for this specific application. The framework is based on the following ideas:

a. The predictability of the parameter must first be investigated using a high performance predictive algorithm that is not affected by the presence of unimportant predictor variables, e.g. the herein implemented random forest algorithm.

b. Due to the large number of predictor variables, their importance must be computed by an appropriate algorithm, which here is again random forests.

c. If a linear model with similar predictive performance to the benchmark model exists, then it will be preferable, as it is more interpretable. To find such a model, the number of predictor variables is reduced in an automatic way.

d. Then a semi-automatic procedure that uses importance metrics for linear models and random forests, and examines combinations of predictor variables and their interactions, is implemented using the retained predictor variables. If the predictive performance of the linear model is similar with the performance of random forests, then both aims of the present study will be accomplished.

The new concepts introduced here compared to Tyralis et al. (2018) and Addor et al. (2018) is the examination of interactions, the use of importance metrics for linear models, as well as the extensive investigation of the latter through the inclusion of interactions. This investigation is benchmarked using random forests. The framework based on these ideas is presented in detail in Section 2.2.5. The proposed framework can better reveal possible relationships compared to previous studies. For instance, it can offer an improvement in comparison to Ahn and Palmer (2016), who exclusively used additive terms for predicting the k parameter, while (as results from the present study) interactions could result in a better model. The data used herein were obtained from the CAMELS dataset (Newman et al. 2015; Addor et al. 2017b) and are comprised of daily streamflow, precipitation and other basin attributes of 671 catchments in the contiguous United States (CONUS).

2. Data and methods

2.1 Data

The data used in the present study can be sourced from Newman et al. (2014) and Addor et al. (2017a). Moreover, their documentation is available in Newman et al. (2015) and Addor et al. (2017b), who created their dataset by combining data from Miller and White (1998), Hartmann and Moosdorf (2012), Gleeson et al. (2014), Thornton et al. (2014) and Pelletier et al. (2016). The CAMELS dataset is one of the largest datasets with respect to the number of included basins and the information provided for each basin; therefore, it is suitable for benchmarking purposes (Newman et al. 2017).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

Briefly, the dataset comprises information about daily streamflow and forcing for 671 small- to medium-sized basins. Here we used the precipitation forcing derived from the daily gridded Daymet dataset. The data span in the period 1980–2014. The 671 basins cover the entire CONUS with a wide range of hydroclimatic conditions and having minimal human influence. The catchment attributes used in the present study include topographic characteristics, climatic indices, land cover characteristics, soil characteristics and geological characteristics related to the basin of interest, and are presented in Table 1. Details on the collection of the datasets can be found in Newman et al. (2015) and Addor et al. (2017b), while the explanation of the attributes of Table 1 is presented in Appendix A.

Table 1. Attributes and respective abbreviations of the 671 basins in Addor et al. (2017b) used as predictor variables. The GEV attributes were estimated in the present study. A detailed description of the attributes can be found in Appendix A.

Attribute type	Values as -is		Transformed using log	
	Attribute	Abbreviation	Attribute	Abbreviation
Topographic			Mean elevation	elev_mean
			Mean slope	slope_mean
			Area	area_gages2
Geographical coordinates	Latitude	gauge_lat		
	Longitude	gauge_lon		
Climatic	Seasonality and timing of precipitation	p_seasonality	Mean daily precipitation	p_mean
	Fraction of precipitation falling as snow	frac_snow	Mean daily PET	pet_mean
			Frequency of high precipitation events	high_prec_freq
			Duration of high precipitation events	high_prec_dur
			Duration of low precipitation events	low_prec_dur
Land cover	Forest fraction	forest_frac		
	LAI maximum	lai_max		
	Green vegetation fraction difference	gvf_diff		
Soil	Soil depth	soil_depth_statsgo	Depth to bedrock	soil_depth_pelletier
	Maximum water content	max_water_content	Volumetric porosity	soil_porosity
	Sand fraction	sand_frac	Saturated hydraulic conductivity	soil_conductivity
	Silt fraction	silt_frac		
	Clay fraction	clay_frac		
	Water fraction	water_frac		
	Organic fraction	organic_frac		
Geology	Fraction of carbonate rocks	carb_rocks_frac		
	Subsurface porosity	geol_porosity		
	Subsurface permeability	geol_permeability		
GEV attributes	Precipitation shape parameter	shape_par_prcp	Precipitation location parameter	loc_par_prcp

The data were preprocessed according to the following procedure. Firstly, the data in the year 2014 were omitted, because most basins included many missing streamflow values in this year (e.g. 17 basins without data in the year 2014). Therefore, the period 1980–2013 was examined (34 years of data for each basin). Attributes not used in the analysis are presented in Table 2 for the sake of completeness. The root depth and the second most common geologic class attributes were not used, because of their many missing values (24 and 138 respectively). Categorical attributes were not included in the analysis, since they were found in general unimportant for predicting the shape parameter in a preliminary analysis using importance metrics for random forests (see Section 2.2.2; for instance, the most common geological class was the most important

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

categorical variable; still, it was ranked very low, i.e. 15th in overall), while their inclusion in the linear model was not possible in the testing procedure. The latter is again due to the appearance of missing values in the cross-validation procedure (see Section 2.2.4). Some variables were not considered, because they were highly correlated (correlations higher than 0.9) with other variables; therefore, the inclusion of both variables would not increase the performance of the fitted models. These variables are presented in Table 2 under the term collinearity. From the remaining dataset, basins with more than 100 missing daily streamflow values in the period 1980–2013 or with at least one missing attribute were omitted. The maximum likelihood estimates (Coles 2001, pp. 55, 56) of the parameters of eq. (1) for the annual block maxima of daily streamflow and precipitation were calculated and they were named GEV attributes in Table 1. In particular, the maximum likelihood estimates were obtained using the `SpatialExtremes` R package (Ribatet 2018). Basins with $k \geq 1$ (k denotes the estimate of the shape parameter from hereinafter) were omitted, since $E[\underline{x}]$ (i.e. the first moment) is not defined for $k \geq 1$ (Dey et al. 2016). Here $\underline{x}$ denotes the GEV random variable that models the streamflow annual block maxima. The 591 remained basins are presented in Figure 1, while the histogram of the k estimates is presented in Figure 2. The correlogram of the remained predictor variables is presented in Figure 3.

Table 2. Abbreviations of the attributes of the 671 basins in Addor et al. (2017b) not included in the analysis for reasons explained in the heading of the Table. The GEV attributes were estimated in the present study. The names of the attributes and a detailed description can be found in Appendix A.

Attribute type	Categorical	Used for identification of stations	Many missing values	Without physical meaning	Due to collinearity
Topographic		gauge_id huc_02 gauge_name			area_geospa_fabric
Climatic	high_prec_timing low_prec_timing				aridity low_prec_freq
Land cover	dom_land_cover		root_depth_XX	dom_land_cover_frac	lai_diff gvf_max
Soil				other_frac	
Geology	geol_class_1st		geol_class_2nd	geol_class_1st_frac geol_class_2nd_frac	
GEV attributes					scale_par_prcp

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

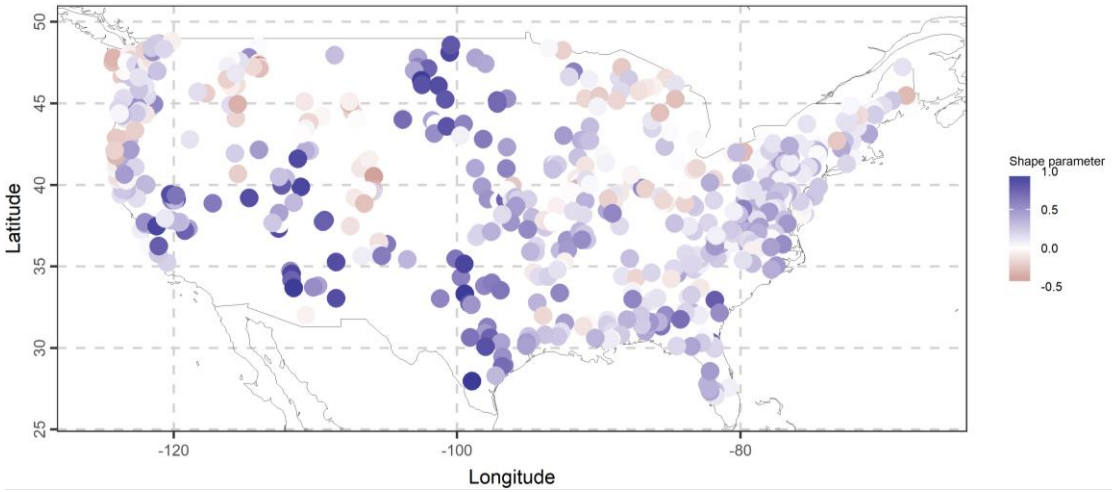

Figure 1. The 591 basins examined and their estimated streamflow GEV shape parameter (or *k* estimates).

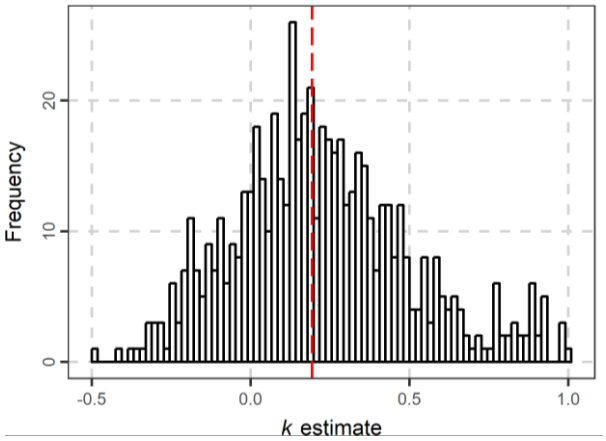

Figure 2. Histogram of *k* estimates of the 591 basins in Figure 1. The median (red dashed line) is equal to 0.19.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

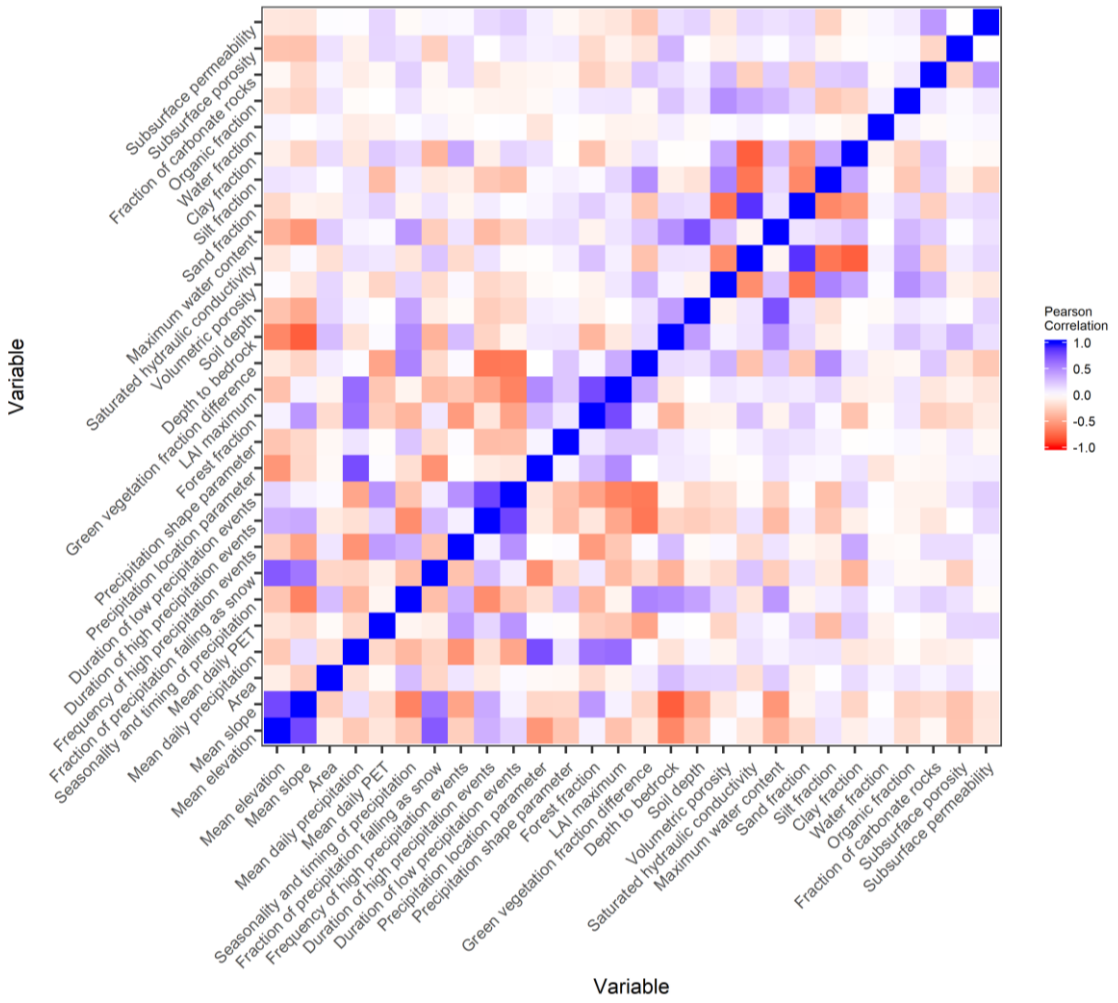

Figure 3. Correlations between the predictor variables included in the analysis (see also Table 1).

2.2 Methods

We used two algorithms (i.e. linear models and random forests) to predict the *k* parameter as a response to the attributes of the basin. All computations were performed using the R Programming Language (R Core Team 2018) and the contributed packages mentioned in Appendix C. Several other utilities accompanying linear models and random forests were implemented in the present study. These are presented in the following.

2.2.1 Linear models

The theory of linear models is well known (see, for instance, the textbook of Casella and Berger, 2002, pp. 577–611). The discussion here will be limited to special topics of interest. We use multiple linear regression models. The model is fitted to the data by implementing the `lm` R function. The complexity of the model increases with the increase of the number of predictor variables. However, the improvement in its predictive

importance for a number of predictor variables higher than a specific value depending on the specific case is not substantial. On the other hand, important predictor variables should not be removed. To this end stepwise backward regression, in which each fitted model is evaluated with the Akaike information criterion (AIC; Akaike 1974), can be used and unimportant predictor variables can be removed according to an automatic procedure. Stepwise backward regression is performed by implementing the `olsrr` R package (Hebbali 2018).

Linear models including interactions are also examined. In common statistical literature, the term interaction denotes the influence of the product of two or more predictor variables to the response. This approach differentiates from the usual approach in which the effects of the predictor variables are additive. The concept of interaction differs from the concept of confounding, which in Gaussian-based settings is equivalent with correlation (Boulesteix et al. 2015). Interactions between two predictor variables $\underline{x}_1$ and $\underline{x}_2$ are notated with $\underline{x}_1 : \underline{x}_2$. The notation $\underline{x}_1 \times \underline{x}_2 := \underline{x}_1 : \underline{x}_2 + \underline{x}_1 + \underline{x}_2$ is used to denote that additive terms are included in the interaction term.

Ranking the relative importance of the predictor variables in the linear model is crucial for understanding how the predictor variables affect the dependent variable (Grömping 2007a). The LMG relative importance metric (abbreviation of Lindeman, Merenda, and Gold 1980 who introduced the metric) is here used through the `relaimpo` R package (Grömping 2007b, 2018). The LMG metric decomposes the r^2 values of the fitted model into contributions from different predictor variables (see Grömping 2007a). While there are many methods to decompose r^2, the LMG metric is amongst the most credible ones (Grömping 2007a, b). For instance, LMG is invariant to the ordering of the predictor variables in the linear model, unlike the most frequently used Analysis of Variance (ANOVA).

It is essential to test the normality of the residuals of fitted linear models. To do so, we used the Shapiro-Wilk test (Shapiro and Wilk 1965). Selection of a linear model between many candidates is possible by using information criteria. Here we implement the Akaike Information Criterion (AIC; Akaike 1974) and the Bayesian Information Criterion (BIC; Schwarz 1978). When two linear models are fitted to a specific dataset, each one including different predictor variables, then the model with lower values of AIC and BIC is preferable.

Despite omitting highly correlated predictor variables in the examined dataset (see Section 2.1), the remaining variables still have some residual correlation. A suitable metric to examine the influence of the correlated variables in the linear model (also termed collinearity) is the Variance Inflation Factor (VIF, O'Brien 2007). Let v_i^2 represent the proportion of variance of the i^{th} predictor variable, which is associated with the other predictor variables in the model. The VIF metric is defined by $1/(1 - v_i^2)$ and intuitively is interpreted as the effect of v_i^2 on the variance of the estimated regression coefficient of the i^{th} predictor variable (O'Brien 2007). As a rule of thumb, common unacceptable values of the VIF metric are those that are higher than 10, albeit in some studies the limit reduces to 4. However, these rules should not be strictly applied (see the discussion in O'Brien 2007) and models including predictor variables with VIF higher than 10 can become acceptable.

2.2.2 Random forests

Random forests are a machine learning algorithm with a few parameters to optimize, while they are simple with high predictive accuracy and successful implementation in practical problems and forecasting competitions (Scornet et al. 2015; Biau and Scornet 2016). A detailed presentation of random forests and related concepts and terminology oriented to the purpose of our study is available in Appendix B. Random forests are used here for regression by implementing the `randomForest` R package (Liaw and Wiener 2002; Breiman et al. 2018). The algorithm has four hyperparameters (see also Appendix B). When increasing the number of trees hyperparameter, predictions become more accurate at the cost of increasing the computational time (Oshiro et al. 2012). The number of trees is set equal to 1 000 in the present study, since the gain in the predictive performance of the algorithm would be small by adding more trees (e.g. Probst and Boulesteix 2018). The other hyperparameters were also not optimized, because their predictive performance using their default values is similar to the predictive performance of the optimized algorithms, while the gain in computational time is high when optimization is not performed (see e.g. Biau and Scornet 2016).

Similarly to the linear model, random forests can be used for ranking the importance of variables in predicting the dependent variable (Verikas et al. 2011) with the aim to select important variables (Genuer et al. 2010). Rankings of variable importance using random forests and linear models exhibit some dissimilarities (Grömping 2009). For this

reason, the examination of both algorithms is useful. An important remark is that in contrast to the linear model, random forests are robust to the inclusion of many and non-important predictor variables (Díaz-Uriarte and De Andres 2006); thus, including all predictor variables would hardly affect the predictive performance of the model.

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, was used as relative variable importance metric. It was computed by implementing the `randomForest` R package. Relevant details are presented in the documentation of Breiman et al. (2018) (see also Appendix B).

2.2.3 Naïve prediction

The predictive performance of the regression models is compared to the naïve approach. In the latter, the predicted value of the k parameter is equal to its median value from the training sample in the 10-fold cross validation (see Section 2.2.4). Naïve prediction is used as worst-case benchmark.

2.2.4 10-fold cross validation

To test the predictive performance of the regression models (naïve, linear or random forests) 10-fold cross validation is performed (Kuhn and Johnson 2013, pp. 69–71). In particular, the sample is randomly divided into ten equal sized subsamples. The model is trained in nine subsamples and tested in the remaining one, while the procedure is repeated ten times. The Root Mean Square Error (RMSE), the Pearson's r and the slope of the regression line between the predicted and testing values are the metrics used for the assessment of the predictive performance of the regression models.

To test whether the differences between the mean RMSE values (each computed using the 10 RMSE values obtained via 10-fold cross validation) are statistically significant between a pair of methods, we implement the Wilcoxon signed-rank test (WSRT; Wilcoxon 1945). WSRT is a non-parametric statistical hypothesis test used to assess whether the population mean ranks between two samples differ. Its use for comparing the performance of machine learning algorithms in k-fold cross validation is suggested by Demšar (2006). A low p-value of the test (e.g. 0.05) denotes that the means of the two samples are different at a significance level 0.05.

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

2.2.5 The proposed framework

The proposed framework consists of the following sequential steps.

Step 1: Application of stepwise backward linear regression to the whole dataset. Let n_1 be the number of the retained predictor variables.

Step 2: Computation of LMG importance metrics for the retained predictor variables of Step 1.

Step 3. Computation of relative importance metrics for random forests when using all predictor variables.

Step 4. 10-fold cross validation with random forests using (a) all predictor variables (b) all predictor variables excluding geographical coordinates (c) the groups of predictor variables defined in Table 1.

Step 5. Again 10-fold cross validation with random forests. In this new model, training the most important variable according to the variable importance metric for random forests is included. Subsequently, the cross validation is repeated by adding one predictor variable at the time according to their importance. The procedure terminates when the performance of the last trained model is similar to the performance of the model that uses all predictor variables.

Step 6. The same procedure (procedure of Step 5) is repeated with the linear model, but here it terminates when using all predictor variables of Step 1. The LMG variable importance metric is used for ranking the variable importance and selecting the additional predictor variable in each iteration. Furthermore, AIC and BIC are computed for each fitted model.

Step 7. 10-fold cross validation is also performed for the naïve method.

Step 8. From the results of the steps 4–7 we understand (a) the performance of the models, (b) the importance of variables using two available metrics and (c) how the inclusion of more predictor variables increases (or decreases) the predictive performance of the models.

Step 9. Since the performance of the best linear model is expected to be worse compared to the best random forest model (the one that terminates the procedure in step 5) then we seek for interactions. The inclusion of these interactions could potentially increase the performance of the linear model. Here the procedure is semi-automatic. The main idea is

to first examine interactions between climatic attributes, since (as will be shown in the following Sections, when following steps 1–3) they are found to be the most important in the predictive model.

Step 10. Finally (and hopefully), a parsimonious (i.e. with few predictor variables) linear model with interacting terms and high predictive performance (slightly worse than the best random forest model) appears. According to the criteria set, other linear models may slightly outperform the proposed model; however, they are too complicated, because they include many predictor variables. The selected linear model is investigated by computing VIF and *p*-values.

2.2.6 Some remarks on the proposed framework

Some remarks on the steps of the procedure of the proposed framework are presented here:

Steps 1–3: Regarding the selection of important variables, we mention that, when many predictor variables (i.e. more than approximately 20 in our dataset) are included in the model fitting, the LMG cannot be computed in a regular home PC. Random forests are used in predictive modelling, in which variable selection is required through e.g. recursive procedures. In this case, these procedures are informative and can accompany other predictive models (Boulesteix et al. 2012). Variable importance, when including all predictor variables in the regression model, can be computed in a reasonable time if random forests are implemented. Consequently, variable importance computation with random forests is convenient when compared, e.g. with linear models, due to computational speed advantages. Therefore, a proposed strategy by Ziegler and König (2014) is to select important predictor variables using random forests in the beginning and, subsequently, use more computationally intensive methods (e.g. related to linear models) in the following. Here we decided to remove predictor variables in the linear model and then compare the results with the random forests. In our opinion, this strategy is equally reasonable.

Variable importance metrics rank the predictor variables, but the values of the metrics do not provide full knowledge about how significant the predictor variables are (Boulesteix et al. 2012). A conservative rule of thumb for selecting predictor variables based on importance metrics for random forests is presented by Strobl et al. (2009). Variables with negative, zero or small positive value of importance can be excluded. This

decision is based on the assumption that the importance of non-important variables is randomly distributed around zero.

Steps 4, 5: Regarding the selection of random forests as best case benchmark predictive model we mention that random forests are fast, flexible, robust, they can cope with high-dimensional data (i.e. few observations but many predictor variables), highly correlated variables, interactions between predictor variables, non-linear relationships between the response and the predictor variables and are non parametric, i.e. the specification of a statistical model is not required (Boulesteix et al. 2012; Ziegler and König 2014). Correlated variables have a very slight influence in the predictive performance of random forests (Boulesteix et al. 2012). They were found to outperform other methods, as well as hydrological models in hydrological signatures predictions (Zhang et al. 2018).

Variable importance metrics can be affected by strongly correlated variables; therefore, in some cases a few representative predictor variables should be selected. However, excluding all correlated variables is also not recommended, since information is lost. In this case, there should be some compromise between all options (Boulesteix et al. 2012). Removal of confounding can be done by adding the effect of the confounder separately in e.g. a multiple regression model. In this case, if for instance the effects of both confounders are positive, then the coefficients of the predictor variables are expected to be smaller compared to the case in which one of them is present (Boulesteix et al. 2015).

Step 10: We mention that the selection of a useful model is not only a matter of objectivity. As mentioned by Gelman and Hening (2017) "*practitioners must apply their subjective judgement in the choice of what method to use, what assumptions to invoke and what data to include in their analyses*". For instance, the choice of a linear model with a significant lower number of predictor variables can be justified over a linear model with a high number of predictor variables, when the AIC value of the latter is slightly lower. A discussion on the subjectivity and objectivity in statistical modelling, and how these concepts can be substituted by concepts such as transparency, consensus, impartiality and correspondence to observable reality, awareness of multiple perspectives, context dependence and stability can be found in Gelman and Hennig (2017).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

3. Application

3.1 Application of linear model

We applied a linear model to better understand the effect of the predictor variables of Table 1. In this application, we excluded the geographical coordinates, because this would not have a physical meaning, unless spatial models such as kriging were used in the modelling procedure. When including all variables of interest of Table 1 the computation of the LMG metric was not possible due to the high computational cost (see Section 2.2.6). Therefore, by applying the stepwise backward regression we excluded some variables not important for the prediction of *k*. The remaining variables, as well as their respective LMG metric values, are presented in Figure 4.

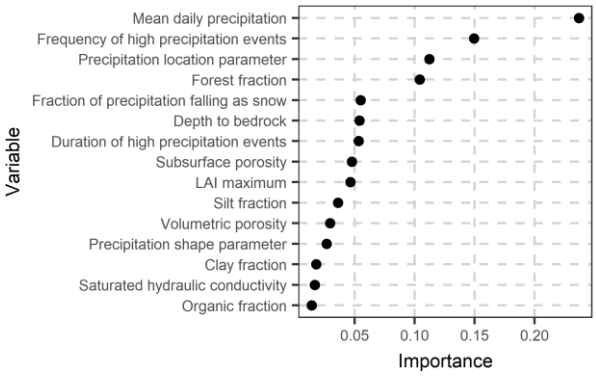

Figure 4. LMG relative importance metric for the predictor variables presented in the y-axis when a linear model is used to predict the shape parameter.

3.2 Application of random forests

We applied random forests to predict the *k* parameter. The predictor variables are presented in Table 1. Neighbouring basins share similar attributes, while this information is included in their geographical coordinates. Consequently, inclusion of geographical coordinates may mask the influence of other attributes in the prediction of *k*. Hence, two cases were examined, i.e. in the first case the geographical coordinates (longitude and latitude of the basin) were omitted from the set of the predictor variables, while in the second case they were included in the set. The importance of the predictor variables in predicting *k* is presented in Figure 5 for both cases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

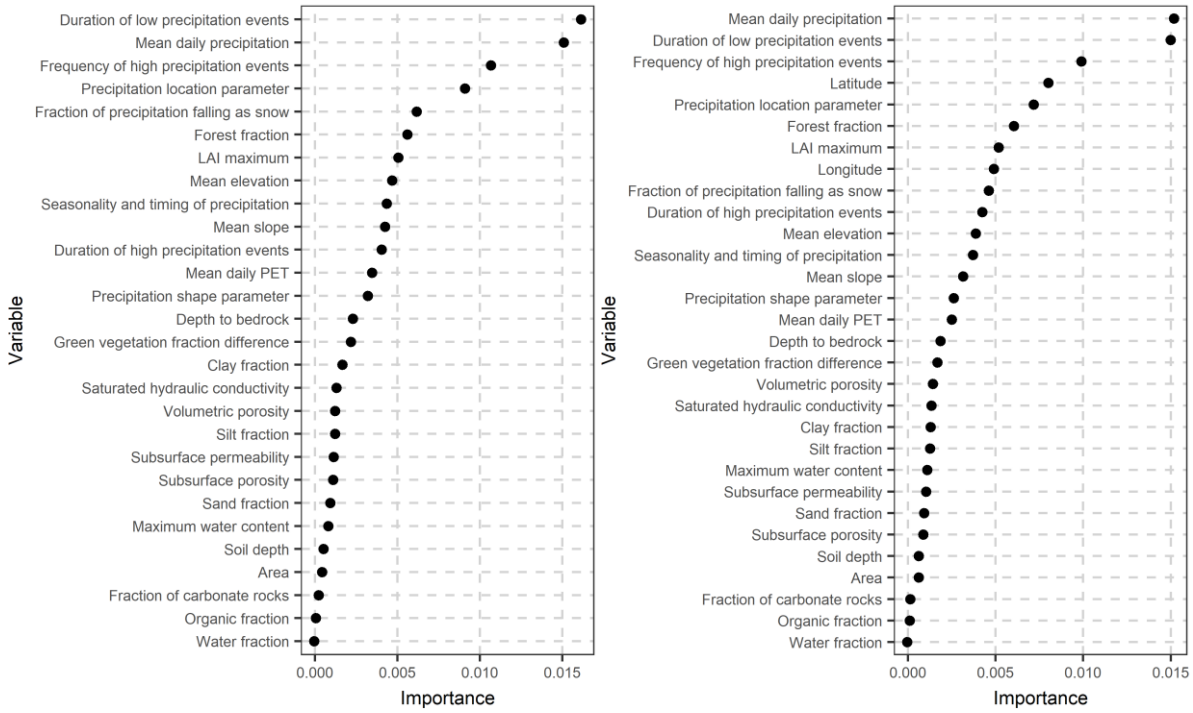

Figure 5. Variable importance of explanatory variables of interest in Table 1 (the geographical coordinates are excluded in the left and included in the right) when random forests are applied to the dataset of the 591 stations to predict the shape parameter. The variable importance of a particular variable is the percentage of increase in mean square error observed in out-of-bag (OOB) prediction when this variable is randomly permuted (Breiman et al. 2018, see also Appendix B).

The ranking of the variables with respect to their importance is slightly different in the two cases. When excluding the geographical coordinates, the most important variables are the mean daily precipitation and the duration of low precipitation events. They are followed by the frequency of high precipitation events and the precipitation GEV location parameter. The fraction of precipitation falling as snow and the forest fraction are also important variables. We note here again that variable importance metrics rank the predictor variables, but the values of the metrics are less informative (see Section 2.2.6). Therefore, they should be combined with the predictive performance of the models, to understand their absolute contribution to the *k* parameter. This examination follows in Sections 3.3 and 3.4. Here we mention that the increase in the performance of the random forest based predictive models flattens after including 7 to 8 predictor variables (see again Section 2.2.6), and this is the criterion used here to characterize a predictor variable as important.

When including the geographical coordinates, the latitude and, to a lesser extent, the longitude of the basins are important variables as well, albeit not as important as most of

the earlier mentioned ones. The maximum monthly mean of the leaf area index (LAI maximum) also becomes an important variable.

3.3 General results

In both models, the most important variables are climatic indices (the GEV parameters of precipitation can also be considered as climatic indices). Important variables of other types are the forest fraction, the LAI maximum, the catchment mean elevation, the catchment mean slope and the depth to bedrock depending on the employed model. The duration of low precipitation events was excluded when applying the stepwise backward regression, albeit it is an important variable in the random forest model.

To understand how k and important predictor variables are related we provide some representative scatterplots in Figure 6. We selected predictor variables based primarily on the computations presented in Figure 5 and secondly on the computations presented in Figure 4. It appears that there is a significant linear relationship between k and some variables (p-values lower than 0.05). Furthermore, k is rather dispersed around the regressions lines (see relative low r values).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

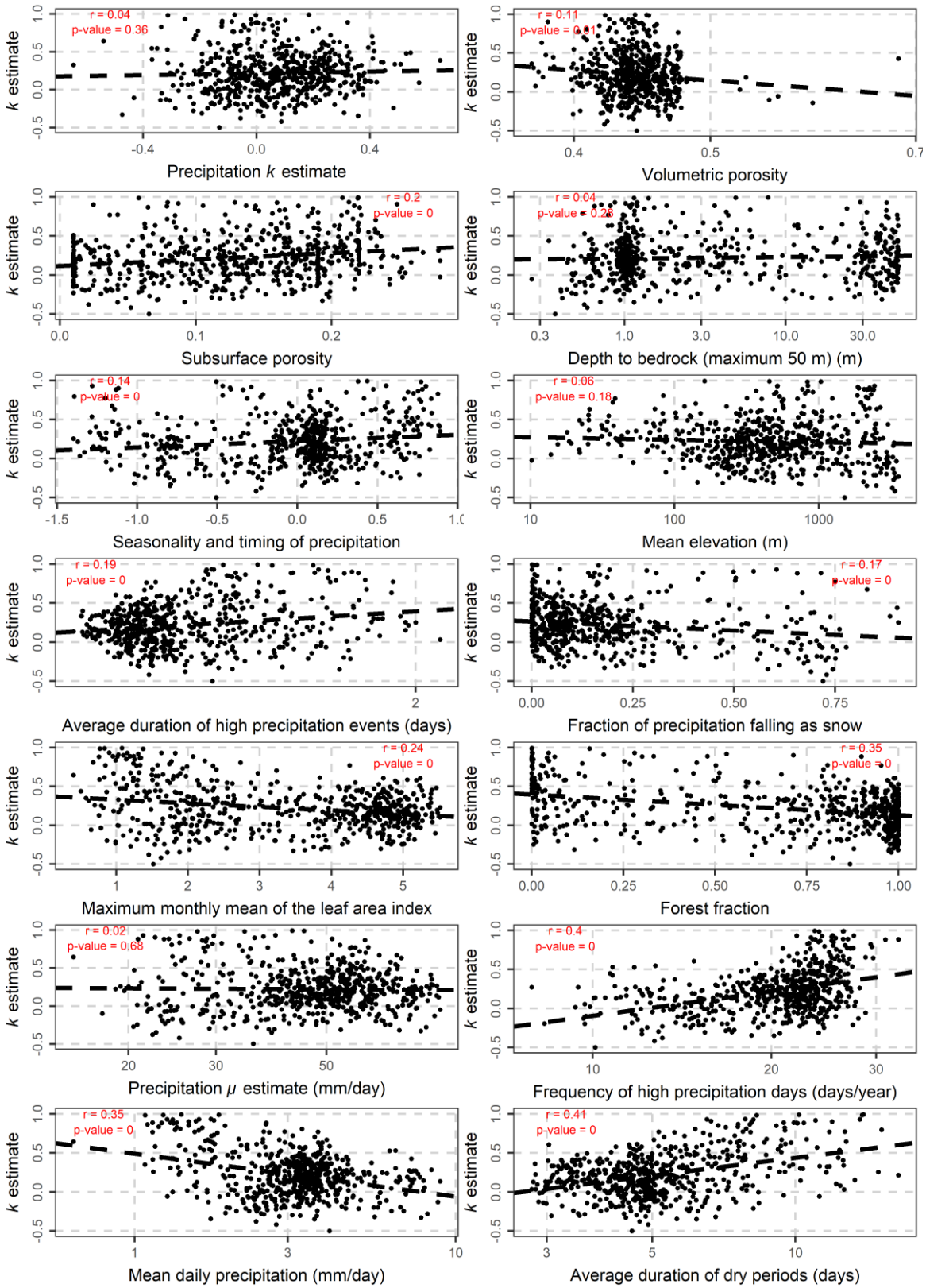

Figure 6. Scatterplots of the shape parameter and predictor variables of interest. The line is obtained by the linear regression of the shape parameter with the predictor variable. The *p*-values and Pearson's *r* of the linear model are also depicted.

To understand how the predictor variables improve the predictive performance of the fitted models we fit a sequence of models presented in Table 3 using random forests. All

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

models are evaluated using 10-fold cross validation. The models are trained on the 90% of the data and predict *k* for the remaining 10% of the data. The procedure is repeated 10 times, while the respective metrics are equal to the mean of their 10 values obtained from the 10-fold cross-validation. Optimal values of the RMSE should be near 0, of Pearson's *r* should be near 1 and the slope should be near 1. When the slope is equal to 1, the regression line between the predicted and the real values of *k* makes an angle of 45° with the x-axis.

The rf1–rf11 models include important predictor variables based on Figure 5. The first model includes the most important predictor variable, while an important predictor variable based on the ranking of Figure 5 is added in the model at each step. The rf12 model includes the predictor variables of the rf11model and the geographical coordinates. The topographic, climate, land, soil, geology and geographical coordinates random-forest-based models include the respective variables defined in Table 1. The reason is that we aim to understand how each particular type of attributes influences the *k* parameter. Two additional models, which are based on random forests and include all predictor variables of Figure 5, are examined with the aim to estimate the best prediction of *k* using the available data. The results of the naïve model are also presented in Table 3.

Table 3. Mean model errors (see Section 2.2.4) on the test set of the 10-fold cross-validation for predicting the shape parameter for each method and metric using random forests. The naïve method is also presented.

Name	Predictor Variables	RMSE	*r*	slope
rf1	low_prec_dur	0.291	0.284	0.196
rf2	low_prec_dur, p_mean	0.242	0.518	0.338
rf3	low_prec_dur, p_mean, high_prec_freq	0.233	0.552	0.347
rf4	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp	0.227	0.577	0.363
rf5	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow	0.223	0.599	0.368
rf6	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow, forest_frac	0.221	0.607	0.383
rf7	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow, forest_frac, lai_max	0.219	0.620	0.384
rf8	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow, forest_frac, lai_max, elev_mean,	0.217	0.628	0.391
rf9	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow, forest_frac, lai_max, elev_mean, p_seasonality	0.215	0.634	0.404
rf10	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow, forest_frac, lai_max, elev_mean, p_seasonality, slope_mean	0.215	0.632	0.401
rf11	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow, forest_frac, lai_max, elev_mean, p_seasonality, slope_mean, high_prec_dur	0.216	0.632	0.397
rf12	low_prec_dur, p_mean, high_prec_freq, loc_par_prcp, frac_snow, forest_frac, lai_max, elev_mean, p_seasonality, slope_mean, high_prec_dur, gauge_lat, gauge_lon	0.214	0.641	0.409
Topographic	Topographic attributes, see Table 1	0.262	0.355	0.161
Climate	Climatic attributes, see Table 1	0.218	0.621	0.390
Land	Land cover attributes, see Table 1	0.243	0.500	0.252
Soil	Soil attributes, see Table 1	0.260	0.362	0.157
Geology	Geology attributes, see Table 1	0.276	0.241	0.091
Geographical coordinates	gauge_lat, gauge_lon	0.224	0.615	0.396
All attributes 1	All attributes of Table 1, excluding geographical coordinates	0.214	0.641	0.387
All attributes 2	All attributes of Table 1	0.213	0.642	0.393
naïve		0.281	–	0.000

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

The sequence of fitted linear models is presented in Table 4. The lm1–lm16 models include predictor variables according to their ranking of Figure 4. The AIC and BIC values of the linear models when fitted to the whole dataset are also presented.

Table 4. Mean model errors (see Section 2.2.4) on the test set of the 10-fold cross-validation for predicting the shape parameter for linear models. AIC and BIC are computed when the linear model is fitted to the whole dataset.

Name	Predictor Variables	RMSE	*r*	slope	AIC	BIC
lm1	p_mean	0.263	0.340	0.118	101.13	114.27
lm2	p_mean, high_prec_freq	0.254	0.419	0.179	59.10	76.63
lm3	p_mean, high_prec_freq, loc_par_prcp	0.238	0.518	0.276	–17.64	4.27
lm4	p_mean, high_prec_freq, loc_par_prcp, forest_frac	0.238	0.518	0.276	–16.10	10.19
lm5	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow	0.238	0.516	0.276	–14.32	16.35
lm6	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier	0.233	0.544	0.306	–37.31	–2.26
lm7	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur	0.232	0.551	0.315	–43.62	–4.18
lm8	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity	0.230	0.564	0.330	–55.65	–11.83
lm9	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity, lai_max	0.230	0.563	0.330	–55.72	–7.52
lm10	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity, lai_max, silt_frac	0.228	0.572	0.342	–63.74	–11.16
lm11	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity, lai_max, silt_frac, soil_porosity	0.227	0.578	0.351	–70.92	–13.95
lm12	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity, lai_max, silt_frac, soil_porosity, shape_par_prcp	0.223	0.598	0.374	–91.02	–29.68
lm13	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity, lai_max, silt_frac, soil_porosity, shape_par_prcp, clay_frac	0.223	0.598	0.375	–90.20	–24.48
lm14	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity, lai_max, silt_frac, soil_porosity, shape_par_prcp, clay_frac, soil_conductivity	0.223	0.599	0.377	–90.49	–20.38
lm15	p_mean, high_prec_freq, loc_par_prcp, forest_frac, frac_snow, soil_depth_pelletier, high_prec_dur, geol_porosity, lai_max, silt_frac, soil_porosity, shape_par_prcp, clay_frac, soil_conductivity, organic_frac, as in Figure 4	0.222	0.601	0.382	–96.26	–21.77

Finally, we fit linear models that include interactions, as presented in Table 5. Practically, we firstly tested all interactions between the mean daily precipitation, the frequency of high precipitation days and the precipitation location parameter. We found that newlm4 combination of predictor variables includes 2 terms and performs similarly or better compared to the newlm1–newlm7 combinations, while it includes less predictor variables. The procedure continued by adding (and then removing if found useless) in a stepwise mode the most important variables found in Figure 4. The newlm17 model includes 6 terms and reduces the RMSE compared to the previous best fitted model (newlm16) by 0.06. The next models (newlm21, newlm22, newlm23) further decrease the RMSE by 0.03, but they include at least 5 more terms (see newlm21, which includes 3-way interactions and their additive effects). Finally, the newlm24–27 models present a significant increase in the RMSE when some terms from the newlm17 model are omitted.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

As representative model we select the newlm17 one (further reasoning along with other details can be found later in Section 4.3).

Table 5. Mean model errors (see Section 2.2.4) on the test set of the 10-fold cross-validation for predicting the shape parameter for linear models with interactions. AIC and BIC are computed when the linear model is fitted to the whole dataset. Here $a : b$ denotes interaction while $a \times b := a + b + a : b$ (includes interactions and additive effects, see Section 2.2.1).

Name	Predictor Variables	RMSE	r	slope	AIC	BIC
newlm1	p_mean × high_prec_freq	0.240	0.504	0.261	−5.47	16.44
newlm2	p_mean : high_prec_freq	0.267	0.297	0.089	120.12	133.27
newlm3	p_mean × high_prec_freq, loc_par_prcp	0.231	0.553	0.317	−53.44	−27.15
newlm4	p_mean : high_prec_freq, loc_par_prcp	0.233	0.542	0.303	−41.82	−24.29
newlm5	p_mean + loc_par_prcp	0.239	0.514	0.268	−11.39	6.13
newlm6	high_prec_freq + loc_par_prcp	0.257	0.395	0.157	75.13	92.66
newlm7	p_mean × high_prec_freq × loc_par_prcp	0.232	0.552	0.322	−52.63	−13.19
newlm8	p_mean : high_prec_freq, loc_par_prcp, frac_forest	0.233	0.541	0.303	−39.88	−17.97
newlm9	p_mean : high_prec_freq, loc_par_prcp, frac_snow	0.233	0.542	0.304	−40.70	−18.79
newlm10	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier	0.231	0.554	0.318	−52.70	−30.79
newlm11	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, high_prec_dur	0.231	0.553	0.318	−51.00	−24.70
newlm12	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity	0.228	0.570	0.337	−67.15	−40.86
newlm13	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, lai_max	0.228	0.569	0.336	−65.24	−34.57
newlm14	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, silt_frac	0.227	0.576	0.344	−72.20	−41.52
newlm15	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, silt_frac, soil_porosity	0.226	0.581	0.352	−77.74	−42.68
newlm16	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity	0.225	0.583	0.352	−79.27	−48.60
newlm17	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp	0.219	0.612	0.387	−110.54	−75.48
newlm18	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp, clay_frac	0.219	0.611	0.387	−108.58	−69.14
newlm19	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp, soil_conductivity	0.219	0.612	0.388	−110.17	−70.73
newlm20	p_mean : high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp, organic_frac	0.219	0.614	0.390	−112.31	−72.87
newlm21	p_mean × high_prec_freq × shape_par_prcp, loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity	0.216	0.628	0.412	−129.91	−72.95
newlm22	p_mean × high_prec_freq × shape_par_prcp, loc_par_prcp, soil_depth_pelletier × geol_porosity × soil_porosity	0.217	0.622	0.410	−122.76	−48.27
newlm23	p_mean × high_prec_freq × shape_par_prcp, loc_par_prcp, soil_depth_pelletier × geol_porosity, soil_porosity	0.216	0.627	0.412	−128.47	−67.13
newlm24	p_mean, high_prec_freq, loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp	0.225	0.586	0.356	−79.39	−39.95
newlm25	p_mean : high_prec_freq, soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp	0.260	0.372	0.143	90.92	121.60
newlm26	loc_par_prcp, soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp	0.274	0.215	0.048	153.28	183.95
newlm27	soil_depth_pelletier, geol_porosity, soil_porosity, shape_par_prcp	0.274	0.218	0.048	151.96	178.25

To understand how differences in the RMSE, which can be perceived as small, can largely influence predictive uncertainties, we compute prediction intervals for the 500-year floods for all basins. The T-year flood is defined by (Dey et al. 2016; Tyralis and Langousis 2018):

$$q_T = \mu + (\sigma/k)\,\left(\,(-\log(1-1/T))^{-k} - 1\right) \quad (2)$$

Two models are compared, i.e. the selected newlm17 and the newlm4 models. The newlm4 model was selected for comparison reasons because compared to the newlm17 model does not include geologic and soil attributes (see also the relevant discussion in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0265.v1

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

Section 4.3). The difference in the mean RMSE of the 10-fold cross validation between the two models is 0.014. The *p*-value of the WSRT is equal to 0.075, i.e. it is lower than the significance level 0.10, indicating that the difference is significant. Since the focus here is to isolate the influence of the k parameter, the μ and σ parameters are set equal to their known values (i.e. the maximum likelihood estimates). Then 10-fold cross validation is implemented for both linear models and 95% prediction intervals for k are computed at the independent sets using the `lm` R function. Since the quantile is an increasing function of k, as can be derived by eq. (2), 95% prediction intervals can be obtained for q_{500} by simply substituting k in eq. (2) with its prediction limits. Coverage probabilities for newlm17 and newlm4 are equal to 0.949 and 0.956 respectively. However, the prediction intervals of the newlm17 model are considerably narrower compared to those produced by newlm4. In particular, we computed the relative decrease in the width of the prediction interval between the two models in each basin (in the 10 independent test sets of the 10-fold cross validation) according to:

$$a = (\text{width}_{\text{newlm4}} - \text{width}_{\text{newlm17}})/\ \text{width}_{\text{newlm4}} \quad (3)$$

The mean relative decrease in the sample of all basins is 4.99%, while the histogram of the relative improvements per basin can be found in Figure 7. To understand the difference between the two models, it is mentioned that the mean width of the prediction intervals are 4 785 m^3/s and 5 460 m^3/s for the newlm17 and newlm4 models respectively, while 500-year floods range up to 20 000 m^3/s.

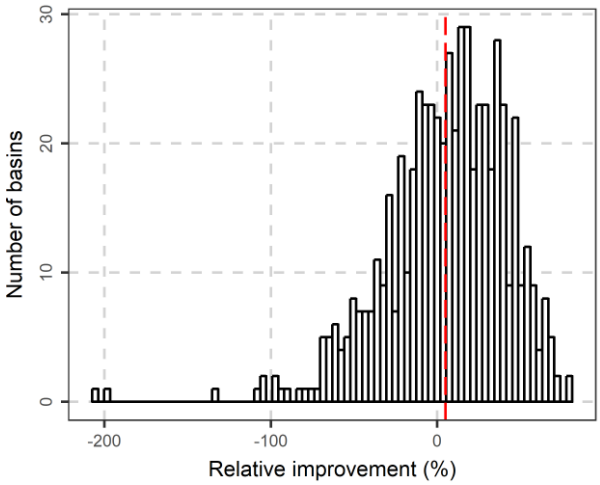

Figure 7. Histogram of relative improvement of the 95% prediction interval width of the model newlm17 against the model newlm4. The mean (red dashed line) is equal to 4.99%.

Peer-reviewed version available at Journal of Hydrology 2019, 574, 628-645; doi:10.1016/j.jhydrol.2019.04.070

3.4 Overview of results

3.4.1 Naïve method

We limit our following discussion mainly to the assessment of the predictive performance with respect to the value of the RMSE reported in Section 3.3. The naïve method serves as a benchmark; therefore, all combinations of predictor variables and all applied models should be assessed based on their relative performance compared to it. The RMSE in the estimation of k when applying the naïve method is equal to 0.281. It could be said informally that the naïve method is equivalent to not using any predictor variables for k in the models introduced, for example, by Northrop (2004).

3.4.2 Linear models

When applying the linear model by adding one variable at a time the increase in performance is small. However, the use of 12 variables (lm12) leads to a performance that is approximately equal to the one of the rf model with the geographical coordinates, i.e. the RMSE is 0.223 (21% increase in performance compared to the naïve method). The use of all predictor variables presented in Figure 4 results in a RMSE of 0.222, while when including 3–5 predictor variables (lm3–lm5) the RMSE becomes 0.238. The increase in performance is 15%. Finally, the inclusion of six predictor variables results in a 17% increase in performance. These values are important benchmarks for understanding the importance of the predictor variables when added in stepwise mode.

3.4.3 Random forests

The predictive performance of the random forest based models increases with the increase of the number of predictor variables, but it seems to stabilize when using nine (rf9, see Table 3 for the abbreviation) or more predictor variables. When moving from the use of one predictor variable (rf1) to nine predictor variables (rf9), the RMSE in the prediction of k decreases from 0.291 to 0.215. Its optimal value is equal to 0.213 and it is observed for the use of all predictor variables of Figure 5, including the geographical coordinates of the catchments. When excluding the geographical coordinates the RMSE increases to 0.214. This change can be considered negligible. When using the geographical coordinates of the catchments the RMSE is 0.224. This value can be surpassed by using five predictor variables (rf5). An explanation is that the information gained through utilising the proximity of the catchments can be compensated by the information obtained

through the predictor variables of rf5. This is important, since the information from the CAMELS dataset can be transferred to geographical coordinates not included in the CONUS, using just five predictor variables.

It is further important to understand how the different types of attributes of the catchments can be used to increase the information for *k*. The climatic indices result in an RMSE equal to 0.218, which is significantly low compared to the naïve approach, while its difference compared to the optimal model is also very small. The other types of predictor variables (see Table 1) do not seem to be particularly useful for the prediction model. The land cover characteristics seem to improve the performance of the model with an RMSE equal to 0.243. They are followed by the soil and topographic characteristics with RMSE values equal to 0.260 and 0.262 respectively. The improvement using geological characteristics is negligible.

Compared to the naïve approach, the optimal model results in an increase in performance of the RMSE equal to 24%, which is a considerable improvement. The respective improvement in performance when applying the rf5 model is 21%. This improvement is fair as well, if we also consider the fact that it can be achieved by using only five predictor variables.

3.4.4 Summary of results

The mean Pearson's *r* in the 10-fold cross-validation is approximately equal to 0.60, while its exact value depends on the combination of predictor variables and the selected model. The patterns of change in performance are similar to the patterns observed for RMSE. Here again, we highlight that in this kind of studies the relative importance compared to naïve methods is of high importance, and therefore the approach should not be exclusively assessed based on criteria related to the absolute performance. A Pearson's *r* equal to 0.60 may not be close to 1, yet the improvement is considerable compared to its respective value when explanatory information is not used. The case for the slope of the regression line is also similar. The slope of the naïve method is 0, while it increases to 0.40 for rf12.

By comparing the random forests with the linear model we observe that the former have better predictive performance. This is due to the flexibility of the algorithm, which can reveal possible non-linear relationships. Therefore, random forests that use less predictor variables can have equal performance with linear models. It is also of

significance that the most important variables entered firstly in the models are climate indices.

4. Discussion

4.1 Some additional remarks on the experimental design

Shmueli (2010) identifies three modelling perspectives, i.e. predictive, explanatory and descriptive modelling. Breiman (2001) makes a distinction between two cultures in statistical modelling. In the first culture, it is assumed that the data are generated by a statistical model, while in the second culture that the data are modelled by a non-parametric model, since the data mechanism is considered unknown. According to Boulesteix and Schmid (2014) these two approaches are related, i.e. the statistical approach should be preferred when descriptive modelling is required, while non-parametric approaches (termed algorithmic approaches in Breiman 2001) are suitable in the second case. In some cases, it is possible that a statistical model can also perform equally well to an algorithmic model; therefore, it can simultaneously answer questions related to the description of the model and its predictability. If such a model can be found, as is the case here, then it can answer multiple questions.

Genuer et al. (2010) identify two variable selection objectives. These are the finding of important variables that are highly related to the dependent variable for interpretation purposes and the designing of a parsimonious prediction model by retaining a small and sufficient number of predictor variables. The two objectives are parallel to the distinction between explanatory/causal importance and predictive importance in typical regression models (Grömping 2009) and are related to the descriptive modelling perspective (Shmueli 2010). The combination of these two objectives can lead to a better understanding of the influence of the predictor variables on k (Grömping 2009), albeit a theory-driven explanatory model should be preferred, if it exists (Grömping 2007a). In the absence of such model important variables that result from metrics based on data-driven methods should be preferred to explain the nature of the response variable (Grömping 2007a). Here we employed a similar strategy to the ones proposed by Díaz-Uriarte and De Andres (2006) and Genuer et al. (2010) for variable selection by ranking the predictor variables according to their importance and by introducing variables in the prediction problem in a stepwise strategy. Ad-hoc interpretation of the importance of the predictor variables can then take place, while the selection of a parsimonious model

depends on the specific case examined. The latter involves comparison with naïve methods and intercomparison of models with varying number of predictor variables with respect to their predictive performance.

4.2 New findings on the nature of the *k* parameter

A general assessment is that the *k* parameter depends on climatic indices, while the other attributes of the catchments are less important. This result is in agreement with Beck et al. (2015) and Addor et al. (2018), who argue that hydrological signatures mostly depend on climatic indices, albeit Addor et al. (2018) claim that this may be a result of an insufficient summary of the catchments attributes by the implemented indices. What is particularly important is that hydrological signatures related to the magnitude of flow (e.g. the μ and σ parameters here) mostly depend on the area of the catchment (see also Northrop 2004), with a high influence in their values, while other attributes have less influence in the response variable. However, the *k* parameter has a different nature. Firstly, the uncertainties in its estimation are higher, resulting in the dispersed scatterplots observed in Figure 6. Secondly, assuming that the uncertainties are mitigated by the large sample, the influence of the area of the catchment is less profound compared to the cases of the μ and σ parameters. The large CAMELS dataset (Newman et al. 2015; Addor et al. 2017b) helped in finding such relationships for the *k* parameter (see e.g. Northrop 2004; Villarini and Smith 2010; Villarini et al. 2011a, b; Lima et al. 2016, Wallis et al. 2007; Ahn and Palmer 2016), which earlier studies could not identify due to limited data availability. In particular, Northrop (2004) and Lima et al. (2016) did not use a regression model for the *k* parameter, while Villarini and Smith (2010) and Villarini et al. (2011a, b) found relationships with the catchment area. The latter is here shown to be less important compared to at least 10 other predictor variables. Wallis et al. (2007) found a relationship between *k* and the mean annual precipitation, which was also found here, but it is not sufficient for a good prediction of *k*. On the other hand, Ahn and Palmer (2016) found that *k* depends on the latitude, the mean basin slope and the precipitation seasonality. These attributes were found less important here compared to other attributes. He et al. (2015) also did not find any relationship between the *k* parameter and the catchment area, and suggested that hydrological heterogeneity is implicitly reflected in the shape parameter. Apparently, results of different studies are not directly comparable, especially when data from different regions are used; however, the present

study includes a higher number of examined basins and attributes, while the basins represent a large diversity of climate types.

4.3 The final model

Random forests is an algorithm with high predictive performance and an ability to reveal interactions between the predictor variables and non-linear relationships (see Section 2.2.6). Therefore, the here lowest predictive performance of the linear model should be expected. Considering that the improvement of other algorithms is expected to be low compared to random forests, it is reasonable to assume that an optimal benchmark regarding the prediction of k would be a result from the implementation of random forests. Considering also the need for obtaining an interpretable and parsimonious model, a linear model with a small number of predictor variables should be selected in the model. Such a model is the newlm17, which includes five predictor variables and the interaction between other two attributes, when fitted to the sample of the 591 catchments as shown in the next equation:

k = – 2.61 + 0.87 log (precipitation location parameter) – 0.03 log(depth to bedrock) + 0.46 subsurface porosity – 0.66 log(volumetric porosity) + 0.30 precipitation GEV shape parameter – 0.32 log(mean daily precipitation) log(frequency of high precipitation events) (4)

It is obvious that the newlm17 model has good predictive properties, since it is better than all linear models in Table 4, highlighting the role of interactions. It is also slightly worse compared to the rf8–rf12 random forest models with respect to its predictive performance; however, it is more interpretable and includes less predictor variables. It is also notable that the rate of increase in the predictive performance of random forests decreases rapidly as more predictor variables are added in the models. Therefore, starting from an RMSE equal to 0.291 (rf1 model in Table 3), an intermediate RMSE equal to 0.219 is reached (rf7 in Table 3), while the terminating RMSE is equal to 0.214 (rf14 model in Table 3). A delivered RMSE equal to 0.219, together with a small number of predictor variables and a simple model structure, are good reasons to select the newlm17 model for the given data.

All coefficients of the model of eq. (4) where statistically significant at the 0.05-level. The VIF of precipitation location parameter and the interaction term were 4.41 and 4.52 respectively, which are far lower compared to 10 (see Section 2.2.1); therefore, they are

acceptable, especially if we consider that their exclusion results in significant decrease in performance. The VIF of the predictor variables were in the range 1−1.5 (1 is the lower limit of VIF). The residuals of the newlm17 model were also found normally distributed according to the Shapiro-Wilk test. The model of eq. (4) uses seven predictor variables and its RMSE was 0.219 in the 10-fold cross-validation, while the Pearson's r was equal to 0.612. Its adjusted r^2 was 0.39 when fitted to the dataset of the 591 catchments. Furthermore, its performance is equal to the one of the rf7 model, which also includes seven predictor variables.

When looking at eq. (4) one sees that k is a decreasing function of the product of mean daily precipitation and frequency of high precipitation events. The inclusion of the interaction played a crucial role in the considerable increase in performance compared to the models that do not include interactions, i.e. the models of Table 4. Additionally, k increases with the location parameter of the GEV distribution of precipitation extremes and with the increase of their shape parameter. The latter seems also sensible, because extreme precipitation should result in streamflow extremes. Lastly, k increases with increasing subsurface porosity and decreases with increasing depth to bedrock and volumetric porosity. To the authors' knowledge, there is not a theory-driven explanatory model for the relationship between k and geological or soil attributes. However, the benefits of using such model have been shown in Section 3.3 (see Figure 7 and the relevant discussion on the comparison in the predictive performance between newlm4 (which includes the interaction term and the location parameter of precipitation) and newlm17).

5. Conclusions

The shape parameter of the generalized extreme value distribution of daily annual block maxima of streamflow is important because it is related to how extreme the floods are. For this specific reason, it should be attentively examined with the aim to reduce its high impact on uncertainty, when incorporated in statistical models of extremes.

Here we propose a framework to find significant relationships between the shape parameter and basin attributes in the context of flood frequency analysis, as well as to predict the shape parameter given the attributes in ungauged or sparsely gauged basins. The framework is based on multiple linear regression, incorporation of interactions between the attributes, assessment of the importance of attributes in predicting the shape parameter within a linear framework and comparison with a high performance non-

linear model (random forests), which is herein used as best case prediction algorithm, aiming to validate the proposed linear model. We applied the framework to 591 basins in the contiguous US.

We found that the shape parameter is influenced by the interactions between the mean daily precipitation and the frequency of high precipitation days, the precipitation GEV location parameter and the precipitation GEV shape parameter. It also depends on geological and soil characteristics of the catchment, albeit to a smaller extent.

The RMSE of the linear model in a 10-fold cross-validation scheme was found to be 0.219, i.e. 22% smaller than the RMSE computed for a naïve model, while its adjusted r^2 when the model is fitted to the whole dataset is 0.39. Its performance was similar to the more complex benchmark model, i.e. negligible improvements can be found, by further modification of the model. The incorporation of this model into relevant Bayesian frameworks or regression-based models for regional flood frequency analysis may result in considerable reduction of the predictive uncertainties.

Conflicts of interest: The authors declare no conflict of interest.

Appendix A Description of catchment attributes

In Tables A-1-A-6 we describe the attributes of the basins.

Table A-1. Name, location and topographic characteristics (adapted from Addor et al. 2017b).

Attribute	Abbreviation	Description
Gauge id	gauge_id	catchment identifier (eight-digit USGS hydrologic unit code)
Region	huc_02	region (two-digit USGS hydrologic unit code)
Gauge name	gauge_name	gauge name, followed by the state
Latitude	gauge_lat	gauge latitude
Longitude	gauge_lon	gauge longitude
Mean elevation	elev_mean	catchment mean elevation
Mean slope	slope_mean	catchment mean slope
Area	area_gages2	catchment area (GAGESII estimate)
	area_geospa_fabric	catchment area (geospatial fabric estimate)

Table A-2. Climatic indices (adapted from Addor et al. 2017b).

Attribute	Abbreviation	Description
Mean daily precipitation	p_mean	mean daily precipitation
Mean daily PET	pet_mean	mean daily PET, estimated by N15 using Priestley–Taylor formulation calibrated for each catchment
Aridity	aridity	aridity (PET / P, ratio of mean PET, estimated by N15 using Priestley–Taylor formulation calibrated for each catchment, to mean precipitation)
Seasonality and timing of precipitation	p_seasonality	seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and precipitation cycles; positive (negative) values indicate that precipitation peaks in summer (winter); values close to 0 indicate uniform precipitation throughout the year)
Fraction of precipitation falling as snow	frac_snow	fraction of precipitation falling as snow (i.e., on days colder than 0°C)
Frequency of high precipitation events	high_prec_freq	frequency of high precipitation days (≥ 5 times mean daily precipitation)
Duration of high precipitation events	high_prec_dur	average duration of high precipitation events (number of consecutive days ≥ 5 times mean daily precipitation)
Season of high precipitation events	high_prec_timing	season during which most high precipitation days (≥ 5 times mean daily precipitation) occur
Frequency of low precipitation events	low_prec_freq	frequency of dry days (< 1 mm day^{-1})
Duration of low precipitation events	low_prec_dur	average duration of dry periods (number of consecutive days < 1 mm day^{-1})
Season of low precipitation events	low_prec_timing	season during which most dry days (< 1 mm day^{-1}) occur

Table A-3. Land cover characteristics (adapted from Addor et al. 2017b).

Attribute	Abbreviation	Description
Forest fraction	forest_frac	forest fraction
LAI maximum	lai_max	Maximum monthly mean of the leaf area index (based on 12 monthly means)
LAI difference	lai_diff	difference between the maximum and minimum monthly mean of the leaf area index (based on 12 monthly means)
Green vegetation fraction maximum	gvf_max	maximum monthly mean of the green vegetation fraction (based on 12 monthly means)
Green vegetation fraction difference	gvf_diff	difference between the maximum and minimum monthly mean of the green vegetation fraction (based on 12 monthly means)
Dominant land cover	dom_land_cover	dominant land cover (Noah-modified 20-category IGBP-MODIS land cover)
Dominant land cover fraction	dom_land_cover_frac	fraction of the catchment area associated with the dominant land cover
Root depth	root_depth_XX	root depth (percentiles XX = 50 and 99 % extracted from a root depth distribution based on IGBP land cover)

Table A-4. Soil characteristics (adapted from Addor et al. 2017b).

Attribute	Abbreviation	Description
Depth to bedrock	soil_depth_pelletier	depth to bedrock (maximum 50 m)
Soil depth	soil_depth_statsgo	soil depth (maximum 1.5 m; layers marked as water and bedrock were excluded)
Volumetric porosity	soil_porosity	volumetric porosity (saturated volumetric water content estimated using a multiple linear regression-based on sand and clay fraction for the layers marked as USDA soil texture class and a default value (0.9) for layers marked as organic material; layers marked as water, bedrock, and "other" were excluded)
Saturated hydraulic conductivity	soil_conductivity	saturated hydraulic conductivity (estimated using a multiple linear regression-based on sand and clay fraction for the layers marked as USDA soil texture class and a default value (36 cm h^{-1}) for layers marked as organic material; layers marked as water, bedrock, and "other" were excluded)
Maximum water content	max_water_content	maximum water content (combination of porosity and soil_depth_statsgo; layers marked as water, bedrock, and "other" were excluded)
Sand fraction	sand_frac	sand fraction (of the soil material smaller than 2 mm; layers marked as organic material, water, bedrock, and "other" were excluded)
Silt fraction	silt_frac	silt fraction (of the soil material smaller than 2 mm; layers marked as organic material, water, bedrock, and "other" were excluded)
Clay fraction	clay_frac	clay fraction (of the soil material smaller than 2 mm; layers marked as organic material, water, bedrock, and "other" were excluded)
Water fraction	water_frac	fraction of the top 1.5 m marked as water (class 14)
Organic fraction	organic_frac	fraction of soil_depth_statsgo marked as organic material (class 13)
Other fraction	other_frac	fraction of soil_depth_statsgo marked as "other" (class 16)

Table A-5. Geological characteristics (adapted from Addor et al. 2017b).

Attribute	Abbreviation	Description
Common geologic class	geol_class_1st	most common geologic class in the catchment
Fraction of common geologic class	geol_class_1st_frac	fraction of the catchment area associated with its most common geologic class
Second most common geologic class	geol_class_2nd	second most common geologic class in the catchment
Fraction of second most common geologic class	geol_class_2nd_frac	fraction of the catchment area associated with its second most common geologic class
Fraction of carbonate rocks	carb_rocks_frac	fraction of the catchment area characterized as "carbonate sedimentary rocks"
Subsurface porosity	geol_porosity	subsurface porosity
Subsurface permeability	geol_permeability	subsurface permeability (log10)

Table A-6. GEV attributes.

Attribute	Abbreviation	Description
Shape parameter of streamflow extremes	shape_par	GEV shape parameter estimate of the streamflow annual block maxima
Precipitation location parameter	loc_par_prcp	GEV location parameter estimate of the precipitation annual block maxima
Precipitation scale parameter	scale_par_prcp	GEV scale parameter estimate of the precipitation annual block maxima
Precipitation shape parameter	shape_par_prcp	GEV shape parameter estimate of the precipitation annual block maxima

Appendix B Random Forests

Here we present aspects of Random Forests (RF), an algorithm introduced by Breiman (2001). The presentation is based on the classical textbook of Hastie et al. (2018, Chapter

15). RF are a classification and regression algorithm. Here we use it for regression. The algorithm uses regression trees (see Hastie et al. 2018, Chapter 9) and a modification of bootstrap aggregating (bagging). Breiman's (2001) RF use the CART decision trees (see Hastie et al. 2018, Chapter 9), while other tree versions also exist. Trees have low bias and can model interactions. The idea of bagging is to average many noisy but approximately unbiased models aiming to reduce the variance. Consequently, a good option is to average many trees. The bias of the average of trees is equal to the bias of each tree; however, bagging reduces the variance of the average of trees. Further reduction of the variance is achieved when a modification of bagging is used. In this modification, each tree grows by a random selection of the input variables. The notation `mtry` is commonly used to denote the number of variables randomly selected at each tree due to the most frequently used software implementation of the algorithm, i.e. the `randomForest` (Liaw and Wiener 2002; Breiman et al. 2018) R package. The training of the algorithm is performed by minimizing the out-of-bag (oob) error, i.e. the error of the internal (within the training set) cross-validation of the algorithm.

The algorithm needs little tuning, while its performance is very good when using the default parameters, i.e. `mtry`, the number of trees, the maximum number of terminal nodes of the trees and the minimum size of terminal nodes. The number of trees is a critical parameter. Growing a large number of trees results in better predictions but the performance flattens asymptotically.

Estimation of the variable importance, i.e. the contribution of each input variable in predicting the response (see Hastie et al. 2018, Chapter 10; see also Grömping 2015) is also possible with RF. Variable importance of RF is computed by (a) growning a tree, (b) computing the prediction accuracy of the tree in the oob sample are passed down, (c) randomly permuting the j^{th} variable in the oob sample and recomputing the prediction accuracy. The variable importance of the j^{th} variable is equal to the decrease in accuracy after permuting in all trees and averaging the results. Negative variable importance means that inclusion of the predictor variables results in decrease of the performance of the algorithm. Positive values indicate positive contribution in the prediction of the algorithm, while the magnitude of the contribution is related to the relative contribution of all variables, as estimated from their respective variable importance.

Appendix C Used software

All analyses and visualizations were conducted in R Programming Language (R Core Team 2018) using the following packages: `caret` (Kuhn 2008, 2018), `devtools` (Wickham et al. 2018), `gdata` (Warnes et al. 2017), `ggplot2` (Wickham 2016; Wickham et al. 2018), `knitr` (Xie 2014, 2015, 2018), `olsrr` (Hebbali 2018), `randomForest` (Liaw and Wiener 2002; Breiman et al. 2018), `readr` (Wickham et al. 2017), `relaimpo` (Grömping 2007b, 2018), `reshape2` (Wickham 2007, 2017), `rmarkdown` (Allaire et al. 2018), `SpatialExtremes` (Ribatet 2018), `stringi` (Gagolewski 2018).

References

[1] Addor N, Nearing G, Prieto C, Newman AJ, Le Vine N, Clark MP (2018) A ranking of hydrological signatures based on their predictability in space. Water Resources Research. https://doi.org/10.1029/2018WR022606

[2] Addor N, Newman AJ, Mizukami N, Clark MP (2017a) Catchment attributes for large-sample studies. Boulder, CO: UCAR/NCAR. https://doi.org/10.5065/D6G73C3Q

[3] Addor N, Newman AJ, Mizukami N, Clark MP (2017b) The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences 21: 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

[4] Ahn KH, Palmer R (2016) Regional flood frequency analysis using spatial proximity and basin characteristics: Quantile regression vs. parameter regression technique. Journal of Hydrology 540:515–526. https://doi.org/10.1016/j.jhydrol.2016.06.047

[5] Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

[6] Allaire JJ, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W (2018) rmarkdown: Dynamic Documents for R. R package version 1.10. https://CRAN.R-project.org/package=rmarkdown

[7] Beck HE, De Roo A, van Dijk AIJM (2015) Global Maps of Streamflow Characteristics Based on Observations from Several Thousand Catchments. Journal of Hydrometeorology 16:1478–1501. https://doi.org/10.1175/JHM-D-14-0155.1

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

[9] Blöschl G, Sivapalan M (1997) Process controls on regional flood frequency: Coefficient of variation and basin scale. Water Resources Research 33(12):2967–2980. https://doi.org/10.1029/97WR00568

[10] Blum AG, Archfield SA, Vogel RM (2017) On the probability distribution of daily streamflow in the United States. Hydrology and Earth System Sciences 21:3093–3103. https://doi.org/10.5194/hess-21-3093-2017

[11] Boulesteix AL, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mining and Knowledge Discovery 2(6):493–507. https://doi.org/10.1002/widm.1072

[12] Boulesteix AL, Janitza S, Hapfelmeier A, Van Steen K, Strobl C (2015) Letter to the Editor: On the term 'interaction' and related phrases in the literature on Random Forests. Briefings in Bioinformatics 16(2):338–345. https://doi.org/10.1093/bib/bbu012

[13] Boulesteix AL, Schmid M (2014) Machine learning versus statistical modeling. Biometrical Journal 56(4):588–593. https://doi.org/10.1002/bimj.201300226

[14] Bourgin F, Andréassian V, Perrin C, Oudin L (2015) Transferring global uncertainty estimates from gauged to ungauged catchments. Hydrology and Earth System Sciences 19:2535–2546. https://doi.org/10.5194/hess-19-2535-2015

[15] Breiman L (2001) Statistical Modeling: The Two Cultures. Statistical Science 16(3):199–231. https://doi.org/10.1214/ss/1009213726

[16] Breiman L (2001) Random Forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324

[17] Breiman L, Cutler A, Liaw A, Wiener M (2018) randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-14. https://CRAN.R-project.org/package=randomForest

[18] Burlando P, Rosso R (1996) Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. Journal of Hydrology 187(1–2):45–64. https://doi.org/10.1016/S0022-1694(96)03086-7

[19] Casella G, Berger RL (2002) Statistical Inference. Duxbury, Pacific Grove, California

[20] Coles GS (2001) An Introduction to Statistical Modeling of Extreme Values. Springer, New York. https://doi.org/10.1007/978-1-4471-3675-0

[21] Demšar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7:1–30

[22] Dey DK, Roy D, Yan J (2016) Univariate Extreme Value Analysis. In: Dey DK, Yan J (Eds) Extreme Value Modeling and Risk Analysis, Methods and Applications. CRC Press, pp. 1–22

[23] Díaz-Uriarte R, De Andres SA (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(3). https://doi.org/10.1186/1471-2105-7-3

[24] Gagolewski M (2018) stringi: Character String Processing Facilities. R package version 1.2.4. https://CRAN.R-project.org/package=stringi

[25] Gaume E, Gaál L, Viglione A, Szolgay J, Kohnová S, Blöschl G (2010) Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. Journal of Hydrology 394(1–2):101–117. https://doi.org/10.1016/j.jhydrol.2010.01.008

[26] Gelman A, Hennig C (2017) Beyond subjective and objective in statistics. Journal of the Royal Statistical Society: Series A 180(4):967–1033. https://doi.org/10.1111/rssa.12276

[27] Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognition Letters 31(14):2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

[28] Gleeson T, Moosdorf N, Hartmann J, Beek LPH (2014) A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity. Geophysical Research Letters 41(11):3891–3898. https://doi.org/10.1002/2014GL059856

[29] Griffis VW, Stedinger JR (2007) Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. I: Distribution Characteristics. Journal of Hydrologic Engineering 12(5). https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(482)

[30] Grömping U (2007a) Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. The American Statistician 61(2):139–147. https://doi.org/10.1198/000313007X188252

[31] Grömping U (2007b) Relative Importance for Linear Regression in R: The Package relaimpo. Journal of Statistical Software 17(1). https://doi.org/10.18637/jss.v017.i01

[32] Grömping U (2009) Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician 63(4):308–319. https://doi.org/10.1198/tast.2009.08199

[33] Grömping U (2015) Variable importance in regression models. WIREs Computational Statistics 7(2):137–152. https://doi.org/10.1002/wics.1346

[34] Grömping U (2018) relaimpo: Relative Importance of Regressors in Linear Models. R package version 2.2-3. https://CRAN.R-project.org/package=relaimpo

[35] Gupta HV, Wagener T, Liu Y (2008) Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrological Processes 22(18):3802–3813. https://doi.org/10.1002/hyp.6989

[36] Gupta VK, Dawdy DR (1995) Physical interpretations of regional variations in the scaling exponents of flood quantiles. Hydrological Processes 9(3-4):347–361. https://doi.org/10.1002/hyp.3360090309

[37] Gupta VK, Mesa OJ, Dawdy DR (1994) Multiscaling theory of flood peaks: Regional quantile analysis. Water Resources Research 30(12):3405–3421. https://doi.org/10.1029/94WR01791

[38] Gupta VK, Troutman BM, Dawdy DR (2007) Towards a Nonlinear Geophysical Theory of Floods in River Networks: An Overview of 20 Years of Progress. In: Tsonis AA, Elsner JB (Eds) Nonlinear Dynamics in Geosciences. Springer, New York, NY, pp.121–151. https://doi.org/10.1007/978-0-387-34918-3_8

[39] Gupta VK, Waymire E (1990) Multiscaling properties of spatial rainfall and river flow distributions. Journal of Geophysical Research 95(D3):1999–2009. https://doi.org/10.1029/JD095iD03p01999

[40] Gvoždíková B, Müller M (2017) Evaluation of extensive floods in western/central Europe. Hydrology and Earth System Sciences 21:3715–3725. https://doi.org/10.5194/hess-21-3715-2017

[41] Haddad K, Rahman A, Stedinger JR (2012) Regional flood frequency analysis using Bayesian generalized least squares: a comparison between quantile and parameter regression techniques. Hydrological Processes 26(7):1008–1021. https://doi.org/10.1002/hyp.8189

[42] Hartmann J, Moosdorf N (2012) The new global lithological map database GLiM: A representation of rock properties at the Earth surface. Geochemistry, Geophysics, Geosystems 13(Q12004). https://doi.org/10.1029/2012GC004370

[43] Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning. Springer-Verlag New York. https://doi.org/10.1007/978-0-387-84858-7

[44] He J, Anderson A, Valeo C (2015) Bias compensation in flood frequency analysis. Hydrological Sciences Journal 60(3):381–401. https://doi.org/10.1080/02626667.2014.885651

[45] Hebbali A (2018) olsrr: Tools for Building OLS Regression Models. R package version 0.5.1. https://CRAN.R-project.org/package=olsrr

[46] Hosking JRM, Wallis JR (1997) Regional frequency analysis. Cambridge University Press, New York. https://doi.org/10.1017/CBO9780511529443

[47] Hrachowitz M, Savenije HHG, Blöschl G, McDonnell JJ, Sivapalan M, Pomeroy JW, Arheimer B, Blume T, Clark MP, Ehret U, Fenicia F, Freer JE, Gelfan A, Gupta HV, Hughes DA, Hut RW, Montanari A, Pande S, Tetzlaff D, Troch PA, Uhlenbrook S, Wagener T, Winsemius HC, Woods RA, Zehe E, Cudennec C (2013) A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological Sciences Journal 58(6):1198–1255. https://doi.org/10.1080/02626667.2013.803183

[48] James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning. Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7

[49] Koutsoyiannis D (2004) Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. Hydrological Sciences Journal 49(4):575–590. https://doi.org/10.1623/hysj.49.4.575.54430

[50] Kuhn M (2008) Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28(5). https://doi.org/10.18637/jss.v028.i05

[51] Kuhn M (2018) caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret

[52] Kuhn M, Johnson K (2013) Applied Predictive Modeling. Springer-Verlag New York. https://doi.org/10.1007/978-1-4614-6849-3

[53] Kuzuha Y, Tomosugi K, Kishii T, Komatsu K (2009) Coefficient of variation of annual flood peaks: variability of flood peak and rainfall intensity. Hydrological Processes 23(4):546–558. https://doi.org/10.1002/hyp.7184

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

[55] Lindeman RH, Merenda PF, Gold RZ (1980) Introduction to Bivariate and Multivariate Analysis. Scott, Foresman, Glenview, IL

[56] Lima CHR, Lall U (2010) Spatial scaling in a changing climate: A hierarchical Bayesian model for non-stationary multi-site annual maximum and monthly streamflow. Journal of Hydrology 383(3–4):307–318. https://doi.org/10.1016/j.jhydrol.2009.12.045

[57] Lima CHR, Lall U, Troy T, Devineni N (2016) A hierarchical Bayesian GEV model for improving local and regional flood quantile estimates. Journal of Hydrology 541(Part B):816–823. https://doi.org/10.1016/j.jhydrol.2016.07.042

[58] McMillan H, Westerberg I, Branger F (2017) Five guidelines for selecting hydrological signatures. Hydrological Processes 31(26):4757–4761. https://doi.org/10.1002/hyp.11300

[59] Merz R, Blöschl G (2005) Flood frequency regionalisation—spatial proximity vs. catchment attributes. Journal of Hydrology 302(1–4):283–306. https://doi.org/10.1016/j.jhydrol.2004.07.018

[60] Miller DA, White RA (1998) A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling. Earth Interactions 2(2):1–26. https://doi.org/10.1175/1087-3562(1998)002<0001:ACUSMS>2.3.CO;2

[61] Morrison JE, Smith JA (2001) Scaling Properties of Flood Peaks. Extremes 4(1):5–22. https://doi.org/10.1023/A:1012268216138

[62] Morrison JE, Smith JA (2002) Stochastic modeling of flood peaks using the generalized extreme value distribution. Water Resources Research 38(12):41-1–41-12. https://doi.org/10.1029/2001WR000502

[63] Newman AJ, Clark MP, Sampson K, Wood A, Hay LE, Bock A, Viger RJ, Blodgett D, Brekke L, Arnold JR, Hopson T, Duan Q (2015) Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences 19:209–223. https://doi.org/10.5194/hess-19-209-2015

[64] Newman AJ, Mizukami N, Clark MP, Wood AW, Nijssen B, Nearing G (2017) Benchmarking of a Physically Based Hydrologic Model. Journal of Hydrometeorology 18:2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1

[65] Newman AJ, Sampson K, Clark MP, Bock A, Viger RJ, Blodgett D (2014) A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. Boulder, CO: UCAR/NCAR. https://doi.org/10.5065/D6MW2F4D

[66] Northrop PJ (2004) Likelihood-based approaches to flood frequency estimation. Journal of Hydrology 292(1–4): 96–113. https://doi.org/10.1016/j.jhydrol.2003.12.031

[67] Northrop PJ, Attalides N (2016) Posterior propriety in Bayesian extreme value analyses using reference priors. Statistica Sinica 26(2):721–743. https://doi.org/10.5705/ss.2014.034

[68] O'Brien RM (2007) A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality and Quantity 41(5):673–690. https://doi.org/10.1007/s11135-006-9018-6

[69] Odry J, Arnaud P (2017) Comparison of Flood Frequency Analysis Methods for Ungauged Catchments in France. Geosciences 7(3):88. https://doi.org/10.3390/geosciences7030088

[70] Oshiro TM, Perez PS, Baranauskas JA (2012) How many trees in a random forest?. In: Perner P (Ed) Machine Learning and Data Mining in Pattern Recognition (Lecture Notes in Computer Science). Springer-Verlag Berlin Heidelberg, IBaI, Leipzig, Germany, 2012; Volume 7376, pp. 154–168. https://doi.org/10.1007/978-3-642-31537-4

[71] Ouali D, Chebana F, Ouarda TBMJ (2016) Quantile Regression in Regional Frequency Analysis: A Better Exploitation of the Available Information. Journal of Hydrometeorology 17:1869–1883. https://doi.org/10.1175/JHM-D-15-0187.1

[72] Ouali D, Chebana F, Ouarda TBMJ (2017) Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites. Journal of Advances in Modeling Earth Systems 9(2):1292–1306. https://doi.org/10.1002/2016MS000830

[73] Papacharalampous GA, Tyralis H (2018) Evaluation of random forests and Prophet for daily streamflow forecasting. Advances in Geosciences 45:201–208. https://doi.org/10.5194/adgeo-45-201-2018

[74] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018a) One-step ahead forecasting of geophysical processes within a purely statistical framework. Geoscience Letters. https://doi.org/10.1186/s40562-018-0111-1

[75] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018b) Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. Preprints 2017100133. https://doi.org/10.20944/preprints201710.0133.v2

[76] Parkes B, Demeritt D (2016) Defining the hundred year flood: A Bayesian approach for using historic data to reduce uncertainty in flood frequency estimates. Journal of Hydrology 540:1189–1208. https://doi.org/10.1016/j.jhydrol.2016.07.025

[77] Pelletier JD, Broxton PD, Hazenberg P, Zeng X, Troch PA, Niu G-Y, Williams Z, Brunke MA, Gochis D (2016) A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. Journal of Advances in Modeling Earth Systems 8(1):41–65. https://doi.org/10.1002/2015MS000526

[78] Probst P, Boulesteix AL (2018) To Tune or Not to Tune the Number of Trees in Random Forest. Journal of Machine Learning Research 18(181):1–18

[79] R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[80] Rahman A, Charron C, Ouarda TBMJ, Chebana F (2018) Development of regional flood frequency analysis techniques using generalized additive models for Australia. Stochastic Environmental Research and Risk Assessment 32(1):123–139. https://doi.org/10.1007/s00477-017-1384-1

[81] Reiss RD, Thomas M (2007) Statistical Analysis of Extreme Values. Birkhäuser Basel. https://doi.org/10.1007/978-3-7643-7399-3

[82] Requena AI, Chebana F, Ouarda TBMJ (2017) Heterogeneity measures in hydrological frequency analysis: review and new developments. Hydrology and Earth System Sciences 21:1651–1668. https://doi.org/10.5194/hess-21-1651-2017

[83] Ribatet M (2018) SpatialExtremes: Modelling Spatial Extremes. R package version 2.0-7. https://CRAN.R-project.org/package=SpatialExtremes

[84] Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54(3):507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

[85] Scornet E, Biau G, Vert JP (2015) Consistency of random forests. The Annals of Statistics 43(4):1716–1741. https://doi.org/10.1214/15-AOS1321

[86] Shafii M, Tolson BA (2015) Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. Water Resources Research 51(5):3796–3814. https://doi.org/10.1002/2014WR016520

[87] Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3–4):591–611. https://doi.org/10.2307/2333709

[88] Shmueli G (2010) To Explain or to Predict?. Statistical Science 25(3):289–310. https://doi.org/10.1214/10-STS330

[89] Singh R, Archfield SA, Wagener T (2014) Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments – A comparative hydrology approach. Journal of Hydrology 517:985–996. https://doi.org/10.1016/j.jhydrol.2014.06.030

[90] Smith JA, Villarini G, Baeck ML (2011) Mixture Distributions and the Hydroclimatology of Extreme Rainfall and Flooding in the Eastern United States. Journal of Hydrometeorology 12:294–309. https://doi.org/10.1175/2010JHM1242.1

[91] Stedinger JR, Tasker GD (1985) Regional Hydrologic Analysis: 1. Ordinary, Weighted, and Generalized Least Squares Compared. Water Resources Research 21(9):1421–1432. https://doi.org/10.1029/WR021i009p01421

[92] Stedinger JR, Griffis VW (2008) Flood Frequency Analysis in the United States: Time to Update. Journal of Hydrologic Engineering 13(4). https://doi.org/10.1061/(ASCE)1084-0699(2008)13:4(199)

[93] Stedinger JR, Vogel RM, Foufoula-Georgiou E (1993) Frequency Analysis of Extreme Events. In: Maidment DR (ed) Handbook of Hydrology, 1st edn. McGraw Hill Education, pp 18.1–18.66

[94] Strobl C, Malley J, Tutz G (2009) An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychological Methods 14(4):323–348

[95] Schwarz G (1978) Estimating the Dimension of a Model. The Annals of Statistics 6(2):461–464. https://doi.org/10.1214/aos/1176344136

[96] Thorarinsdottir TL, Hellton KH, Steinbakk GH, Schlichting L, Engeland K (2018) Bayesian Regional Flood Frequency Analysis for Large Catchments. Water Resources Research. https://doi.org/10.1029/2017WR022460

[97] Thornton PE, Thornton MM, Mayer BW, Wilhelmi N, Wei Y, Devarakonda R, Cook RB (2014) Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. ORNL DAAC, Oak Ridge, Tennessee, USA. Date accessed: 2016/01/20. https://doi.org/10.3334/ORNLDAAC/1219

[98] Tyralis H, Dimitriadis P, Koutsoyiannis D, O'Connell PE, Tzouka K, Iliopoulou T (2018) On the long-range dependence properties of annual precipitation using a global network of instrumental measurements. Advances in Water Resources 111:301–318. https://doi.org/10.1016/j.advwatres.2017.11.010

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

[100] Tyralis H, Langousis A (2018) Estimation of intensity–duration–frequency curves using max-stable processes. Stochastic Environmental Research and Risk Assessment. https://doi.org/10.1007/s00477-018-1577-2

[101] Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: A survey and results of new tests. Pattern Recognition 44(2):330–349. https://doi.org/10.1016/j.patcog.2010.08.011

[102] Veneziano D, Langousis A (2010) Scaling and fractals in hydrology. In: Sivakumar B, Berndtsson R (Eds) Advances in data-based approaches for hydrologic modeling and forecasting. World Scientific, Singapore, pp.107–243. https://doi.org/10.1142/9789814307987_0004

[103] Viglione A, Merz R, Salinas JL, Blöschl G (2013a) Flood frequency hydrology: 3. A Bayesian analysis. Water Resources Research 49(2):675–692. https://doi.org/10.1029/2011WR010782

[104] Viglione A, Parajka J, Rogger M, Salinas JL, Laaha G, Sivapalan M, Blöschl G (2013b) Comparative assessment of predictions in ungauged basins – Part 3: Runoff signatures in Austria. Hydrology and Earth System Sciences 17:2263–2279. https://doi.org/10.5194/hess-17-2263-2013

[105] Villarini G, Smith JA (2010) Flood peak distributions for the eastern United States. Water Resources Research 46(W06504). https://doi.org/10.1029/2009WR008395

[106] Villarini G, Smith JA, Baeck ML, Krajewski WF (2011a) Examining Flood Frequency Distributions in the Midwest U.S.. Journal of the American Water Resources Association 47(3): 447–463. https://doi.org/10.1111/j.1752-1688.2011.00540.x

[107] Villarini G, Smith JA, Serinaldi F, Ntelekos AA (2011b) Analyses of seasonal and annual maximum daily discharge records for central Europe. Journal of Hydrology 399(3–4):299–312. https://doi.org/10.1016/j.jhydrol.2011.01.007

[108] Villarini G, Smith JA, Serinaldi F, Ntelekos AA, Schwarz U (2012) Analyses of extreme flooding in Austria over the period 1951–2006. International Journal of Climatology 32(8):1178–1192. https://doi.org/10.1002/joc.2331

[109] Vogel RM, Wilson I (1996) Probability Distribution of Annual Maximum, Mean, and Minimum Streamflows in the United States. Journal of Hydrologic Engineering 1(2). https://doi.org/10.1061/(ASCE)1084-0699(1996)1:2(69)

[110] Vogel RM, Sankarasubramanian A (2000) Spatial scaling properties of annual streamflow in the United States. Hydrological Sciences Journal 45(3):465–476. https://doi.org/10.1080/02626660009492342

[111] Wagener T, Montanari A (2011) Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. Water Resources Research 47(W06301). https://doi.org/10.1029/2010WR009469

[112] Wallis JR, Schaefer MG, Barker BL, Taylor GH (2007) Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington State. Hydrology and Earth Systems Sciences 11:415–442. https://doi.org/10.5194/hess-11-415-2007

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

[114] Westerberg IK, McMillan HK (2015) Uncertainty in hydrological signatures. Hydrology and Earth System Sciences 19:3951–3968. https://doi.org/10.5194/hess-19-3951-2015

[115] Westerberg IK, Wagener T, Coxon G, McMillan HK, Castellarin A, Montanari A, Freer J (2016) Uncertainty in hydrological signatures for gauged and ungauged catchments. Water Resources Research 52(3):1847–1865. https://doi.org/10.1002/2015WR017635

[116] Wickham H (2007) Reshaping Data with the reshape Package. Journal of the Statistical Software 21(12). https://doi.org/10.18637/jss.v021.i12

[117] Wickham H (2016) ggplot2. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

[118] Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K (2018) ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.1.0. https://CRAN.R-project.org/package=ggplot2

[119] Wickham H (2017) reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package. R package version 1.4.3. https://CRAN.R-project.org/package=reshape2

[120] Wickham H, Hester J, Chang W (2018) devtools: Tools to Make Developing R Packages Easier. R package version 2.0.1. https://CRAN.R-project.org/package=devtools

[121] Wickham H, Hester J, Francois R (2017) readr: Read Rectangular Text Data. R package version 1.1.1. https://CRAN.R-project.org/package=readr

[122] Wilcoxon F (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 1(6):80–83. https://doi.org/10.2307/3001968

[123] Wu Y, Lall U, Lima CHR, Zhong P (2018) Local and regional flood frequency analysis based on hierarchical Bayesian model: application to annual maximum streamflow for the Huaihe River basin. Hydrology and Earth System Sciences Discussions. https://doi.org/10.5194/hess-2018-22

[124] Xie Y (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In: Stodden V, Leisch F, Peng RD (Eds) Implementing Reproducible Computational Research. Chapman and Hall/CRC

[125] Xie Y (2015) Dynamic Documents with R and knitr, 2nd edition. Chapman and Hall/CRC

[126] Xie Y (2018) knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.20. https://CRAN.R-project.org/package=knitr

[127] Xu YP, Booij MJ, Tong YB (2010) Uncertainty analysis in statistical modeling of extreme hydrological events. Stochastic Environmental Research and Risk Assessment 24(5):567–578. https://doi.org/10.1007/s00477-009-0337-8

[128] Yan H, Moradkhani H (2015) A regional Bayesian hierarchical model for flood frequency analysis. Stochastic Environmental Research and Risk Assessment 29(3):1019–1036. https://doi.org/10.1007/s00477-014-0975-3

[129] Yan H, Moradkhani H (2016) Toward more robust extreme flood prediction by Bayesian hierarchical and multimodeling. Natural Hazards 81(1):203–225. https://doi.org/10.1007/s11069-015-2070-6

[130] Zhang Y, Chiew FHS, Li M, Post D (2018) Predicting Runoff Signatures Using Regression and Hydrological Modeling Approaches. Water Resources Research. https://doi.org/10.1029/2018WR023325

[131] Ziegler A, König IR (2014) Mining data with random forests: current options for real-world applications. WIREs Data Mining and Knowledge Discovery 4(1):55–63. https://doi.org/10.1002/widm.1114