Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR030272

A Posteriori Random Forests for Stochastic Downscaling of Precipitation by Predicting Probability Distributions

M. N. Legasa[1], R. Manzanas[1], A. Calviño[2], and J. M. Gutiérrez[3]

[1]Meteorology Group, Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Santander, Spain, [2]Department of Statistics and Data Science, Complutense University of Madrid, Madrid, Spain, [3]Meteorology Group, Instituto de Física de Cantabria (IFCA), Universidad de Cantabria-CSIC, Santander, Spain

Key Points:

- We thoroughly assess different configuration options of the random forest technique for statistical downscaling of precipitation
- We introduce a posteriori random forests, which reliably estimate the probability distribution of rainfall on wet days
- As a consequence, this new methodology allows us to generate realistic stochastic precipitation amounts without loss of predictive power

Correspondence to:
M. N. Legasa,
mikel.legasa@unican.es

Citation:
Legasa, M. N., Manzanas, R., Calviño, A., & Gutiérrez, J. M. (2022). A posteriori random forests for stochastic downscaling of precipitation by predicting probability distributions. *Water Resources Research*, *58*, e2021WR030272. https://doi.org/10.1029/2021WR030272

Received 21 APR 2021
Accepted 16 MAR 2022

Abstract This work presents a comprehensive assessment of the suitability of random forests, a well-known machine learning technique, for the statistical downscaling of precipitation. Building on the experimental and validation framework proposed in the Experiment 1 of the COST action VALUE—the largest, most exhaustive intercomparison study of statistical downscaling methods to date—we introduce and thoroughly analyze a posteriori random forests (AP-RFs), which use all the information contained in the leaves to reliably predict the shape and scale parameters of the gamma probability distribution of precipitation on wet days. Therefore, as opposed to traditional random forests, which typically provide deterministic predictions, our AP-RFs allow realistic stochastic precipitation samples to be generated for wet days. Indeed, as compared to one particular implementation of a generalized linear model that exhibited an overall good performance in VALUE, our AP-RFs yield better distributional similarity with observations without loss of predictive power. Noteworthy, the new methodology proposed in this paper has substantial potential for hydrologists and other impact communities which are in need of local-scale, reliable stochastic climate information.

Plain Language Summary Statistical downscaling methods aim to improve the limited spatial resolution of current climate models by linking a set of key large-scale predictor variables (e.g., geopotential, winds, etc.) to the predictand of interest (e.g., precipitation). Recently, the Experiment 1 of the COST action VALUE carried out the most comprehensive intercomparison of statistical downscaling methods to date. However, it lacked the inclusion of machine learning techniques, whose popularity has rapidly grown during the last years. Therefore, building on the same data and experimental framework used in VALUE, this work aims to partially fill this knowledge gap by introducing a modification of random forests—a well-known machine learning technique—for stochastic downscaling of precipitation at 86 European locations. As opposed to traditional random forests, which typically provide deterministic predictions, our proposed model predicts a probability distribution of precipitation for each predictors' state. This is key to appropriately characterize the uncertainty of the downscaled predictions, allowing us to produce realistic samples of precipitation for wet days and to answer questions such as "What is the probability of getting more than 40 mm of precipitation today?," relevant for many impact activities.

1. Introduction

Despite being the main tool used nowadays to simulate the evolution of the climate system, the spatial resolution of the current global climate models (GCMs)—typically up to around a hundred kilometers—is still insufficient for most practical applications (see e.g., Doblas-Reyes et al., 2013 and references therein). To alleviate this limitation, statistical downscaling (SD) methods aim to build models linking a set of key large-scale predictors (e.g., geopotential or winds) with the target predictand (e.g., precipitation or temperature) over the area of interest (see e.g., von Storch et al., 1993). Under the perfect prognosis (PP) approach (Charles et al., 1999; Bürger & Chen, 2005; Fowler et al., 2007; Gutiérrez et al., 2013; Haylock et al., 2006; Hertig & Jacobeit, 2008; Sauter & Venema, 2011; Timbal et al., 2003), these models are trained/calibrated based on observed data over a historical reference period, relying on some reanalysis for the predictors. Typically, these trained/calibrated models are subsequently applied to (future) GCM large-scale predictors in order to obtain the corresponding downscaled values. In this work we focus exclusively on the training/calibration phase—i.e., using reanalysis predictors in a PP framework—and leave for a next study the application of the models to GCM predictors.

Many SD methods have been proposed in the literature during the last decades, ranging from techniques like linear or generalized linear models (GLMs), which normally involve a relatively low number of parameters, to highly non-linear deep neural networks, which need to optimize a large number of weights and are thus typically prone to overfitting. In the Experiment 1 of the COST action VALUE (http://www.value-cost.eu), Gutiérrez et al. (2019) intercompared the performance of more than 50 standard (i.e., well-established within the climate community) SD methods in PP conditions to downscale precipitation and temperature at 86 European locations. Although this study provides the largest and most comprehensive intercomparison of SD methods to date, it lacked the inclusion of machine learning techniques, which still need to be fairly compared against more classical alternatives such as GLMs (Nelder & Wedderburn, 1972) or analogs (Lorenz, 1969; Zorita & von Storch, 1999). Recently, Baño-Medina et al. (2020) completed a first step to fill this gap by assessing the performance of convolutional neural networks using the experimental/validation framework proposed in Gutiérrez et al. (2019). The present work extends the VALUE experiment to include additional machine learning techniques by analyzing the suitability of random forests (RFs; Breiman, 2001) for SD of precipitation on wet days, a key variable for most hydrometeorological applications. To do so, we build as well on the experimental/validation framework proposed in Gutiérrez et al. (2019).

It is important to highlight that the predictive capability of the large-scale predictors used to build statistical models for climate downscaling is by nature limited, especially for precipitation. This implies the existence of a high degree of uncertainty in the predictions, something that can be seen from the limited correlations attained in Figure 2 of Gutiérrez et al. (2019). Hence, the main purpose of this work is to accurately predict the full probability distribution of precipitation instead of just its expected value, which would in turn allow for appropriately characterizing the uncertainty of the downscaled values. This kind of information provides enormous added value for hydrologists and other impact communities, which typically need high-quality, local-scale climate information to run their models subject to a set of different plausible scenarios (Ailliot et al., 2015; Booij, 2005; Chandler, 2020; Farmer & Vogel, 2016; Schlabing et al., 2014). Indeed, several studies have pointed out the importance of having access to reliable stochastic time-series of precipitation, since impact models tend to fail to reproduce the extreme events (e.g., floods) as a consequence of the underestimated variance typically exhibited by deterministic predictions, that *collapse* the uncertainty into a single value (Acharya et al., 2017; Mukundan et al., 2019; Nogueira & Barros, 2015). As suggested by several recent works, this limitation can be mitigated by introducing a meaningful stochastic component of uncertainty in the models (Farmer & Vogel, 2016; Haberlandt et al., 2015; Langousis & Kaleris, 2014; Neves et al., 2020).

Therefore, the potential of RFs to produce reliable stochastic precipitation is assessed in this work. So far, RFs already proved to be useful in regression problems dealing with ecological and environmental variables (Chan & Paelinckx, 2008; Vincenzi et al., 2011; Wei et al., 2010; Yu et al., 2011). Moreover, Ibarra-Berastegi et al. (2011) used RFs to predict three variables—including precipitation—at two locations and tested them against multilinear regression. Based also on RFs, Shi and Song (2015) and He et al. (2016) provided local precipitation fields over the Tibetan Plateau and the United States, respectively. Jing et al. (2016) found that RFs provided better results than three other machine learning techniques for SD over 378 meteorological stations in China encompassing very heterogeneous climate conditions. Furthermore, more recent studies (Sattari et al., 2020; Xiang et al., 2020) have analyzed the applicability of RFs to precipitation prediction, suggesting their potential to detect extreme events. Additionally, Xu et al. (2020) combined multiple machine learning methods—including RFs—to provide downscaled climate change projections using model output statistics (MOS), that is, building on GCM predictors. Still, to our knowledge, none of these studies has performed a comprehensive analysis of the different parameters and configuration options available for RFs, including the choice of the split function, which is a key element. Moreover, these studies focus mostly on specific regions and/or particular aspects of precipitation, and thus a comprehensive assessment of their potential added value for SD of precipitation is still lacking.

This work fills this gap by testing the suitability of a novel modification of traditional RFs, called a posteriori random forests (AP-RFs), for SD of daily precipitation on wet days. As opposed to traditional RFs, which predict a single precipitation value at each time-step following a bagging procedure (Hastie et al., 2009), our AP-RFs allow us to accurately predict the shape and scale parameters of the underlying gamma probabil-

Table 1
Subset of Nine Representative Stations Considered for Some of the Experiments Carried Out in This Work

	ID	Name	(Lon, Lat)	Region	Altitude	Wet days	Annual rainfall
1	003946	MADRID-BARAJAS	(–3.56, 40.47)	Iberia (IP)	609 masl	14.4%	370.2 mm
2	000176	ROMA-CIAMPINO	(12.58, 41.78)	Mediterranean (ME)	105 masl	21.1%	837.8 mm
3	000450	SIBIU	(24.15, 45.8)	Eastern Europe (EE)	444 masl	24.4%	635.4 mm
4	000032	BOURGES	(2.37, 47.07)	France (FR)	161 masl	31.2%	756.5 mm
5	000013	INNSBRUCK	(11.4, 47.27)	Alps (AL)	577 masl	31.7%	884.4 mm
6	002006	BROCKEN	(10.62, 51.8)	Central Europe (CE)	1,142 masl	55%	1881.9 mm
7	000272	ESKDALEMUIR	(–3.2, 55.32)	British Isles (BI)	242 masl	50.1%	1735.5 mm
8	000028	HELSINKI-KAISANIEMI	(24.95, 60.18)	Scandinavia (SC)	4 masl	29.2%	651.9 mm
9	000191	KJOEREMSGRENDE	(9.05, 62.1)	Scandinavia (SC)	626 masl	23.5%	433.9 mm

Note. ID is the station code used in VALUE. The *Wet Days* column corresponds to the percentage of wet days (>1 mm) and *Annual Rainfall* is the mean rainfall accumulated each year along the period 1979–2008. See Figure 1 for a map with the location of the 86 meteorological stations considered.

ity distribution without loss of predictive power, which in turn makes feasible to produce reliable stochastic wet-day series.

The paper is structured as follows: Section 2 describes the experimental framework and data used. Section 3 explains the particularities of the RFs developed. The results obtained are presented through Section 4. Finally, Section 5 gathers the main conclusions drawn.

2. Experimental Framework and Data

The COST action VALUE (Maraun et al., 2015, http://www.value-cost.eu) was designed with the goal of providing a common, transparent, reproducible and comprehensive framework for intercomparing SD methods for climate. In its Experiment 1, Gutiérrez et al. (2019) presented an assessment of the relative strengths and weaknesses of more than 50 well-established techniques—mostly linear and analog-based—to downscale temperature and precipitation over Europe in *perfect* conditions, that is, using quasi-observed predictors from reanalysis. Therefore, in this work we use the same experimental framework and data proposed in Gutiérrez et al. (2019).

In particular, our target predictand is daily precipitation at the same 86 stations considered in the latter reference, which are provided by the European Climate Assessment & Data set project (ECA&D). Nevertheless, for the experiments presented in Sections 4.1, 4.2 and 4.3 (which deal with optimal model parameters selection and are computationally expensive), we consider a subset of 9 illustrative stations which are representative of the 8 PRUDENCE regions covering Europe (Christensen & Christensen, 2007) and exhibit different rainfall regimes (see Table 1 for more details).

Table 2
ERA-Interim Large-Scale Predictors Used in This Work

Description	Height level
Air pressure	Sea level
Geopotential	500, 700, 850 mbar
Temperature	2 m, 500, 700, 850 mbar
Eastward wind	500, 700, 850 mbar
Northward wind	500, 700, 850 mbar
Specific humidity	500, 700, 850 mbar

Note. The reader is referred to Gutiérrez et al. (2019) for details. Since we use the 4 nearest gridboxes, we have 68 individual predictor variables for each station.

As predictors, we have considered the same large-scale variables used in Gutiérrez et al. (2019), which are listed in Table 2. In particular, for each predictand location, ERA-Interim reanalysis (Dee et al., 2011) information at the 4 nearest gridboxes was used. Throughout the manuscript we use X to refer to the set of large-scale predictors and Y to refer to our target downscaled variable (precipitation on wet days).

In addition, we also use the same 5-fold cross-validation employed in Gutiérrez et al. (2019), in which the total period of study, 1979–2008, is divided into 5 non-overlapping folds $\{\mathcal{F}_1, \ldots, \mathcal{F}_5\}$, with each fold comprising 6 years:

Table 3
Subset of the Validation Metrics Used in Gutiérrez et al. (2019) Which Have Been Considered to Assess the Performance of the Different Statistical Downscaling Methods Used in This Work

Code	Description
Mean	Mean rainfall
SD	Standard deviation of rainfall
P05	5th percentile of rainfall
P95	95th percentile of rainfall
P95A	Sum of total rainfall in days exceeding the 95th percentile value
P99	99th percentile of rainfall
P99A	Sum of total rainfall in days exceeding the 99th percentile value
KS	Kolmogorov-Smirnov statistic between the observed and simulated rainfall
Cor	Spearman correlation between the observed and (deterministic) predicted rainfall

Note. Except for correlation and the Kolmogorov-Smirnov statistic, we use the simulated/observed ratio for direct comparison among metrics.

$\mathcal{F}_1 = 1979-1984$, $\mathcal{F}_2 = 1985-1990$, $\mathcal{F}_3 = 1991-1996$, $\mathcal{F}_4 = 1997-2002$, $\mathcal{F}_5 = 2003-2008$. For each of the 5 folds, predictions are obtained based on the method being trained using the remaining 4 folds. By doing so, we end up with an out-of-sample (i.e., based on previously *unseen* data) predicted wet-day series which covers 1979–2008. Then, the so-obtained predictions/simulations are validated in terms of the metrics described in Table 3, which allow us to assess their predictive power (as measured by the Spearman correlation: Cor), ability to reproduce the mean and standard deviation (Mean, SD), low values (P05) and the tail of the distribution (P95, P95A, P99, P99A). Finally, the Kolmogorov-Smirnov statistic, (KS, see Chapter 3 in Cox and Hinkley (1974)), which measures the maximum distance between the observed and simulated cumulative distribution functions, provides a summary for the overall distributional performance of our methods. Note that, except for correlation and the KS statistic, we use the predicted/observed ratio for direct comparison among different metrics and, therefore, higher/lower than 1 values indicate an overestimation/underestimation.

3. Methods

Precipitation is a semi-continuous variable with a probability distribution characterized by a spike at zero (dry days) followed by a continuous distribution with positive support (wet days). For this reason, it is common to treat separately the binary event *occurrence* (dry/wet days) and the continuous event *amount* for wet days (see e.g., Chandler and Wheater, 2002; Manzanas et al., 2015). This is the approach followed here for the two downscaling techniques considered, which are next described in detail: random forests and GLMs. Note however that the present work focuses exclusively on the prediction of the precipitation amount distribution on wet days.

Here the rainfall on each wet day is assumed to follow a two parameter gamma distribution (Husak et al., 2007; Martínez-Villalobos & Neelin, 2019; Richardson, 1981; Wilks & Eggleston, 1992), which can be characterized by the parameters α and β (shape and rate, respectively), with its probability density function defined for positive values of y as

$$f(y) = \frac{\beta^{\alpha} y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)},$$

where $\Gamma(\alpha) = \int_0^{\infty} z^{\alpha-1} e^{-z} dz$ is the gamma function.

As we will later explain, the methodology introduced in this article requires computationally efficient and reliable methods for estimating the parameters of the gamma distribution (notice that the methodology can be easily adapted to other probability distributions). This poses a problem, since no closed form unbiased estimators for α and β exist. In particular, and despite the fact that both are *consistent* estimators (i.e., $\tilde{\theta} \to \theta$ as sample size increases for an estimator $\tilde{\theta}$ of θ), the widely used maximum likelihood estimators (MLE) and the moments matching estimators (MME) are positively biased (Choi & Wette, 1969; Smith et al., 2009). A positive bias for an

estimator $\tilde{\theta}$ of θ means that $E\left(\tilde{\theta}\right) > \theta$ (Cox & Hinkley, 1974, Chapter 8). Furthermore, there is no closed form for MLE and it must thus be computed numerically solving two simultaneous equations, which involve the *digamma* function (Choi & Wette, 1969; Song, 2008). MME estimators are defined as $\alpha_{\text{MME}} = \overline{y}^2/s$, $\beta_{\text{MME}} = \overline{y}/s$ for the mean $\overline{y}$ and standard deviation s of a sample $\{y\}$. Ye and Chen (2017) and Louzada et al. (2019) both proposed alternatives for less biased closed-form estimators. In particular, the latter work proposed an estimator based on the *hybrid* MLE that exhibits less bias for very small samples. This estimator, which is called BC3 by the authors, is defined for α as

$$\alpha_{\text{BC3}} = \dot{\alpha} - \frac{1}{n}\left(3\dot{\alpha} - \frac{2\dot{\alpha}}{3(1+\dot{\alpha})} - \frac{4\dot{\alpha}}{5(1+\dot{\alpha})^2}\right), \tag{1}$$

with $\dot{\alpha} = \overline{y}/\sum_{y\in\{y\}}(y-\bar{y})\log(y)$. β can be trivially estimated as $\beta_{\text{BC3}} = \alpha_{\text{BC3}}/\overline{y}$.

3.1. Random Forests

Classification and regression trees (CART) are predictive models that separate the target variable into homogeneous groups according to a sequence of *if-else* rules applied on the predictors' space. Being attractively simple and interpretable, CART are capable of capturing non-linearities in the data and can thus be used in problems of different complexities and nature. However, these models present several weaknesses that hinder their application in real-life problems, mainly their lack of stability and tendency to overfit. The well-known machine learning technique called random forest (Breiman, 2001; Hastie et al., 2009) alleviates these issues by combining an ensemble of K CART to predict the random variable Y (in this work rainfall on wet days at a meteorological station) from the predictors X (in this work a set of m reanalysis large-scale variables). Note that CART (and thus random forests) are not hindered by the different scales and variances of the predictors, since they are based on recursively splitting the predictors space. Therefore, reanalysis predictors are not standardized in the RFs used in this work.

Each of these t_k trees maps a combination of predictors $x \in X$ to a *leaf* containing a set $\{y \in Y\}_{t_k(x)}$ of observed records. This mapping is constructed by means of a *split function*, which recursively chooses both the predictor variable X^0 and its threshold value $X^0 = x_0^0$ that best separate a subset Y_0 of observations drawn from Y into $Y_+ = \left\{y \in Y_0 \mid X^0 \geq x_0^0\right\}$ and $Y_- = \left\{y \in Y_0 \mid X^0 < x_0^0\right\}$. In particular, for the mean observed value of each subset ($\overline{y}_0, \overline{y}_+$ and $\overline{y}_-$ for Y_0, Y_+ and Y_-, respectively), $X^0 = x_0^0$ is appropriately selected to maximize the reduction of the average error committed in the leaves $Err\left(\overline{y}_0, y\right) - \left(Err\left(\overline{y}_+, y\right) + Err\left(\overline{y}_-, y\right)\right)/2$, where $Err\left(\overline{y}, y\right)$ measures how well $\overline{y}$ represents (or *summarizes*) the observed values for each set (hence we call $\overline{y}$ the *summary measure*). Note that we will use the term *split function* to refer to the combination of the error function Err and summary measure $\overline{y}$ computed on the leaves henceforth.

In regression problems, the mean squared error (MSE) is typically used as Err (see e.g., He et al., 2016 or Xu et al., 2020), that is, $Err\left(\overline{y}_0, y\right) = \frac{1}{|Y_0|}\sum_{y\in Y_0}\left(y - \overline{y}_0\right)^2$ and the individual predictions $\overline{y}_{t_k(x)}$ produced by each tree are directly averaged to obtain a final prediction $\hat{y}$, that is,

$$\hat{y}(x) = \frac{1}{K}\sum_{k=1}^{K}\overline{y}_{t_k(x)}. \tag{2}$$

However, it is well known that the use of MSE as split function maximizes the likelihood function assuming $Y \mid X$ follows a gaussian distribution with constant variance (Hastie et al., 2009, Chapter 2), which is not the case for precipitation. Indeed, we assume precipitation to follow a two-parameter gamma distribution here, which poses an additional problem: the mean no longer suffices as summary measure $\overline{y}$, we must provide both α and β for each $X = x$ to fully characterize $Y \mid X = x$.

Using the log-likelihood as error function is a natural choice to tackle these issues. For a continuous variable Y and a sample $\{y\}$ drawn from Y, the log-likelihood function is defined as $L(\theta, \{y\}) = \sum_{y\in\{y\}}\log(f(y \mid \theta))$, where f is the probability density function of Y and θ are the parameters of this distribution. Note that this definition assumes all elements in the sample to be drawn independently, which is a reasonable assumption, since the lag-1 autocorrelation (i.e., the correlation between precipitation amount on one wet day and the previous wet day) is already small. Indeed, its median value across the 86 stations is ≈0.089, being smaller than 0.23 in all cases.

Table 4
The Different Split Functions and Their Summary Measures Used in This Work

Split name	Summary measure	Description
MSE	Mean $\overline{y}$	Mean squared error
Deviance	Mean $\overline{y}$	Deviance for the gamma distribution
NLL-MME	$\alpha_{\text{MME}} = \overline{y}^2/s$, $\beta_{\text{MME}} = \overline{y}/s$	Negative log-likelihood using moments matching estimators
NLL-BC3	See formula 1	Negative log-likelihood using BC3 estimators defined in Louzada et al. (2019)

In this work, the actual error function *Err* corresponds to the *negative* log-likelihood (NLL), which is *lower* for better-split subsets. For the gamma distribution f, the negative log-likelihood is derived from the definition given in Section 2,

$$NLL(\alpha, \beta, \{y\}) = \sum_{y \in \{y\}} -\alpha \log(\beta) - (\alpha - 1)\log(y) + \beta y + \log(\Gamma(\alpha)).$$

α and β need to be estimated for each split candidate (these are the summary measures used instead of the mean $\overline{y}$), and the final prediction of the RF is the average of these two estimates predicted by each individual tree. For the log-likelihood split function, we consider in this work the MME and BC3 estimators discussed in the previous section (see Table 4).

However, as an alternative that does not require the estimation of α and β at each split in each tree, we have also considered the gamma deviance (Dunn & Smyth, 2018, Chapter 11), whose formula is

$$deviance\left(\overline{y}, \{y\}\right) = 2 \sum_{y \in \{y\}} \left[-\log\left(\frac{y}{\overline{y}}\right) + \frac{y - \overline{y}}{\overline{y}} \right].$$

It can be used as a split function with the mean of all the elements in the node $\overline{y}$ as summary measure. Although it is an approximation which assumes one of the parameters of the gamma distribution to be fixed, the deviance has the advantage of making the split function faster and potentially less likely to be biased by estimation error at each candidate split. We analyze in Section 4.1 the effect that the choice of split function may have to obtain reliable estimates of $Y \mid X$. To do so, the 4 split functions listed in Table 4 are considered.

Beyond the issue of the choice of split function, trees are known to be prone to overfitting to the training data set (Hastie et al., 2009, Chapter 9). Thus, their depth must be carefully prevented to grow too much. This is typically accomplished by either directly applying some restriction to the allowed maximum depth or by setting a minimum number of elements that each leaf must have (we call this parameter the minimum leaf size). As we will explain in Section 4.3, setting up the optimal configuration/complexity of the forest based on these two parameters is not a straightforward task. Note that additional techniques like pruning and regularization (Hastie et al., 2009, see Sections 9.2.2 and 10.12) are an option typically used for individual trees, but they are not usually employed in RFs and will not be considered in this work.

Moreover, RFs introduce two additional mechanisms to reduce the risk of overfitting, make the individual tress be less correlated and increase the predictive capability of the ensemble model: First, each individual tree is trained with a bootstrap sample drawn from the original data set. Second, for each split in each individual tree, only a random subset of the total available predictors, m, is considered as split candidate. In this work we use $m = 22$, that is, one third of the total 68 predictor variables available, which corresponds to the standard choice for this parameter in regression problems (note that other values of m were also tested, but none yielded substantially better performance than $m = 22$).

3.2. A Posteriori Random Forests

For a given instance of the predictors $X = x$, a random forest with K trees produces K subsets of *predictive observations* $\{y \in Y\}_{t_k(x)}$, whose size depends on the minimum leaf size. Instead of producing a single prediction by averaging all the summary measures computed for each $\{y \in Y\}_{t_k(x)}$, which we call the *averaging* (AVG-RF)

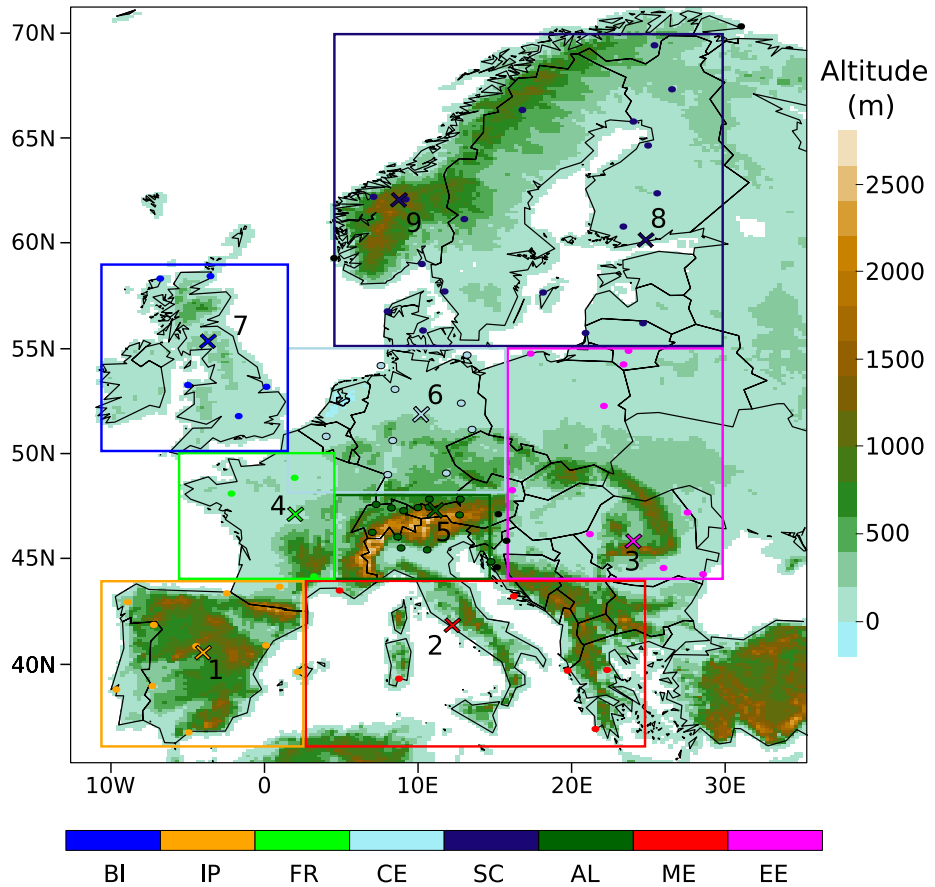

Figure 1. Location of the 86 meteorological stations considered in the Experiment 1 of VALUE, which are also used here. The different colors identify the 8 PRUDENCE regions covering Europe. The 9 illustrative stations selected for the some of the experiments presented in this work are marked with the symbol × (see Table 1 for details). This figure has been adapted from Figure 1 in Gutiérrez et al. (2019).

approach, one can instead merge all of them into a common set $\bigcup_{k=1\ldots K}\{y \in Y\}_{t_k(x)}$, from which the parameters of the predicted probability distribution—in this work α and β—can be estimated. Formula 2 thus becomes

$$\hat{y}(x) = \Theta\left(\bigcup_{k=1}^{K}\{y \in Y\}_{t_k(x)}\right), \tag{3}$$

where Θ represents the particular procedure followed to estimate α and β, which is carried out as a posterior step once the trees have been built. We call this approach a posteriori (AP-RF) random forests. Since the number of samples available to estimate α and β increases notably, the choice of estimation method is much less limited, and more costly or biased procedures such as MLE become feasible via a numeric method, provided they are consistent. Throughout this work, and unless otherwise specified, we use the BC3 estimators defined in Section 4 as the estimation procedure Θ when using AP-RFs.

Therefore, the AP-RFs produce a gamma probability distribution $Y \mid X = x$ for each large-scale predictors' state $X = x$, thus fully characterizing the uncertainty of the predictions. From this *predicted distribution* we can compute its expected value α/β to provide a single deterministic prediction, or we can instead draw random samples to produce predictive simulations, which can in turn be used to validate the reliability of the downscaled probability distributions. Figure 2 illustrates the AP-RF approach as compared to the traditional AVG-RF approach.

Note that we do not use bootstrap samples to train each tree for the AP-RFs presented in this work. Instead, we draw samples without replacement with 0.632 times the number of instances in the original data set, since it can be shown that ~36.8% is the expected percentage of duplicated instances that are found in a sample drawn

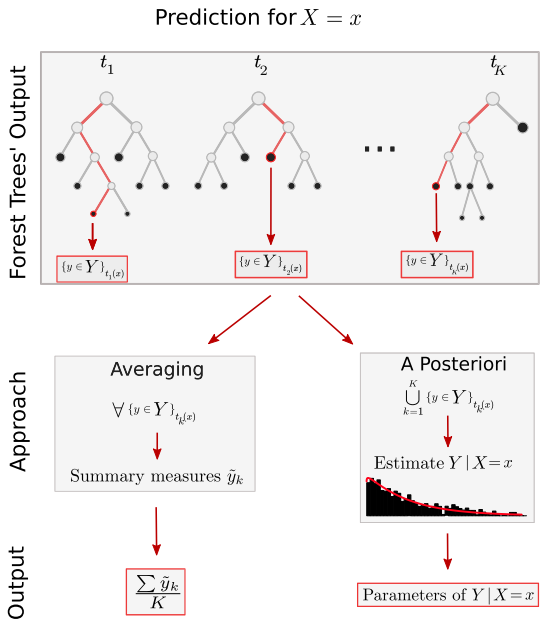

Figure 2. Chart illustrating the a posteriori approach (AP-RF) introduced in this work for random forests versus traditional averaging approach (AVG-RF).

with replacement. This avoids unnecessary repetitions in the training data that could potentially affect the estimation of α and β, while also reducing the computational cost.

3.3. Generalized Linear Models

We consider in this work GLMs for benchmarking purposes. GLMs (Chandler, 2005) are a generalization of linear regression which allow for modeling non-normally distributed variables, providing an estimate $\hat{y}$ for $E(Y \mid X = x)$ using a non-linear link function g as $\hat{y} = g^{-1}(Bx)$, where B are the linear coefficients.

They have been extensively used for statistical downscaling and rainfall modeling at different time-scales (Abaurrea & Asín, 2005; Chandler & Wheater, 2002; Manzanas et al., 2015; Manzanas, Gutiérrez, et al., 2020; Manzanas, Fiwa, et al., 2020; Nikulin et al., 2018; San-Martín et al., 2017) and provide competitive results compared to other machine learning models (Gutiérrez et al., 2019). In addition, they can also be used to obtain both deterministic and stochastic predictions, which makes them ideal as baseline model.

The two configurations used here correspond to GLM-DET and GLM methods used in Gutiérrez et al. (2019) (rows 38 and 39 in Table 3, respectively) to provide deterministic and stochastic predictions of precipitation amounts. Both methods use the gamma distribution as error function and the logarithm as link function. For the sake of comparability with AP-RFs, the two GLMs applied in this work consider the actual predictor fields at the 4 nearest grid boxes instead of the 20 leading principal components used in Gutiérrez et al. (2019). Moreover, we have verified for the case study presented in Section 4.4 that the use of local predictor information yields overall slightly better results than the 20 leading principal components used in Gutiérrez et al. (2019) (not shown here for brevity).

In particular, we use in this work standardized (mean 0 and standard deviation 1) predictor fields for the GLMs considered. Standardization is made over time at each gridbox separately. Importantly, GLMs do not incorporate any predictor information that explicitly accounts for seasonality, low troposphere's stability and/or any interactions (neither do the RFs), which could doubtless improve the performance of the predictions. Nevertheless, obtaining the best possible GLM predictions is out of the scope of the present work, which is aimed at introducing the novel AP-RFs and providing a first comparison with a simple implementation of the well-known GLMs. In this regard, note that, whilst GLMs in their standard form can only relate the predictors to one of the two parameters of the gamma distribution—the other is assumed to be fixed—AP-RFs allow to simultaneously predict both, which constitutes a key advantage. Indeed, as we will show later, this allows for generating realistic stochastic predictions, including high precipitation values.

4. Results

We present in this section a thorough analysis which allows us to better understand the performance of RFs and to assess their suitability for downscaling of precipitation on wet days. Section 4.1 is devoted to evaluate the influence that the choice of split function can have on the predictions, a relevant issue which has not been deeply studied before. In Section 4.2, we demonstrate that AP-RFs outperform traditional AVG-RFs in capturing $Y \mid X$. Section 4.3 shows that AP-RFs and AVG-RFs have the same predictive power and deals with the optimization process of the RFs, focusing on two key parameters: the number of trees and the minimum leaf size. To conclude, Section 4.4 shows that the AP-RF model produces reliable predictions and overall better results than the implementation of the GLM technique used in the Experiment 1 of VALUE in terms of distributional similarity. Note that Sections 4.1, 4.2 and 4.3 focus on the 9 stations listed in Table 1, whereas in Section 4.4 we consider the complete set of 86 stations from the Experiment 1 of VALUE.

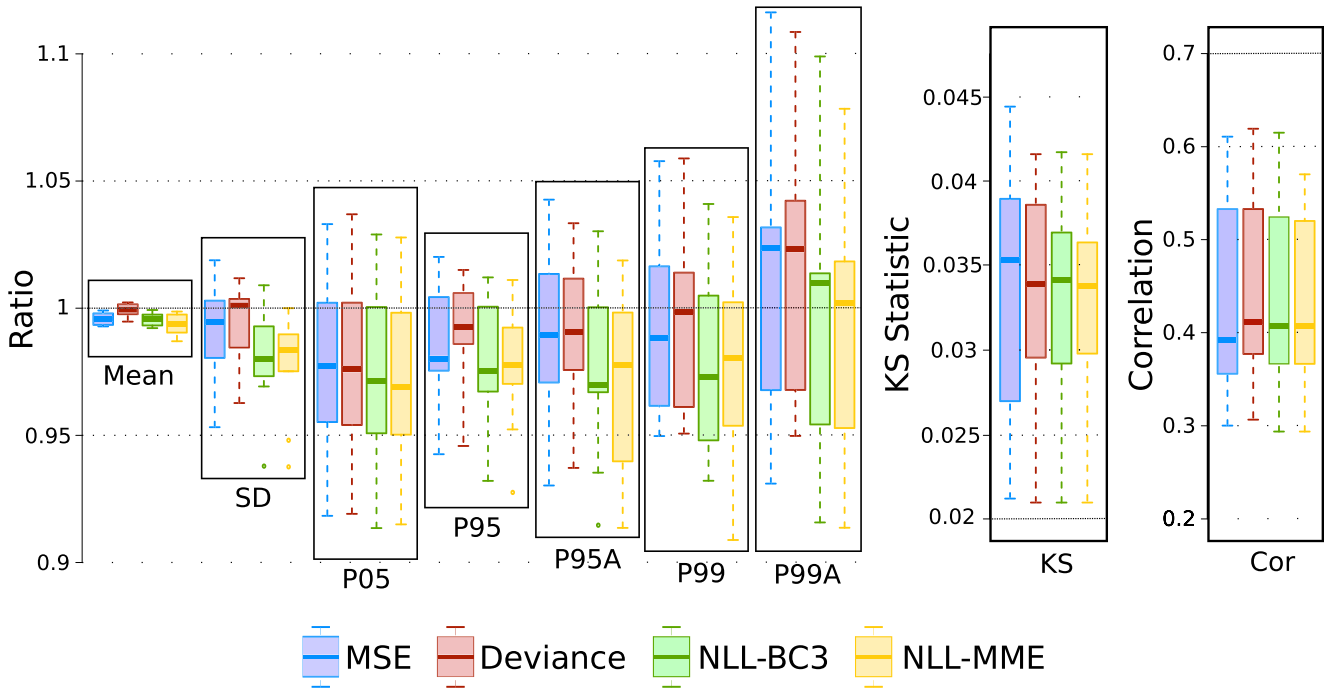

Figure 3. Cross-validated results obtained along the 9 stations of Table 1 when the different split functions of Table 4 are used, in terms of different metrics. In all cases, an AP-RF with 200 trees and minimum leaf size 5 was used, with BC3 estimators as estimation function Θ. Except for correlation, 250 stochastic simulations were first produced, leading to 250 values for each metric which were finally averaged. For the case of correlation, in order to comprehensively assess the predictive power of the different techniques, deterministic predictions (obtained as the expected value α/β of the modeled distributions) are employed.

All the results shown hereafter correspond to out-of-sample predictions/simulations, that is, to the complete time-series (covering the period 1979–2008) obtained according to the 5-fold cross-validation scheme explained in Section 2. Note that, except correlation, all measures in Table 3 are computed for stochastic simulations drawn from the predicted distribution, and thus the validation metrics shown correspond to the mean value obtained from 250 realizations, allowing us to assess the performance of the model for simulating downscaled rainfall. For the case of correlation, in order to comprehensively assess the predictive power of the different techniques, deterministic predictions (obtained as the expected value α/β of the modeled distributions) are employed.

4.1. Sensitivity to the Split Function in a Posteriori Random Forests

As outlined in Section 3, a relevant question when working with AP-RFs is how the choice of split function may affect the results. In AP-RFs the split function is used to iteratively divide the predictor space in order to predict $Y \mid X$ and, even though the estimation procedure is carried out after the trees have been built, the split function should ideally take into account the probability distribution of Y.

In order to assess the effect of the split function on AP-RFs, we put to the test the four alternatives listed in Table 4, considering BC3 estimators as the estimation procedure Θ in all cases. The boxplots in Figure 3 show the cross-validated results obtained with 200 trees and minimum leaf size 5 (similar conclusions were found for other configurations of the forest) in terms of the five metrics shown in Table 3 along the 9 stations of Table 1.

Although the differences are subtle, the gamma deviance outperforms the rest of split functions. This can be seen both for the distributional measures and for correlation. MSE performs slightly worse than the deviance, which gives place to less centered boxplots with larger spread. Interestingly, for the case of the negative log-likelihood, no clear improvement is found when using the BC3 estimators (with respect to the more biased MME estimators). Nevertheless, the negative log-likelihood (both with MME and BC3 estimators) exhibits generally worse performance than both MSE and deviance. Therefore, unless stated otherwise, the gamma deviance will be the split function used hereafter for yielding overall the best results.

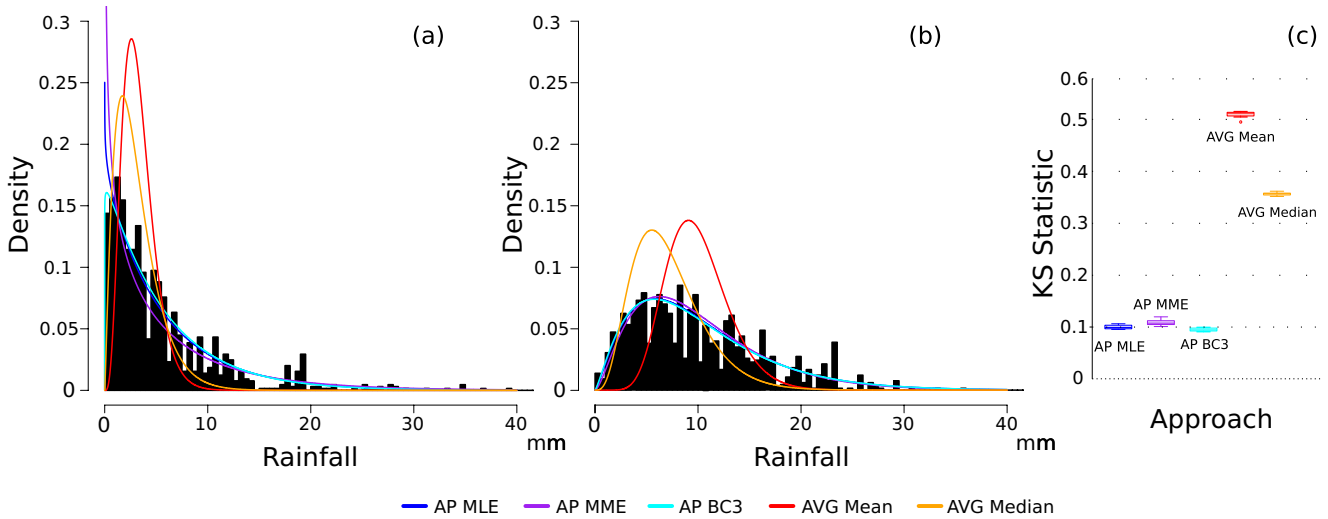

Figure 4. Panels (a and b) show, for two illustrative predictions (i.e., days), the predictive observations $\bigcup_{k=1}^{K}\{y \in Y\}_{t_k(x)}$ (black histograms) along with the predicted gamma distributions, as obtained according to different implementations of the AVG-RF and AP-RF approaches (colored curves). In the case of the AP-RF approach, only the estimator used varies. The boxplots in panel (c), which encompass the 9 stations described in Table 1, assess the performance of each of the alternatives in terms of distributional consistency, by showing the mean Kolmogorov-Smirnov statistic (averaged over all days) between the cumulative distribution function of the predictive observations and predicted gamma distributions.

4.2. Distributional Performance: Averaging Versus a Posteriori

This section illustrates how the AP-RF approach introduced in this work (see Section 3.2), as compared with traditional AVG-RF, much better captures the probability distribution of $Y \mid X$. To do this, panels (a and b) in Figure 4 show, for two illustrative predictions, the predictive observations $\{y \in Y\}_{t_k(x)}$ (black histograms) along with the predicted gamma distributions obtained according to different implementations of the AVG-RF and AP-RF approaches (colored curves).

For the AVG-RF approach, in addition to the mean (in red), we have also considered the median (in orange) as averaging function to see if it yields less biased results. For the case of the AP approach, three different options, which only vary in the type of estimator Θ used—MLE (blue), MME (purple) and BC3 (cyan)—are assessed. Note that, for the sake of consistency, the same RF with 200 trees and minimum leaf size of 5 is used in all cases, although similar results are obtained for other configurations. Moreover, for a direct comparison between AVG-RF and AP-RF approaches, the NLL-BC3 split function is used in all cases.

For completeness, the boxplots in panel (c) allow us to assess the similarity between the predictive observations and predicted distributions obtained along the 9 stations described in Table 1. To do this, we rely on the Kolmogorov-Smirnov (KS) statistic (see Table 3), which measures the maximum distance between two cumulative distribution functions and, therefore, the lower the statistic, the better. For each station, all KS statistics (one per day) have been averaged.

Our results show a clear difference between AVG-RF and AP-RF approaches, with the latter exhibiting a much better ability to capture the whole distribution of predictive observations, regardless of the estimator used. Indeed, the KS statistic is around 0.1 for all AP-RFs—with BC3 estimators yielding the best results by a small margin—whilst it increases to approximately 0.35 (0.5) when the median (mean) is used in traditional AVG-RFs.

4.3. Sensitivity to the Model Complexity

As explained in Section 3.1, two main parameters control the overall complexity of a RF model: the number of trees and the minimum leaf size. Figure 5 assesses their relative effect for the case of deterministic and stochastic RFs (panels a and b, respectively). In particular, panel (a) shows the results obtained, in terms of daily correlation with observations for deterministic predictions, for AVG-RFs and AP-RFs (topleft and bottomright triangle,

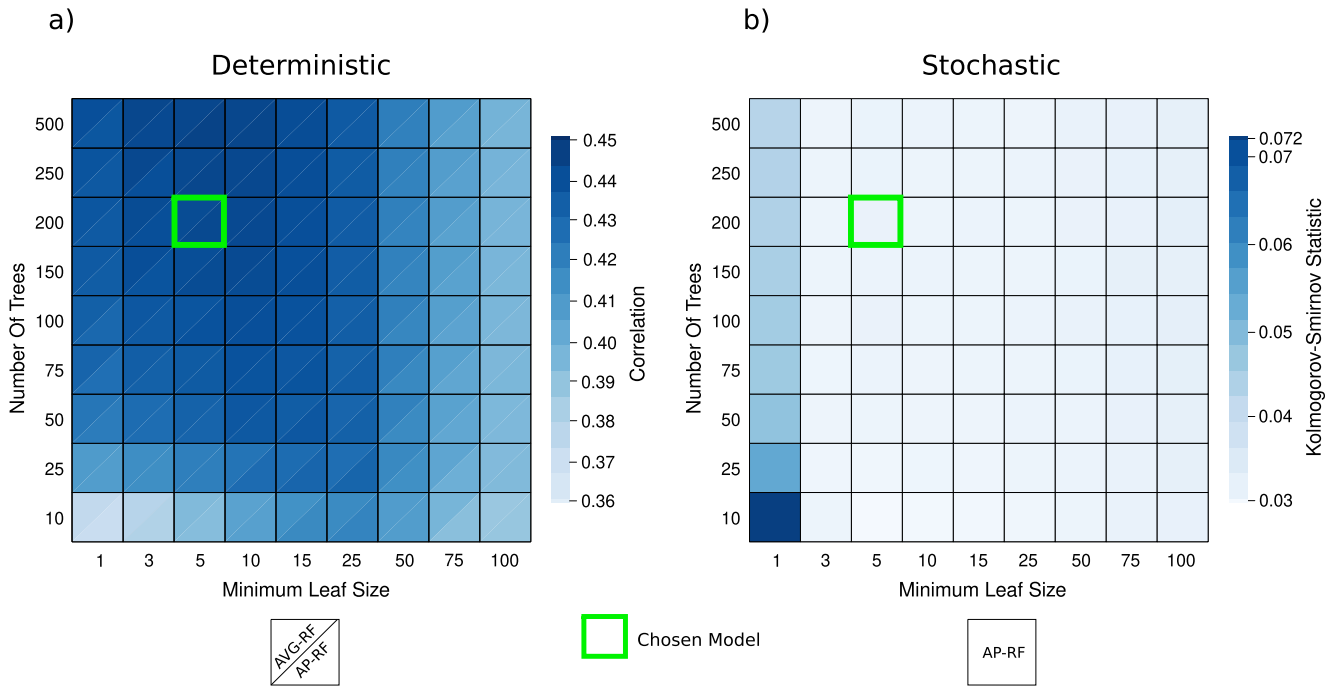

Figure 5. Cross-validated results averaged over the 9 stations listed in Table 1, as a function of the *minimum leaf size* and *number of trees* parameters (axis x and y, respectively). Panel (a) corresponds to deterministic predictions as obtained by averaging (topleft triangle, AVG-RF) and by considering the expected value of the predicted distribution using the AP-RF approach with BC3 estimators (bottomright triangle). In order to assess the distributional performance, panel (b) shows the mean Kolmogorov-Smirnov statistic obtained from 250 stochastic simulations drawn from the distributions predicted by the AP-RFs. In all cases, the RFs were built using the gamma deviance as split function. The model configuration leading to the best correlation (on average for the 9 stations) is highlighted in green in both panels. Note that, although it is not shown, there is some variability across stations, with correlations differences of up to 0.3 units between the best and the worst performing case (in general, better results are found for stations with larger sample sizes available for training).

respectively). For the latter, predictions were obtained as the expected value of the predicted gamma probability distribution, that is, α/β. Panel (b) corresponds to the mean Kolmogorov-Smirnov statistic (KS, see Table 3) value of 250 stochastic realizations following the AP approach using BC3 estimators. In all cases, the RFs were built using the gamma deviance as split function because of its previously shown good performance.

In general, the best predictive performance in terms of correlation is obtained for minimum leaf sizes ranging between 3 and 15. When the minimum leaf size is set to 1, RFs suffer from overfitting (which is clearly noticeable in terms of distributional performance in panel b), although this can be partially alleviated by adding more trees. For minimum leaf sizes greater than 15 the models start lacking predictive power (underfitting). In terms of distributional performance, as measured by the Kolmogorov-Smirnov statistic, all configurations with minimum leaf size greater than 1 perform very similarly.

As per the number of trees, the models with small minimum leaf sizes tend to improve their overall performance by adding more trees. Indeed, this is particularly noticeable for minimum leaf size of 1, both in terms of predictive and distributional performance. For minimum leaf sizes greater than 5, all configurations tend to perform similarly for models with 25 or more trees.

The best results in terms of predictive performance, as measured by correlation, are found for the particular case of 200 trees and minimum leaf size of 5. This configuration, which is marked in green in both panels of Figure 5, will be used in the next section.

Importantly, Panel (a) in Figure 5 reveals that traditional AVG-RF and the AP-RF approaches lead to the same results in terms of correlation with the observed precipitation amounts. This indicates that the predictive power of our AP-RFs is maintained with respect to that of traditional AVG-RFs, with the advantage that AP-RFs allow us to accurately estimate the parameters of $Y \mid X$ (see Figure 4 and the next section).

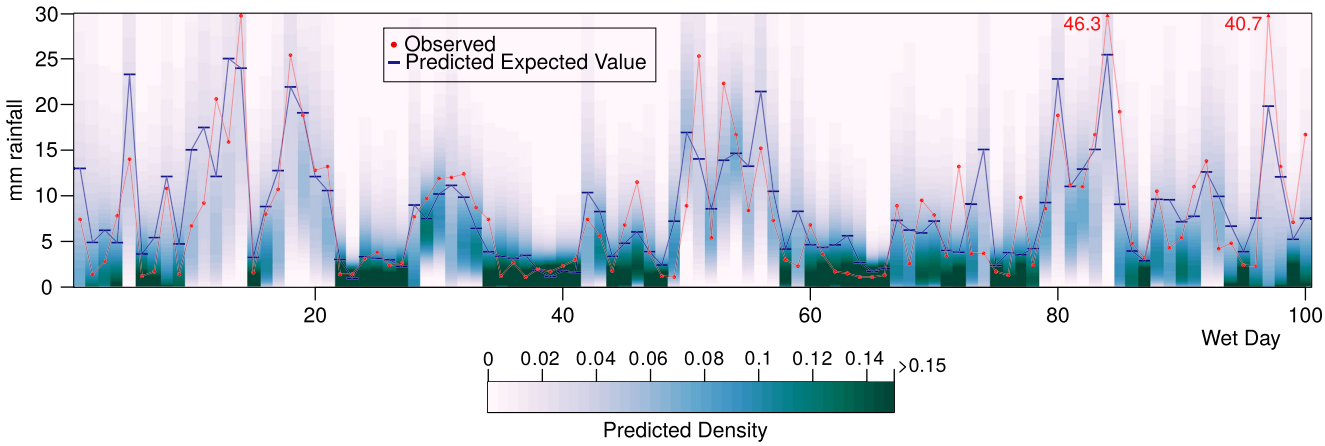

Figure 6. Results obtained for our AP-RFs at station 6 (Brocken, see Table 1), for an illustrative time-series of 100 chronologically ordered wet days. The daily observed precipitation values are indicated with a red dot—note that values above 30 mm are given with numbers—whereas the density of the corresponding (daily) predicted gamma distribution is shown in a white-to-green color scale (which is limited to 0.15 for better visualization). In addition, the (daily) predicted expected values, α/β, are marked in blue.

4.4. Extending the VALUE Experiment 1 With a Posteriori Random Forests

Finally, we compare in this section the performance of AP-RFs and GLMs for the 86 stations considered in the Experiment 1 of VALUE. To do so, we have employed the best configuration found in the previous sections, which consists of 200 trees and minimum leaf size of 5, and considers the gamma deviance as split function and the AP-RF approach with BC3 estimators to estimate the shape and rate parameters of the predicted gamma distribution.

For illustrative purposes, Figure 6 shows the density of the daily gamma distribution predicted by the AP-RF (white-to-green color scale) for a time-series of 100 chronologically ordered wet days at station 6 in Table 1. The corresponding observed precipitation values are marked with red dots, and the predicted expected values (computed as α/β) are given in blue.

This figure illustrates the degree of uncertainty involved in the daily predictions and thus the importance of having fully characterized $Y \mid X$, which allows us, for instance, to calculate the probability of a certain extreme value occurring for a given day. As an example, the probability of receiving more than 40 mm of rainfall for day 97 is 0.10 (notice how the observed value was 40.7 mm and the predicted expected value was only 17.05 mm). For comparison, the probability of getting more than 40 mm of precipitation for day 22 is $0 \approx 10^{-15}$ (the observed precipitation was 1.4 mm). Actually, for this particular day, the probability of having more than 3 mm of rain is 0.0856, thus the uncertainty of this prediction is comparatively low.

Figure 7 shows the cross-validated results for 250 stochastic simulations obtained from the benchmark GLM (blue) and the newly proposed AP-RF (red), in terms of all metrics listed in Table 3. For each metric, the corresponding boxplots encompass the 86 stations from the Experiment 1 of VALUE (see Section 2).

Whilst performing similarly for correlation (the difference in the median correlation is smaller than 0.01), our AP-RF reliably reproduces the observed distributions, providing better results than the benchmark GLM for all metrics in the left panel. Although both models accurately represent the mean precipitation over the considered period, a clear improvement is encountered in the rest of distributional measures, not only in terms of the standard deviation (SD), but also for the tail (P95, P95A, P99, and P99A) and low precipitation values (P05). These results are summarized by the difference in performance as measured by the Kolmogorov-Smirnov statistic (KS).

For the sake of comprehensiveness, the comparison is also presented season by season in Figure 8. Note that neither the GLM nor the AP-RF are trained separately for each season and, thus, the results shown in Figure 8 should not be expected to be as good as those displayed in Figure 7. In general, our AP-RF performs better than the benchmark GLM, particularly for the SD and for the percentiles (especially in DJF). Note that, for both methods, the lowest correlations are found for JJA. Presumably, this may be due to the smaller percentage of wet days in this season, which reduces the sample size available for training. Indeed, for the most extreme case (Málaga,

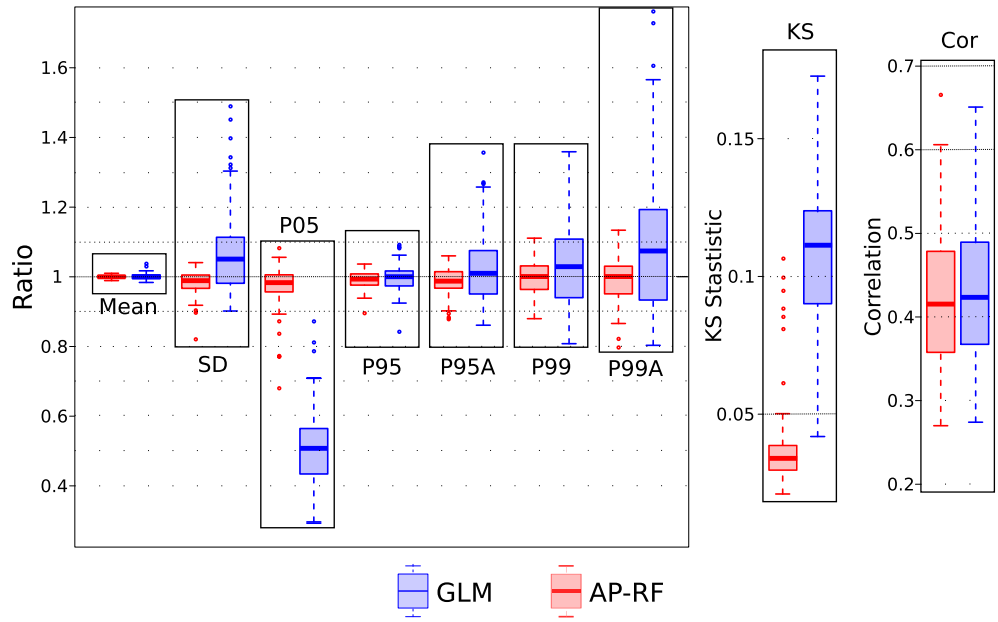

Figure 7. Cross-validated results obtained for the stochastic versions of our optimized AP-RF (see the text for details) and the benchmark GLM. The values shown in the boxplots in the left panel are the ratio between the simulated and observed metrics. The rightmost panel shows the Spearman correlation between observed and predicted wet-day sequences. All the boxplots encompass the 86 stations from the Experiment 1 of VALUE. Except for correlation, the results shown here correspond to the average of 250 metrics which were derived from simulations drawn from the predicted distributions. For the case of correlation, in order to comprehensively assess the predictive power of the different techniques, deterministic predictions (obtained as the expected value α/β of the modeled distributions) are employed.

Iberian Peninsula), there is only 1.66% of wet days in JJA (182 rainfall observations), whereas this proportion increases to 17.57% (1925 rainfall observations) in DJF. A similar situation occurs for some of the Mediterranean stations.

5. Conclusions

This work presents a thorough analysis about the suitability of random forests (RFs) for statistical downscaling of rainfall on wet days, a task for which this popular machine learning technique has been rarely used to date. To do this, we put to the test several configurations that RFs can potentially admit but had not been taken into account before in a climate prediction context. In particular, beyond analyzing the effect that the complexity of the forest (i.e., the number of individual trees and the minimum number of elements that each leaf is required to have) can have in the final results, we test the appropriateness of different split functions and introduce a novel a posteriori (AP-RF) approach, in which the data in the leaves are used to estimate the shape and rate parameters of the underlying gamma distribution. As opposed to other traditional, well-established techniques in the climate community such as GLMs which, in their standard form, can only relate the predictors to one of the two parameters (assuming the other is kept fixed), this is a key feature of the methodology presented here, since it allows us to fully characterize the uncertainty of our downscaled predictions.

Notably, the AP-RF approach here presented can be trivially adapted to any probability distribution, thus making the use of RFs extensible to other variables which may not be necessarily gamma-distributed such as temperature, winds, etc. For this reason, a relevant issue is the choice of the split function for each candidate distribution to model $Y \mid X$. Although in the case of rainfall the effects are somewhat subtle, the results obtained in this work hint that special care has to be put into choosing a split function that suits the probability distribution used to model Y, something that has not received enough attention. For the particular case of precipitation amount in wet days (which is assumed to follow a 2-parameter gamma distribution in this work), we have seen that the gamma deviance—which accounts for the shape of the gamma distribution but is still an approximation that assumes

10.1029/2021WR030272

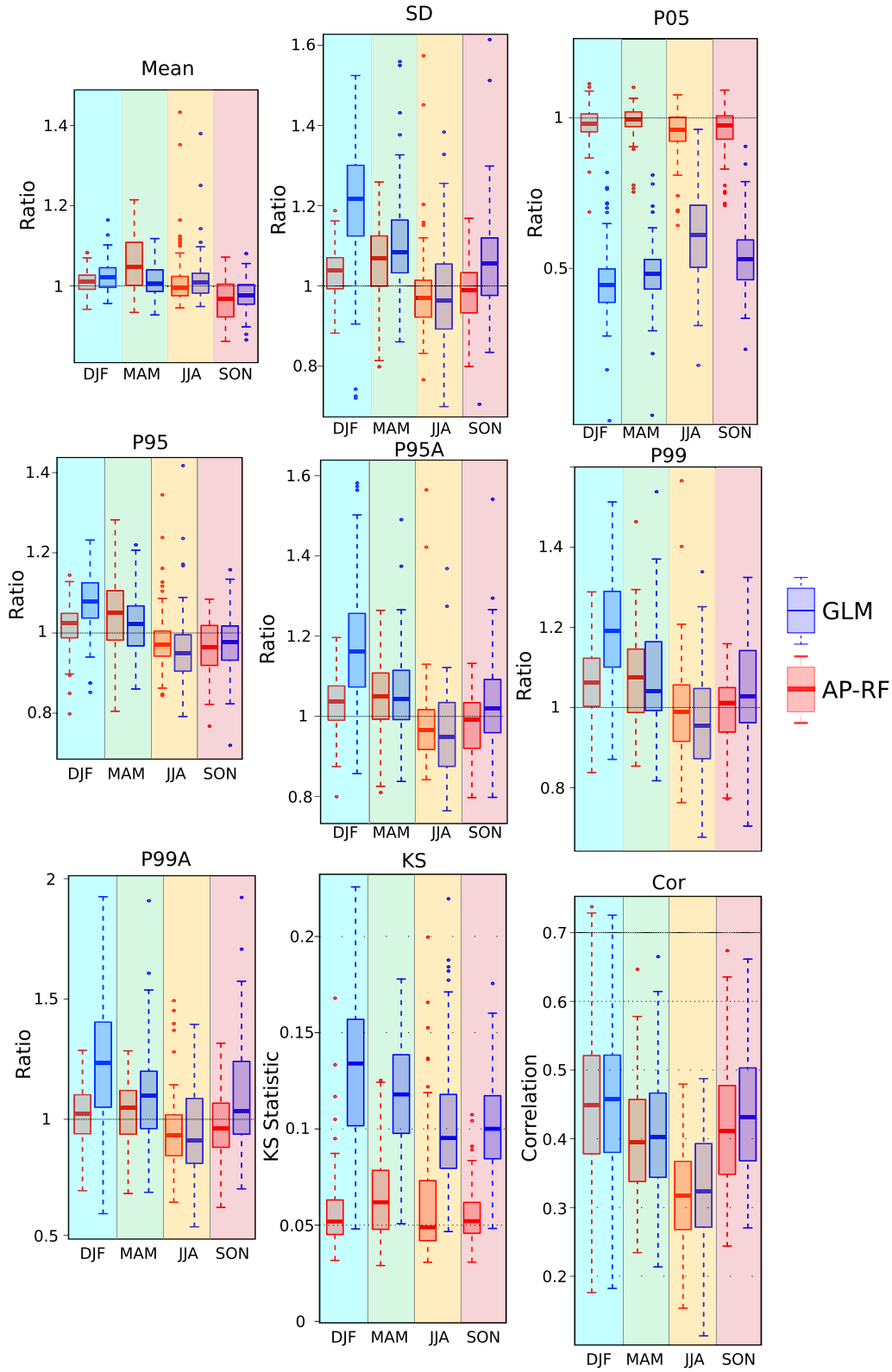

Figure 8. As Figure 7 but for individual boreal seasons: winter (DJF, cyan) spring (MAM, green), summer (JJA, orange) and autumn (SON, red). Note that neither the AP-RF nor the GLM have been separately trained for each season.

one of the parameters to be fixed—provides better results than other common options such as MSE and the log-likelihood.

Since the uncertainty of the predictions in downscaling of precipitation is high, sensibly capturing the whole distribution of downscaled rainfall becomes essential. The ability of the AP-RFs introduced in this work to reliably characterize the entire rainfall distribution (instead of just providing its expected value) without a significant loss of predictive performance with respect to traditional random forests is key for practical applications, since they allow for a much better representation of the uncertainty and variability of each prediction. This, in turn, makes AP-RFs ideal to generate realistic ensembles of stochastic predictions.

Indeed, as compared with a benchmark GLM which was shown to yield overall good results in the Experiment 1 of the COST action VALUE, our AR-RFs provide in general better performance in terms of distributional consistency with observations. This is particularly evident for the standard deviation and modeling the tail of the rainfall distribution, as measured by indicators such as the 99th percentile, which may have important implications for hydrology, crop modeling and other human activities which need reliable rainfall information as input.

In terms of computational costs, AP-RFs are generally more expensive than GLMs, although this depends on the complexity of the forest (the number of trees considered and the depth of each individual tree have a high impact on the time required to train the model). However, our results indicate that AP-RFs do not require a large number of trees, likely due to the fact that they use all the observations falling in the leaves of all trees instead of just a summary measure from each tree. For the datasets considered in this paper, an AP-RF with 200 trees and a minimum leaf size of 5 which considers the gamma deviance as split function can be trained in around 1–5 min (depending on the location) in an Intel i7-8700 processor, making it feasible for personal computers. Note that trees can be built in parallel, which further speeds up computation.

In this regard, we have created an R package called *RandomForestDist* which allows the user to apply all the split functions and estimation approaches here described (plus others). This package, which builds on a modified version of the well-known *rpart* (Therneau & Atkinson, 2019), is publicly available at https://github.com/MNLR/RandomForestDist. For the sake of reproducibility, it includes a *Jupyter* notebook (https://github.com/MNLR/RandomForestDist/blob/master/WorkedExample.ipynb) with worked examples which illustrate how to train/apply the RFs presented in this paper over the 9 stations shown in Table 1. Note that this notebook can be easily modified to cover the particular needs of the interested readers.

Finally, we foresee to extend the present work in various ways. First, we are currently developing a single AP-RF which allows to simultaneously model both the binary and the continuous aspect of precipitation through a mixture of Bernoulli and gamma probability distributions. Second, in order to improve the spatial consistency of our results, we will analyze the potential of AP-RFs for multisite downscaling (i.e., the same model is used to predict at several locations at the same time). Third, we will also test the suitability of AP-RFs to downscale other meteorological variables, such as temperature or winds.

Data Availability Statement

All the data used in this work, both ECA&D observed rainfall and ERA-Interim predictors, are publicly available and can be downloaded from http://www.value-cost.eu/data. The R package *downscaleR* (Bedia et al., 2020, https://github.com/SantanderMetGroup/downscaleR), which is part of the *climate4R* bundle (Iturbide et al., 2019, https://github.com/SantanderMetGroup/climate4R), was used to train and apply the GLM considered as benchmark model.

Acknowledgments
The authors would like to acknowledge the projects MULTI-SDM (CGL2015-66583-R, MINECO/FEDER); IS-ENES3, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement ID 824084 (https://is.enes.org) and Contribución a la nueva generación de proyecciones climáticas regionales de CORDEX mediante técnicas dinámicas y estadísticas (CORDyS): PID2020-116595RB-I00, funded by the Agencia Estatal de Investigación of the Spanish Government.

References

Abaurrea, J., & Asín, J. (2005). Forecasting local daily precipitation patterns in a climate change scenario. *Climate Research*, *28*(3), 183–197. https://doi.org/10.3354/cr028183

Acharya, N., Frei, A., Chen, J., DeCristofaro, L., & Owens, E. M. (2017). Evaluating stochastic precipitation generators for climate change impact studies of New York City's primary water supply. *Journal of Hydrometeorology*, *18*(3), 879–896. https://doi.org/10.1175/JHM-D-16-0169.1

Ailliot, P., Allard, D., Monbet, V., & Naveau, P. (2015). Stochastic weather generators: An overview of weather type models. *Journal de la Société Française de Statistique*, *156*(1), 101–113. http://www.numdam.org/

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124. https://doi.org/10.5194/gmd-13-2109-2020

Bedia, J., Baño Medina, J., Legasa, M. N., Iturbide, M., Manzanas, R., Herrera, S., et al. (2020). Statistical downscaling with the downscale R package (v3.1.0): Contribution to the VALUE intercomparison experiment. *Geoscientific Model Development*, *13*(3), 1711–1735. https://doi.org/10.5194/gmd-13-1711-2020

Booij, M. (2005). Impact of climate change on river flooding assessed with different spatial model resolutions. *Journal of Hydrology*, *303*(1), 176–198. https://doi.org/10.1016/j.jhydrol.2004.07.013

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bürger, G., & Chen, Y. (2005). Regression-based downscaling of spatial variability for hydrologic applications. *Journal of Hydrology*, *311*(1), 299–317. https://doi.org/10.1016/j.jhydrol.2005.01.025

Chan, J. C.-W., & Paelinckx, D. (2008). Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999–3011. https://doi.org/10.1016/j.rse.2008.02.011

Chandler, R. E. (2005). On the use of generalized linear models for interpreting climate variability. *Environmetrics*, *16*(7), 699–715. https://doi.org/10.1002/env.731

Chandler, R. E. (2020). Multisite, multivariate weather generation based on generalised linear models. *Environmental Modelling & Software*, *134*, 104867. https://doi.org/10.1016/j.envsoft.2020.104867

Chandler, R. E., & Wheater, H. S. (2002). Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resources Research*, *38*(10), 1–10. https://doi.org/10.1029/2001WR000906

Charles, S. P., Bates, B. C., & Hughes, J. P. (1999). A spatiotemporal model for downscaling precipitation occurrence and amounts. *Journal of Geophysical Research*, *104*(D24), 31657–31669. https://doi.org/10.1029/1999JD900119

Choi, S. C., & Wette, R. (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, *11*(4), 683–690. https://doi.org/10.1080/00401706.1969.10490731

Christensen, J. H., & Christensen, O. B. (2007). A summary of the PRUDENCE model projections of changes in European climate by the end of this century. *Climatic Change*, *81*(1), 7–30. https://doi.org/10.1007/s10584-006-9210-7

Cox, D., & Hinkley, D. (1974). *Theoretical statistics* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b14832

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. L. (2013). Seasonal climate predictability and forecasting: Status and prospects. *WIREs Climate Change*, *4*(4), 245–268. https://doi.org/10.1002/wcc.217

Dunn, P. K., & Smyth, G. K. (2018). Chapter 1: Statistical models. In P. K. Dunn & G. K. Smyth (Eds.), *Generalized linear models with examples in R* (p. 1–30). Springer. https://doi.org/10.1007/978-1-4419-0118-7_1

Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, *52*(7), 5619–5633. https://doi.org/10.1002/2016WR019129

Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, *27*(12), 1547–1578. https://doi.org/10.1002/joc.1556

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., et al. (2019). An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *International Journal of Climatology*, *39*(9), 3750–3785. https://doi.org/10.1002/joc.5462

Gutiérrez, J. M., San-Martín, D., Brands, S., Manzanas, R., & Herrera, S. (2013). Reassessing statistical downscaling techniques for their robust application under climate change conditions. *Journal of Climate*, *26*(1), 171–188. https://doi.org/10.1175/JCLI-D-11-00687.1

Haberlandt, U., Belli, A., & Bárdossy, A. (2015). Statistical downscaling of precipitation using a stochastic rainfall model conditioned on circulation patterns—An evaluation of assumptions. *International Journal of Climatology*, *35*(3), 417–432. https://doi.org/10.1002/joc.3989

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L., & Goodess, C. M. (2006). Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology*, *26*(10), 1397–1415. https://doi.org/10.1002/joc.1318

He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research*, *52*(10), 8217–8237. https://doi.org/10.1002/2016WR019034

Hertig, E., & Jacobeit, J. (2008). Assessments of Mediterranean precipitation changes for the 21st century using statistical downscaling techniques. *International Journal of Climatology*, *28*(8), 1025–1045. https://doi.org/10.1002/joc.1597

Husak, G. J., Michaelsen, J., & Funk, C. (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, *27*(7), 935–944. https://doi.org/10.1002/joc.1441

Ibarra-Berastegi, G., Saénz, J., Ezcurra, A., Elías, A., Diaz Argandoña, J., & Errasti, I. (2011). Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression. *Hydrology and Earth System Sciences*, *15*(6), 1895–1907. https://doi.org/10.5194/hess-15-1895-2011

Iturbide, M., Bedia, J., Herrera, S., Baño-Medina, J., Fernández, J., Frías, M., et al. (2019). The R-based climate4R open framework for reproducible climate data access and post-processing. *Environmental Modelling & Software*, *111*, 42–54. https://doi.org/10.1016/j.envsoft.2018.09.009

Jing, W., Yang, Y., Yue, X., & Zhao, X. (2016). A comparison of different regression algorithms for downscaling monthly satellite-based precipitation over North China. *Remote Sensing*, *8*, 835. https://doi.org/10.3390/rs8100835

Langousis, A., & Kaleris, V. (2014). Statistical framework to simulate daily rainfall series conditional on upper-air predictor variables. *Water Resources Research*, *50*(5), 3907–3932. https://doi.org/10.1002/2013WR014936

Lorenz, E. N. (1969). Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, *262*, 636–646. https://doi.org/10.1175/1520-0469(1969)26<636:aparbn>2.0.co;2

Louzada, F., Ramos, P. L., & Ramos, E. (2019). A note on bias of closed-form estimators for the gamma distribution derived from likelihood equations. *The American Statistician*, *73*(2), 195–199. https://doi.org/10.1080/00031305.2018.1513376

Manzanas, R., Brands, S., San-Martín, D., Lucero, A., Limbo, C., & Gutiérrez, J. M. (2015). Statistical downscaling in the tropics can be sensitive to reanalysis choice: A case study for precipitation in the Philippines. *Journal of Climate*, *28*(10), 4171–4184. https://doi.org/10.1175/JCLI-D-14-00331.1

Manzanas, R., Fiwa, L., Vanya, C., Kanamaru, H., & Gutiérrez, J. M. (2020). Statistical downscaling or bias adjustment? A case study involving implausible climate change projections of precipitation in Malawi. *Climatic Change*, *162*(3), 1437–1453. https://doi.org/10.1007/s10584-020-02867-3

Manzanas, R., Gutiérrez, J. M., Bhend, J., Hemri, S., Doblas-Reyes, F. J., Penabad, E., & Brookshaw, A. (2020). Statistical adjustment, calibration and downscaling of seasonal forecasts: A case-study for Southeast Asia. *Climate Dynamics*, *54*(5), 2869–2882. https://doi.org/10.1007/s00382-020-05145-1

Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., et al. (2015). Value: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, *3*(1), 1–14. https://doi.org/10.1002/2014EF000259

Martínez-Villalobos, C., & Neelin, J. D. (2019). Why do precipitation intensities tend to follow gamma distributions? *Journal of the Atmospheric Sciences*, *76*(11), 3611–3631. https://doi.org/10.1175/JAS-D-18-0343.1

Mukundan, R., Acharya, N., Gelda, R. K., Frei, A., & Owens, E. M. (2019). Modeling streamflow sensitivity to climate change in New York City water supply streams using a stochastic weather generator. *Journal of Hydrology: Regional Studies*, *21*, 147–158. https://doi.org/10.1016/j.ejrh.2019.01.001

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Neves, G. L., Barbosa, M. A. G. A., Anjinho, P. D. S., Guimarães, T. T., Das Virgens Filho, J. S., & Mauad, F. F. (2020). Evaluation of the impacts of climate change on streamflow through hydrological simulation and under downscaling scenarios: Case study in a watershed in southeastern Brazil. *Environmental Monitoring and Assessment*, *192*(11), 707. https://doi.org/10.1007/s10661-020-08671-x

Nikulin, G., Asharaf, S., Magariño, M. M., Calmanti, S., Cardoso, R. M., Bhend, J., et al. (2018). Dynamical and statistical downscaling of a global seasonal hindcast in Eastern Africa. *Climate Services*, *9*, 72–85. https://doi.org/10.1016/j.cliser.2017.11.003

Nogueira, M., & Barros, A. (2015). Transient stochastic downscaling of quantitative precipitation estimates for hydrological applications. *Journal of Hydrology*, *529*, 1407–1421. https://doi.org/10.1016/j.jhydrol.2015.08.041

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, *17*(1), 182–190. https://doi.org/10.1029/WR017i001p00182

San-Martín, D., Manzanas, R., Brands, S., Herrera, S., & Gutiérrez, J. M. (2017). Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. *Journal of Climate*, *30*(1), 203–223. https://doi.org/10.1175/JCLI-D-16-0366.1

Sattari, M. T., Shaker Sureh, F., & Kahya, E. (2020). Monthly precipitation assessments in association with atmospheric circulation indices by using tree-based models. *Natural Hazards*, *102*(3), 1077–1094. https://doi.org/10.1007/s11069-020-03946-5

Sauter, T., & Venema, V. (2011). Natural three-dimensional predictor domains for statistical precipitation downscaling. *Journal of Climate*, *24*(23), 6132–6145. https://doi.org/10.1175/2011JCLI4155.1

Schlabing, D., Frassl, M., Eder, M., Rinke, K., & Bárdossy, A. (2014). Use of a weather generator for simulating climate change effects on ecosystems: A case study on Lake Constance. *Environmental Modelling & Software*, *61*, 326–338. https://doi.org/10.1016/j.envsoft.2014.06.028

Shi, Y., & Song, L. (2015). Spatial downscaling of monthly TRMM precipitation based on EVI and other geospatial variables over the Tibetan Plateau from 2001 to 2012. *Mountain Research and Development*, *35*(2), 180–194. https://doi.org/10.1659/MRD-JOURNAL-D-14-00119.1

Smith, P. L., Kliche, D. V., & Johnson, R. W. (2009). The bias and error in moment estimators for parameters of drop size distribution functions: Sampling from gamma distributions. *Journal of Applied Meteorology and Climatology*, *48*(10), 2118–2126. https://doi.org/10.1175/2009JAMC2114.1

Song, K.-S. (2008). Globally convergent algorithms for estimating generalized gamma distributions in fast signal and image processing. *IEEE Transactions on Image Processing*, *17*(8), 1233–1250. https://doi.org/10.1109/TIP.2008.926148

Therneau, T., & Atkinson, B. (2019). *Rpart: Recursive partitioning and regression trees (computer software manual)*. Retrieved from https://CRAN.R-project.org/package=rpart

Timbal, B., Dufour, A., & McAvaney, B. (2003). An estimate of future climate change for Western France using a statistical downscaling technique. *Climate Dynamics*, *20*(7–8), 807–823. https://doi.org/10.1007/s00382-002-0298-9

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G. A., & Torricelli, P. (2011). Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecological Modelling*, *222*(8), 1471–1478. https://doi.org/10.1016/j.ecolmodel.2011.02.007

Von Storch, H., Zorita, E., & Cubasch, U. (1993). Downscaling of global climate change estimates to regional scales: An application to iberian rainfall in wintertime. *Journal of Climate*, *6*(6), 1161–1171. https://doi.org/10.1175/1520-0442(1993)006<1161:dogcce>2.0.co;2

Wei, C.-L., Rowe, G. T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M. J., & Narayanaswamy, B. E. (2010). Global patterns and predictions of seafloor biomass using random forests. *PLoS One*, *5*, e15323. https://doi.org/10.1371/journal.pone.0015323

Wilks, D. S., & Eggleston, K. L. (1992). Estimating monthly and seasonal precipitation distributions using the 30- and 90-day outlooks. *Journal of Climate*, *5*(3), 252–259. https://doi.org/10.1175/1520-0442(1992)005<0252:emaspd>2.0.co;2

Xiang, B., Zeng, C., Dong, X., & Wang, J. (2020). The application of a decision tree and stochastic forest model in summer precipitation prediction in Chongqing. *Atmosphere*, *11*, 508. https://doi.org/10.3390/atmos11050508

Xu, R., Chen, N., Chen, Y., & Chen, Z. (2020). Downscaling and projection of Multi-CMIP5 precipitation using machine learning methods in the upper Han River Basin. *Advances in Meteorology*. https://doi.org/10.1155/2020/8680436

Ye, Z.-S., & Chen, N. (2017). Closed-form estimators for the gamma distribution derived from likelihood equations. *The American Statistician*, *71*(2). https://doi.org/10.1080/00031305.2016.1209129

Yu, X., Hyyppä, J., Vastaranta, M., Holopainen, M., & Viitala, R. (2011). Predicting individual tree attributes from airborne laser point clouds based on the random forests technique. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(1), 28–37. https://doi.org/10.1016/j.isprsjprs.2010.08.003

Zorita, E., & Von Storch, H. (1999). The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *Journal of Climate*, *12*(8), 2474–2489. https://doi.org/10.1175/1520-0442(1999)012<2474:tamaas>2.0.co;2