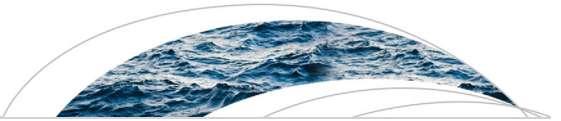
AGU PUBLICATIONS

Water Resources Research



RESEARCH ARTICLE

10.1002/2016WR019034

Spatial downscaling of precipitation using adaptable random forests

Xiaogang He[1], Nathaniel W. Chaney[1,2], Marc Schleiss[1], and Justin Sheffield[1,3]

[1]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey, USA, [2]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, New Jersey, USA, [3]Geography and Environment, University of Southampton, Southampton, UK

Key Points:
- A new machine-learning based algorithm (Prec-DWARF) for spatial precipitation downscaling using random forests (RF) is proposed
- Synthetic experiments are used to examine the performance of Prec-DWARF over different climatic regions in the United States
- Prec-DWARF with a double RF better reproduces observed precipitation structures and distributions

Correspondence to:
X. He,
hexg@princeton.edu

Citation:
He, X., N. W. Chaney, M. Schleiss, and J. Sheffield (2016), Spatial downscaling of precipitation using adaptable random forests, *Water Resour. Res.*, *52*, 8217–8237, doi:10.1002/2016WR019034.

Received 5 APR 2016
Accepted 28 SEP 2016
Accepted article online 3 OCT 2016
Published online 27 OCT 2016

Abstract This paper introduces Prec-DWARF (**Prec**ipitation **D**ownscaling **W**ith **A**daptable **R**andom **F**orests), a novel machine-learning based method for statistical downscaling of precipitation. Prec-DWARF sets up a nonlinear relationship between precipitation at fine resolution and covariates at coarse/fine resolution, based on the advanced binary tree method known as Random Forests (RF). In addition to a single RF, we also consider a more advanced implementation based on two independent RFs which yield better results for extreme precipitation. Hourly gauge-radar precipitation data at 0.125° from NLDAS-2 are used to conduct synthetic experiments with different spatial resolutions (0.25°, 0.5°, and 1°). Quantitative evaluation of these experiments demonstrates that Prec-DWARF consistently outperforms the baseline (i.e., bilinear interpolation in this case) and can reasonably reproduce the spatial and temporal patterns, occurrence and distribution of observed precipitation fields. However, Prec-DWARF with a single RF significantly underestimates precipitation extremes and often cannot correctly recover the fine-scale spatial structure, especially for the 1° experiments. Prec-DWARF with a double RF exhibits improvement in the simulation of extreme precipitation as well as its spatial and temporal structures, but variogram analyses show that the spatial and temporal variability of the downscaled fields are still strongly underestimated. Covariate importance analysis shows that the most important predictors for the downscaling are the coarse-scale precipitation values over adjacent grid cells as well as the distance to the closest dry grid cell (i.e., the dry drift). The encouraging results demonstrate the potential of Prec-DWARF and machine-learning based techniques in general for the statistical downscaling of precipitation.

1. Introduction

Understanding how climate variability and change impact and feedback with ecosystems and human activities is of great importance for natural hazards mitigation and risk management. The use of general circulation models (GCMs) facilitates our understanding of large-scale climate evolution and land surface-atmosphere interactions but the scale mismatch between coarse resolution GCMs and catchment hydrological processes increases the uncertainty of hydrological modeling [*Carter et al.*, 1994; *Hostetler*, 1994; *Fowler and Wilby*, 2007]. One possible way to address this issue is to spatially downscale coarse rainfall estimates. High-resolution products generated from coarse-scale precipitation fields (climate models, reanalysis, or satellite products) are more useful and appropriate for local impact assessments across different sectors. For instance, they can help stakeholders identify risk hotspots of natural hazards (rainfall triggered shallow landslides, flash floods, soil erosion, etc.) [*Hong et al.*, 2007; *He et al.*, 2016; *Zhang et al.*, 2004] and water-related pollution or water-borne disease [*Singh et al.*, 2015]. Precipitation downscaling also acts as an underpinning to investigate subgrid-scale parameterization [*Yano*, 2010], uncertainty propagation [*Bastola and Misra*, 2014], hyperresolution modeling [*Wood et al.*, 2011], and seasonal forecast [*Sheffield et al.*, 2014].

Despite the potential benefits of spatial downscaling of precipitation, its implementation is generally difficult and cumbersome due to the complex characteristics of precipitation (e.g., highly skewed, non-Gaussian distribution, intermittent and complex spatial-temporal structure). Techniques aimed to tackle this can be distinguished as dynamical or statistical downscaling approaches, and both have advantages and disadvantages. Dynamical downscaling relies on a regional climate or numerical weather model to provide high-resolution precipitation and other surface climate variables by simulating the physical processes of the coupled land-atmosphere system, at the expense of often large computational resources [*Rummukainen*, 2010].

In contrast, statistical downscaling methods aim to model the statistical relationships between small and large-scale covariates. They form an attractive alternative to dynamical downscaling methods due to their simplicity and low computational costs. Compared to their purely deterministic counterparts, stochastic models also allow for a clearer and more comprehensive interpretation of the predictable and random parts in a complex system, including crucial information about model uncertainty and probability of extremes [*Koutsoyiannis*, 2010; *Maraun et al*., 2010].

The main challenge of statistical downscaling is to make sure that the mean and temporal dependence of the downscaled process are consistent with the coarse-scale observations. When downscaling is performed at multiple sites, the spatial dependence also has to be modeled with suitable methods. In the case of precipitation, this can be achieved through the use of conceptual/physical models like random cascades [e.g., *Lovejoy and Mandelbrot*, 1985; *Lovejoy et al*., 1987; *Menabde et al*., 1997], weather typing [*Vrac and Naveau*, 2007], filtered autoregressive Gaussian processes [e.g., *Rebora et al*., 2006; *Schleiss and Berne*, 2012; *Jha et al*., 2015], or Poisson cluster models [e.g., *Hershenhorn and Woolhiser*, 1987; *Rodríguez-Iturbe et al*., 1987, 1988; *Cox and Isham*, 1988; *Onof and Wheater*, 1993; *Cowpertwait*, 1995; *Onof et al*., 2000; *Pegram and Clothier*, 2001]. Alternative purely statistical approaches in which physical/conceptual ideas only play a minor role have also been proposed. Examples include multiple linear regression (MLR) [*Jeong et al*., 2012], multivariate adaptive regression splines [*Beuchat et al*., 2011], QQ transforms [*Bárdossy and Pegram*, 2011], and machine learning.

Machine learning, also known as data mining or predictive analytics, is a very general and increasingly popular way to automatically extract information without the need to construct explicit physical or statistical models (e.g., neural networks) [*Olsson et al*., 2001; *Coulibaly et al*., 2005]. Compared to conventional approaches, machine-learning based methods are also easier to generalize beyond the training data samples [*Domingos*, 2012]. The growing popularity of machine learning stemmed from an explosion of big data, which has been regarded as the main driver of the next stage of innovation [*Manyika et al*., 2011]. In the hydrological sciences, sources of big data are from traditional gauge networks, large-scale simulations from GCMs or regional climate models, satellite and radar retrievals, paleoclimate proxies, and reanalysis products. At the same time, the strength of machine-learning algorithms draws from their ability to tackle different types of problems, from classification to prediction and parameter selection. In addition, increases in computer power (memory, storage capacity, and advanced parallel techniques), have made it feasible to use high performance computing (supercomputers) to combine knowledge (machine-learning based algorithms) with big data to develop automated algorithms for a variety of applications. This is particularly appropriate for precipitation downscaling, for which there is increasing demand for data at kilometer or finer spatial scales, such as for hyperresolution hydrological modeling [*Wood et al*., 2011; *Chaney et al*., 2016a].

Among the range of machine-learning algorithms, Random Forests (RF) [*Breiman*, 2001] stands out for its ability to deal with complex nonlinear relationships between variables while minimizing problems with overfitting. Due to its simplicity and capabilities, RF has been used in a wide range of hydrological-related applications, for example, high-resolution soil type classification over the contiguous United States [*Chaney et al*., 2016b], seasonal streamflow forecasting [*Zhao et al*., 2012; *He et al*., 2013], natural flow regime alternation [*Carlisle et al*., 2010], vegetation-type distribution [*Peters et al*., 2007], temperature [*Eccel et al*., 2007], and wind [*Davy et al*., 2010] downscaling, and satellite rainfall estimation from cloud physical properties [*Kühnlein et al*., 2014]. RF also has great potential for statistical precipitation downscaling although there are few studies that have addressed this issue. *Shi and Song* [2015] applied RF to downscale monthly TRMM precipitation by constructing a nonparametric relationship between precipitation and six covariates, including enhanced vegetation index, altitude, slope, aspect, latitude, and longitude. *Ibarra-Berastegi et al*. [2011] combined an analogue method and RF to downscale precipitation from reanalysis data sets. However, these methods did not consider the temporal and spatial dependence in their downscaling frameworks. Overall, the potential of RF and machine-learning techniques for statistical downscaling of precipitation has not been studied to its full extent and there appears to be much room left for further improvement.

In this paper, we present the development of a RF-based precipitation downscaling method called Prec-DWARF (**Prec**ipitation **D**ownscaling **W**ith **A**daptable **R**andom **F**orests) and demonstrate its use over the continental U.S. The method is adaptable because it can be applied in a variety of situations, to downscale data from models, remote sensing retrievals, or gridded observations at different resolutions. It can also be adapted to include any type of covariate (e.g., discrete or continuous) and can be adjusted to different types

of precipitation (e.g., stratiform, convective, and orographic). To our knowledge, this is the first time that RF is applied to downscale precipitation to high-resolution while taking into account the spatial and temporal dependence of the precipitation process. The structure of this paper is organized as follows: Section 2 introduces the data and study area. A comprehensive step-by-step description of the algorithm as well as the experimental design is presented in section 3. Case studies and results are given in section 4. A discussion and summary is provided in section 5.

2. Data and Study Regions

The North-American Land Data Assimilation System Project Phase 2 (NLDAS-2) [*Xia et al*., 2012] products are used in this study as the observational data and covariates for the RF. The spatial resolution of the data is 0.125° and the temporal resolution is hourly. Hourly precipitation data in NLDAS-2 are derived by temporally disaggregating daily rain gauge data based on hourly radar data, CMORPH products [*Joyce et al*., 2004], and CPC hourly CONUS/Mexico gauge data [*Higgins et al*., 1996], which are used to derive the hourly disaggregation weights and do not change the daily precipitation in total [*Cosgrove et al*., 2003]. Data for other near surface climate variables in the NLDAS-2 (e.g., temperature, wind speed, humidity, pressure) and surface short and longwave radiation are derived from the North American Regional Reanalysis (NARR) through the spatial interpolation, temporal disaggregation, and vertical adjustment. Topography-related covariates are derived from GTOPO30 Global 30 Arc Second (∼1 km) and regridded to 0.125°. Soil covariates in NLDAS-2 are reaggregated from the 1 km STATSGO data and only the most dominant soil texture class is selected. The first most predominant vegetation type from the University of Maryland's land cover classification products is chosen as the static vegetation covariate. As the purpose of this paper is to develop and verify the proposed algorithm rather than producing a complete data product, we only process the hourly data in the summer time (June, July, and August) of 2011 focusing on four climate divisions: Southwestern United States (SWUS), Central United States (CUS), Northeastern United States (NEUS), and Southeast United States (SEUS).

3. Methodology

In this section, we describe the algorithm for Prec-DWARF and the development of synthetic experiments to evaluate the algorithm.

3.1. Statistical Model and Physical Covariates

The basic idea of the downscaling algorithm is to set up a transfer function between the response and covariates using the following equations:

$$P_d = f(\mathbf{v}) + error \tag{1}$$

$$\mathbf{v} = (c_1, c_2, \ldots, c_N) \in \mathbb{R}^N \tag{2}$$

where P_d is the downscaled precipitation (response) at high-resolution, $\mathbf{v}$ is a multidimensional feature response vector, c_i represents the individual covariate (e.g., temperature), and N is the dimension of the input feature space ($N = 21$ in this study). f can be either a linear or nonlinear function depending on what assumptions are made. In machine learning, f indicates a black box model and may not have a specific form. One of the fundamental issues of machine learning is how to select the most informative covariates. In this study, precipitation at coarse spatial resolution is used as the main driver for downscaling. Physical predictors including atmospheric covariates and geographical covariates are also utilized. Atmospheric covariates (including coarse resolution precipitation) are included to maintain the temporal dynamics of the downscaled rainfall. Topographic data are included to maintain orographic effects [*Roe*, 2005] and better reproduce the spatial pattern of the downscaled precipitation. Auxiliary covariates at fine resolution, such as soil texture, vegetation type, are also included to take account for land-atmosphere coupling and its impact on the formation of precipitation [*Zhan et al*., 2015]. Latitude, longitude as well as the day of year (DOY) index are also used as covariates as they contain geographical and seasonality information.

As mentioned before, one of the main challenges in precipitation downscaling is to correctly reproduce its spatial structure. To ensure a realistic transition between dry grid cells and rainy grid cells as well as a better

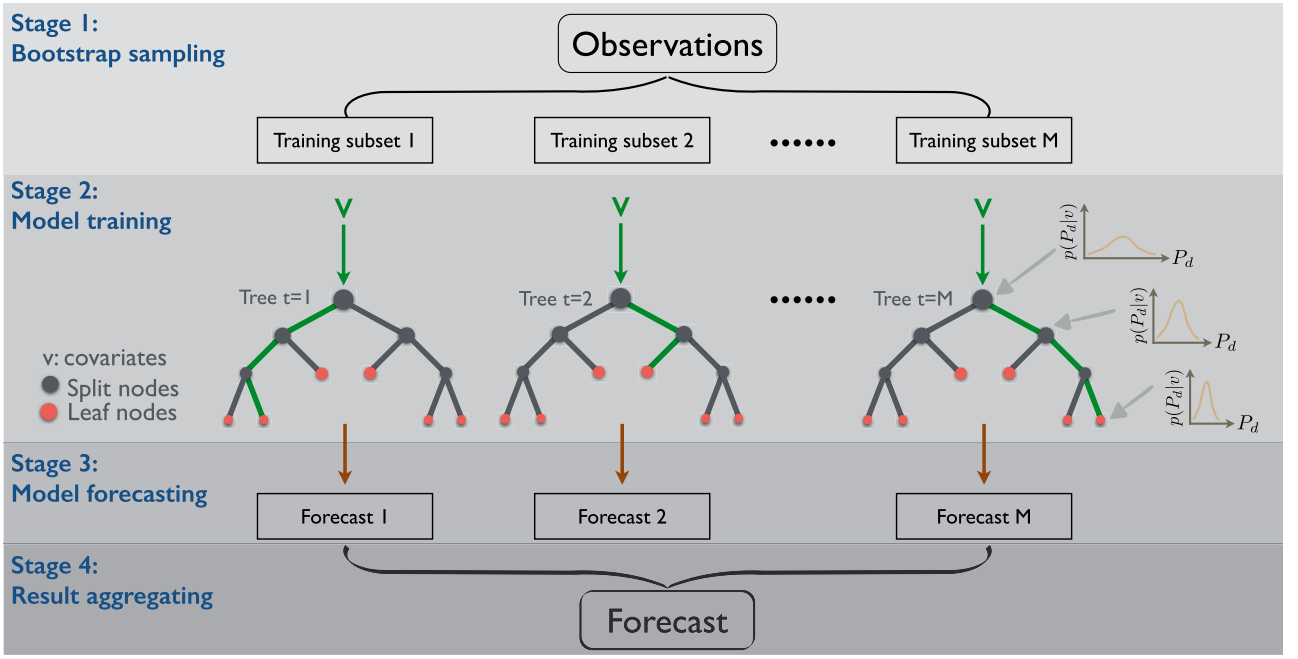


Figure 1. Schematic of the RF algorithm based on the Bagging (Bootstrap + Aggregating) method. (1) Stage 1: Use bootstrap method to sample *M* subsets from the original training data sets. (2) Stage 2: Build *M* independent decision trees for model training using input covariates (**v**). For each individual decision tree, the prediction confidence (posterior probability $p(P_d|\mathbf{v})$ increases from the root toward the leaves. (3) Stage 3: Obtain prediction from each bootstrap tree over *M* replications. (4) Stage 4: Decide the final result by average or majority voting.

distribution of rainfall rates inside the rainy regions, we adopt the "dry drift" formalism proposed by *Schleiss et al*. [2014]. Specifically, we incorporate an additional covariate into the downscaling scheme, which is the distance from any rainy pixel to the closest surrounding dry grid cell (in space). The distance to the closest dry pixel can be calculated efficiently using a binary search tree (e.g., the K-dimensional tree algorithm). Our results show that the inclusion of the dry drift leads to a more realistic spatial distribution of rain rates within the rainy grid cells. To avoid the so-called "border effect" [*Schleiss et al*., 2014], we first calculate the distance to the dry grid cells for the entire CONUS before subsetting it to the four considered regions.

3.2. Random Forests

Although the aforementioned regression-based method described by equations (1) and (2) is relatively easy to implement and does not require a substantial amount of computational resources, one disadvantage of regression-based methods is that they generally provide a single estimate and not a full probability density function (PDF). Sampling a full PDF has the advantage of providing an uncertainty analysis to evaluate the system performance (i.e., employ probabilistic descriptions of the model output using the confidence interval). A way to overcome this issue is to use ensemble learning, which aggregates results from multiple models in order to achieve better performance (greater accuracy and generalization) and reduce the chances of overfitting, whilst quantifying the uncertainty associated with a given estimate/prediction. Moreover, ensemble-based learning methods are relatively easy to train and test in a parallel computing environment. Among the ensemble learning methods, bootstrap [*Efron*, 1979] aggregating (bagging) [*Breiman*, 1996] is widely used, which can minimize variance and help avoid overfitting. Figure 1 is the schematic illustration of how bagging is applied to decision trees.

Random forests (RF) [*Breiman*, 1996, 2001] is an enhanced decision tree model, which is based on the bagging method to add an additional layer of randomness. As shown in Figure 1, in RF, the decision tree model acts as the individual forecast model. A decision tree model is a hierarchical analysis diagram composed of a collection of nodes and edges organized in a tree structure. There are two types of nodes in the decision tree model, split (internal) nodes and leaf (terminal) nodes. Each split node is associated with a test function, which is used to split the incoming data according to different attributes. Whereas, each leaf node corresponds to the final decision (forecast or classification label). Decision tree model is nonparametric and

therefore it is feasible to add either numeric or categorical data layers. What's more, it is also not sensitive to the outliers in the training stage [*Hautaniemi et al*., 2005]. However, the decision tree model can easily overfit the training data set and therefore it may perform poor on the testing data. Compared to the standard decision tree model, which uses the whole data set and whose nodes are split on different attributes among all variables, RF trains each individual tree on bootstrap resamples (M samples) of the total data set. For each split on the node of the individual tree, it only considers N_{try} randomly selected explanatory variables instead of the total explanatory variables (N). In this way, M decision trees are fitted and the final result is decided by average or majority voting. The reason to use bootstrap method and random select the subset of explanatory variables is to inject randomness into the RF such that the redundancy of explanatory variables is reduced and the forecast models (decision trees) are diversified [*Peters et al*., 2007; *Carlisle et al*., 2010; *Criminisi et al*., 2011]. The ensemble posterior is obtained by averaging M posteriors $p_t(P_d|\mathbf{v})$:

$$p(P_d|\mathbf{v}) = \frac{1}{M}\sum_{t=1}^{M} p_t(P_d|\mathbf{v}) \tag{3}$$

where $t = 1, 2, \ldots, M$. $p_t(P_d|\mathbf{v})$ is the leaf output for the *t-th* individual tree specifying the conditional distribution of the downscaled precipitation (P_d) given the multidimensional feature response vector ($\mathbf{v}$).

In this study, the RF algorithm is implemented in Python using the scikit-learn package [*Pedregosa et al*., 2011], which comes with built-in functions to evaluate the importance of each covariate. This is done using OOB (out-of-bag) samples (i.e., samples that are not chosen during the bootstrap split). The prediction strength of each covariate can be measured using the following steps [*Breiman*, 2001; *Liaw and Wiener*, 2002; *Friedman et al*., 2001]: (1) Randomly permuting the value of the *ith* covariate in the OOB samples and leaving other covariates unchanged; (2) Passing down the permuted OOB subsets to the *jth* tree and make a new forecast; (3) Averaging the mean square error over all trees and measuring the importance of *ith* covariate based on how much the prediction errors increase.

The two most important parameters in RF are N_{try} and M. N_{try} determines the variation among different decision trees and M influences the extent of overfitting [*Liaw and Wiener*, 2002]. Typically, $N_{try} = \sqrt{N}$ or $log_2 N$ (N is the number of covariates). Higher values of M are expected to yield better performance but will require more computational resources. In practice, M can be determined through the OOB error.

3.3. Synthetic Experiment Design

The purpose of using synthetic experiments is to constrain errors and avoid uncertainties, which may come from other sources, in order to test the performance of the Prec-DWARF algorithm before applying it to real cases for which the downscaled precipitation values are unknown.

3.3.1. Generate Synthetic Covariates

The synthetic experiments are used to explore how well the algorithm works in downscaling synthetically generated coarse-scale resolution precipitation. To achieve this, the observed NLDAS-2 precipitation at fine resolution (0.125°) is upscaled to coarse resolution using the simple box averaging method. Then the upscaled precipitation is disaggregated to the same resolution as the original observed precipitation using the bilinear interpolation (or the uniform disaggregation) method to ensure that the dimensions of the covariates and response variable are consistent (same grid number for latitude and longitude). Procedures to prepare the synthetic covariates are shown in Figure 2. In addition to the central grid cell, precipitation at adjacent grid cells is also included as covariates to maintain the large-scale spatial dependence of rainfall structure as precipitation at "nearby" locations tend to be similar and therefore is more informative.

For consistency, dynamic atmospheric covariates are regridded in the same way as the precipitation, assuming that in real-word applications these covariates would be available on the same coarse resolution as the precipitation (e.g., from a GCM). As static covariates (e.g., topographic information) in NLDAS-2 are commonly available at the global scale from satellite observations at high-resolution, we keep their original resolution (0.125°) in NLDAS-2. Details about the covariates used in this study are summarized in Table 1.

3.3.2. Scaling Experiments

Scaling experiments are used to test the sensitivity of Prec-DWARF's performance to different scaling ratios, which are defined as the ratio between the coarse resolution and the target resolution (0.125° in this study). We are most interested in evaluating Prec-DWARF's performance for high scaling ratios (e.g., from 1° to

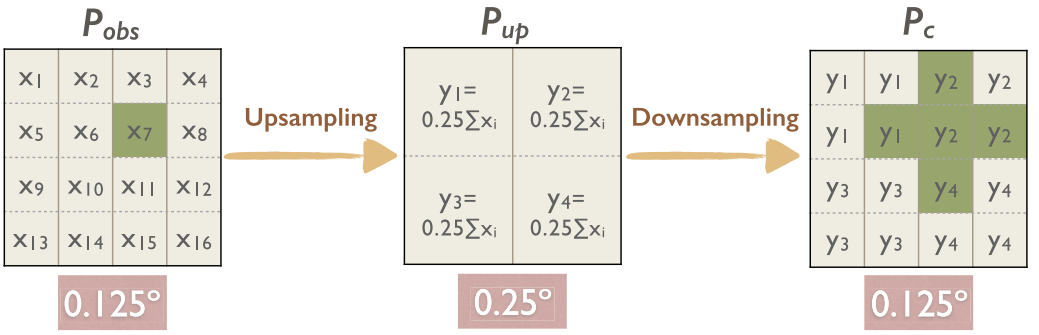


Figure 2. Procedures to prepare synthetic covariates. (a) Upscale the observed precipitation (P_{obs}) from fine resolution (0.125°) to coarse resolution (P_{up}, e.g., 0.25°) using the box averaging method. (b) Disaggregate precipitation from coarse resolution (P_{up}, 0.25°) to fine resolution (P_c, 0.125°) again using the uniform disaggregating method as an example (bilinear interpolation is used in the study). Note that adjacent grid cells (green shaded) are also included as the input covariates.

0.125°), which is typical of state-of-the-art GCMs and reanalysis data sets, although we expect the algorithm to perform better for small downscaling ratios. Besides experiments using coarse resolution (0.25°, 0.5° and 1°) precipitation and coarse resolution dynamic covariates, additional experiments are also tested by replacing the atmospheric covariates at coarse resolution with the original fine-resolution versions (0.125°). We find that only slight improvement in the downscaling performance is achieved in the experiments using high-resolution atmospheric covariates (not shown).

3.3.3. Training and Testing of Random Forests

3.3.3.1. Notations

Let $\mathbf{X}$ be the location in the 3-dimensional (3-D) space, which comprises 2-dimensional (2-D) space and time:

$$\mathbf{X}_{(j)} = [X_{1,j}, X_{2,j}, \ldots, X_{i,j}, \ldots, X_{T,j}] \tag{4}$$

where $X_{i,j} \in \mathbb{R}^2$ is a regular Cartesian grid for time step i at location j, T is the total time step for the simulation period. Downscaled precipitation at $\mathbf{X}_{(j)}$ is denoted as $P_d(\mathbf{X}_{(j)})$ and covariates at $\mathbf{X}_{(j)}$ are denoted as $\mathbf{v}(\mathbf{X}_{(j)})$:

$$\mathbf{v}(\mathbf{X}_{(j)}) = [P_c(\mathbf{X}^m_{(j)}), P_c(\mathbf{X}^u_{(j)}), P_c(\mathbf{X}^d_{(j)}), P_c(\mathbf{X}^l_{(j)}), P_c(\mathbf{X}^r_{(j)}), c_1(\mathbf{X}_{(j)}), c_2(\mathbf{X}_{(j)}), \ldots, c_k(\mathbf{X}_{(j)})] \tag{5}$$

Table 1. Summary of the Covariates Used for Downscaling Precipitation

Covariate Type	Covariate Name	Abbreviation	Resolution
Atmospheric	Precipitation in the target grid cell	Prec (central)	0.25°, 0.5°, 1°
	Precipitation at adjacent grid cell (upward)	Prec (up)	0.25°, 0.5°, 1°
	Precipitation at adjacent grid cell (downward)	Prec (down)	0.25°, 0.5°, 1°
	Precipitation at adjacent grid cell (leftward)	Prec (left)	0.25°, 0.5°, 1°
	Precipitation at adjacent grid cell (rightward)	Prec (right)	0.25°, 0.5°, 1°
	Air temperature at 2 meters above the surface	Temperature	0.25°°, 0.5°, 1°
	Specific humidity at 2 meters above the surface	Humidity	0.25°, 0.5°, 1°
	Meridional wind at 10 meters above the surface	Meridional wind	0.25°, 0.5°, 1°
	Zonal wind at 10 meters above the surface	Zonal wind	0.25°, 0.5°, 1°
	Surface pressure	Pressure	0.25°, 0.5°, 1°
	Convective available potential energy	CAPE	0.25°, 0.5°, 1°
Geographic	Distance to the closest dry grid cells	Distance	0.25°, 0.5°, 1°
	Mean value of elevation	Elevation (mean)	0.125°
	Standard deviation of elevation	Elevation (std)	0.125°
	Slope	Slope	0.125°
	Aspect	Aspect	0.125°
Auxiliary	Vegetation type	Veg type	0.125°
	Soil texture	Texture	0.125°
	Day of year	DOY	/
	Latitude	Lat	/
	Longitude	Lon	/

where $c_k(\mathbf{X}_{(j)})$ is the vector of individual covariates (except precipitation) at location $\mathbf{X}_{(j)}$ and $k=1,\ldots,N-5$. Again, N is the total number of covariates. $P_c(\cdot)$ represents precipitation at coarse resolution, superscript m, u, d, l, r indicate the middle, upward, downward, left and right direction at location j.

3.3.3.2. Selection of Training Samples and Testing of RF

In machine-learning world, training (off-line phase) and testing (on-line phase) are two necessary steps to verify the functionality of an algorithm. In particular, the performance of an algorithm can be very sensitive to how the total data set is split into training and testing samples. To avoid this, *K*-fold cross validation is usually conducted to properly represent the distribution of the population, especially when the total data set is limited [*Friedman et al.*, 2001]. However, conventional *K*-fold cross validation is usually applied either in the spatial domain or in the temporal domain, and so cannot represent the spatial dependence and temporal evolution at the same time when dealing with high-dimensional data samples. We therefore propose a novel way to create 3-D space-time images by stacking the hourly 2-D images in time. Training data sets can therefore be prepared through random sampling of the total T 3-D images using the following random vector $\mathcal{S}$:

$$\mathcal{S}=[S_1, S_2, \ldots, S_i, \ldots, S_T] \in \mathbb{R}^{n_{tr}\times T} \tag{6}$$

where n_{tr} is the number of training grid cells and S_i is the set of grid locations, which are randomly sampled from the total grid cells at time step i. This sampling process implicitly considers the domain-averaged temporal dependence as RF can learn the underlying structure at each time step and therefore the overall temporal structure is expected to be maintained. The response variable (downscaled precipitation) and covariates in the training samples can therefore be denoted as $P_d(\mathcal{S})$ and $\mathbf{v}(\mathcal{S})$. We use RF to construct the underlying relationship between $P_d(\mathcal{S})$ and $\mathbf{v}(\mathcal{S})$ from the training samples. Then the unselected pixels are used as a test data set. After training and testing, the downscaled precipitation for the full spatial domain can be reconstructed through the combination and rearrangement of the grid pixels. A schematic illustration of the Prec-DWARF algorithm is shown in Figure 3.

3.3.4. Single RF Versus Double RF

We find that the Prec-DWARF algorithm works well in downscaling the mean pattern of the precipitation field. The problem with the above approach is that it uses only 10% of the whole spatial domain at each time step. Because of the highly skewed distribution of rainfall rates, there is a high probability of picking solely dry grid cells or low precipitation values (e.g., drizzle). The trained RF is therefore highly unlikely to be representative of extreme precipitation. Taking this into consideration, we add a second RF specifically designed to capture the relationship between the covariates and the target rainfall field for heavy and extreme precipitation. This additional RF is used to train and test extreme precipitation events (precipitation above the 99.95th percentile for all grid cells over the entire simulation period). The remaining grid cells, which do not have heavy precipitation, are used to fit the other RF. As before, only 10% of the total grid cells are used to train each RF. The other 90% is used as testing samples. To determine the number of decision trees for RF, we investigate how OOB errors vary with the number of decision trees (M) in the single RF for different synthetic experiments over SEUS. As shown in Figure 4, there is no significant decrease of OOB errors after 20 individual decision trees (similar results hold for the other three regions), which means adding more trees is not necessary. Therefore, to optimize the computational efficiency and to produce stable predictions, the number of decision trees is set to 50 for both RFs. In summary, six experiments are conducted and their experimental configuration is listed in Table 2.

4. Results

4.1. Example of Spatial Distribution of Downscaled Fields

Prec-DWARF is used to downscale hourly precipitation for the six synthetic experiments (Table 2). We run the algorithm with a single RF and double RF for the selected four climatic regions using the data described in section 2. Here we only present results over SEUS as an example. Figure 5 shows a snapshot of observed precipitation, synthetic upscaled precipitation, and the downscaled precipitation. The original precipitation field at 0.125° (left, reference) shows a very localized pattern with a few spots of heavy precipitation surrounded by moderate to low-intensity regions, typical for warm-season thunderstorms. The upscaled synthetic precipitation captures the mean pattern and location of the reference data but cannot resolve the high-intensity cores

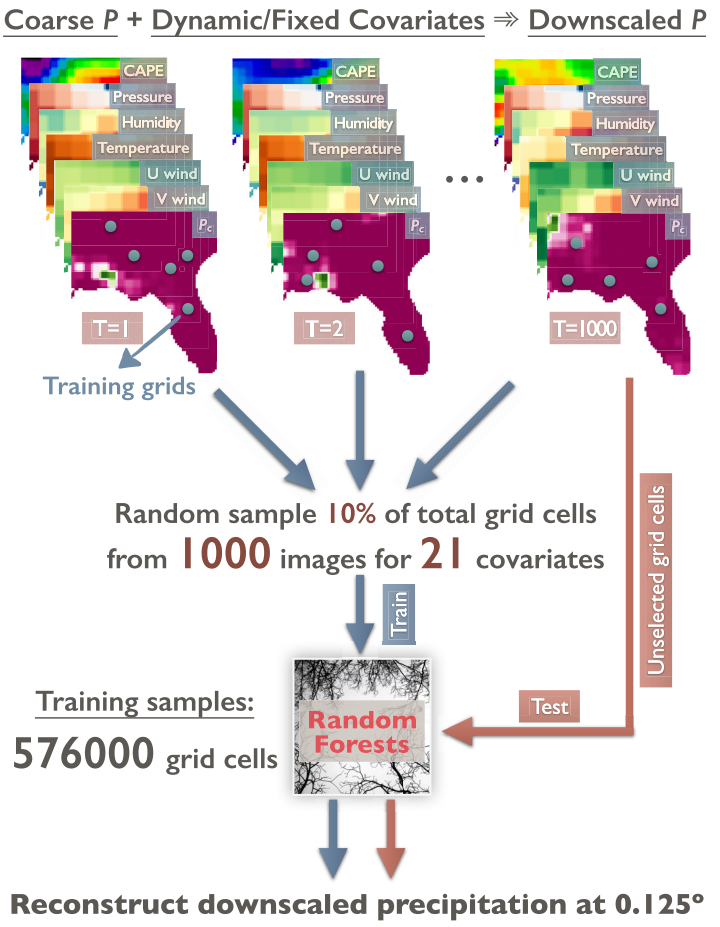


Figure 3. Schematic illustration of the Prec-DWARF algorithm to partition the four-dimensional (2-D spatial structure + 1-D time evolution + 1-D multiple covariates) data structure for training and testing purpose. Training samples are prepared by random sampling 10% of total grid cells at each time step for each covariate.

and fine-scale variability. Using Prec-DWARF with a single RF (Figure 5, middle), the spatial features of downscaled precipitation can be well resolved, especially for the 0.25° experiment. Nevertheless, as the scaling ratio increases (e.g., 0.5° and 1° experiments), the spatial structure of precipitation becomes noisier and patchier as more low intensity precipitation is generated. The dry/wet transitions and spatial distribution of rainfall rates within rainy regions are also not well reproduced. Moreover, both the location and magnitude of the extreme precipitation are not properly resolved. This lost signal of extreme precipitation and the precipitation gradient limits its hydrological application. The double RF model on the other hand (see Figure 5, bottom) performs better, successfully resolving the location and magnitude of the heavy precipitation and clearly outperforms the single RF at larger-scaling ratios (e.g., 1° experiment). In addition, it generates a more realistic area of low intensity rainfall in the state of Alabama (see Figure 5, bottom right).

4.2. Overall Accuracy of the Prec-DWARF

In the following, a comprehensive evaluation of the algorithm's performance over different regions is conducted. Specifically, the bias, distribution, spatial/temporal dependence, dry-wet classification as well as the uncertainty in the downscaled field are compared to the reference data. We also include the bilinear interpolation as the benchmark comparison.

4.2.1. Goodness-of-Fit Metrics

The overall agreement between the observed and downscaled precipitation is quantified using the root-mean square error (RMSE) and correlation coefficient (R). These two metrics are calculated on an hourly basis for the entire time period and for all land grid cells in the considered domains. Results for the four climatic regions are summarized in Figure 6. As expected, as the spatial resolution of the upscaled precipitation (P_{up}) gets coarser, the RMSE increases and R decreases. The smallest RMSE and highest R are obtained

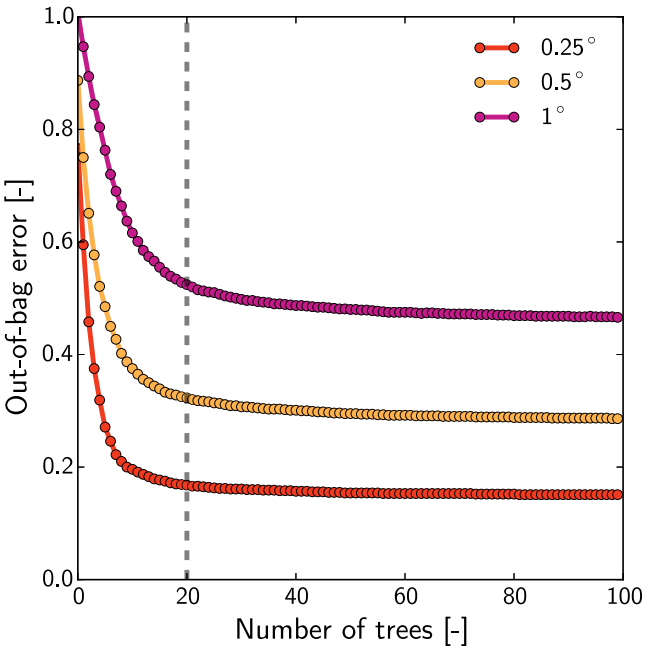


Figure 4. RF applied for different synthetic experiments with different resolutions over SEUS. The curves represent the out-of-bag error as a function of the number of decision trees in RF.

in the 0.25° experiments. This is not surprising as it is more difficult to disaggregate one grid cell into 64 grid cells (downscale from 1° to 0.125°) compared to four grid cells (downscale from 0.25° to 0.125°). But as indicated by the *R* values, there is still good agreement (>0.5) even for the 1° experiment. Comparison between the single RF and double RF shows clear evidence that the latter has superior performance (reduced RMSE and increased *R*). This improvement is more apparent for the coarser resolution experiments (1°) than 0.25° and 0.5° experiments, which further demonstrates the necessity of using a double RF. This significant improvement comes from the fact that double RF can better capture the magnitude and spatial structure of extreme events, which can reduce the bias and increase the correlation. This is consistent with previous findings of the spatial pattern as shown in Figure 5. Compared with the benchmark experiments which use bilinear interpolation for precipitation disaggregation, Prec-DWARF shows systematically lower RMSE and higher *R* values for both single and double RF configuration over all climatic regions, demonstrating that our proposed algorithm can well reduce the downscaling bias.

4.2.2. Precipitation Distribution

We further evaluate the downscaled precipitation in terms of the full distribution using Quantile-Quantile (Q-Q) plots (Figure 7). Compared with the benchmark experiments, the downscaled precipitation from all six synthetic experiments across all regions (except the 1° experiment with single RF over SWUS) correspond much better to the observed distribution through most of the range, yielding data points very close to the reference line. However, the downscaled precipitation underestimates the extreme values, and more prominently for those results simulated using a single RF. In addition, the quantiles of the 1° experiments are further from the 1:1 line, followed by 0.5° and 0.25° experiments. Prec-DWARF configured with double RF shows superior performance compared to a single RF, especially for upper part of the distribution. This improvement is more obvious in the coarse resolution experiments (e.g., 1°). Moreover, the double RF also increases the range/threshold, up to which the downscaled precipitation closely resembles the distribution of the observations. For example, in the 1° experiment over SWUS (Figure 7a), the single RF starts to underestimate the observed precipitation above about 5 mm. For the double RF, the probabilities remain almost identical up to 20 mm. We also observe regional differences in the performance of the downscaling algorithm. For example, Prec-DWARF performs better in CUS, NEUS, and SEUS than SWUS. This is perhaps due to the fact that the current algorithm does not handle the orographic effect very well, as in the southwestern United States precipitation is highly influenced by the Rocky Mountains. The poor performance over SWUS could also be related to the data quality of NLDAS-2, which affects whether the actual statistical relationship between the response and covariates can be realistically reconstructed by the regressions.

Table 2. Description of Synthetic Experiments

Experiment	Coarse Resolution	Number of RF	Scaling Ratio
1RF_0.25°	0.25°	1	2
2RF_0.25°	0.25°	2	2
1RF_0.5°	0.5°	1	4
2RF_0.5°	0.5°	2	4
1RF_1°	1°	1	8
2RF_1°	1°	2	8

4.2.3. Spatial and Temporal Dependence

The ability of the algorithm to reproduce the spatial and temporal dependence structures

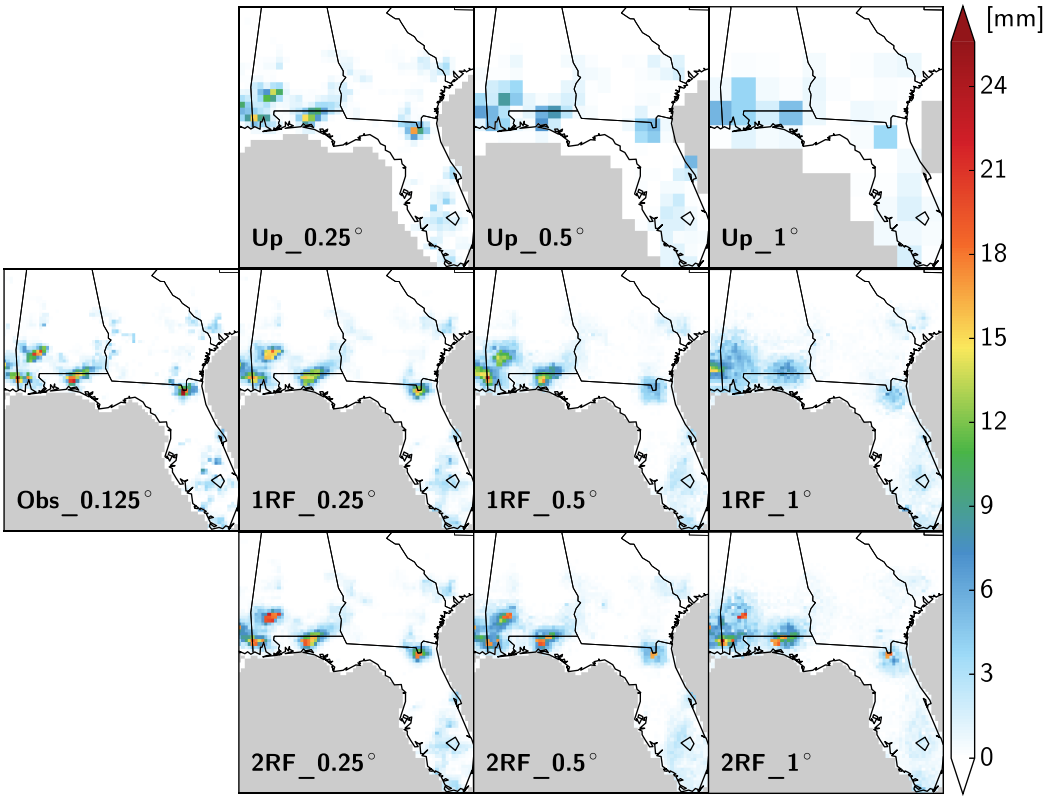


Figure 5. (left column) Example of the spatial distribution of NLDAS-2 observed precipitation at 0.125°, (top) synthetic upscaled precipitation, (middle) downscaled precipitation using a single RF, and (bottom) downscaled precipitation using a double RF for different scaling ratios over the SEUS region at 22:00 on 11 July 2011.

is quantified by comparing the spatial and temporal semivariograms of the observed and downscaled precipitation fields. By definition, the empirical semivariance γ_s (the subscript "s" represents "spatial") can be calculated as:

$$\gamma_s(h_s) = \frac{1}{2}\langle [I(j+h_s) - I(j)]^2 \rangle \tag{7}$$

where $I(j)$ is the precipitation intensity at location j and $h_s \in \mathbb{R}^2$ is a displacement vector (spacing between grid cells). The R library "gstat" [*Pebesma*, 2004] is used to calculate and model the semivariance. Figure 8 shows the observed and downscaled spatial semivariance for different spatial resolutions at one particular time step over SEUS. As expected, the spatial variability increases as a function of the lagged distance. The shape of the variogram indicates that most of the spatial correlation is lost after 15 lags (one lag corresponds to 0.125°). However, for the downscaled precipitation, this decorrelation distance (or range) is larger (around 25 lags), which is a sign that the algorithm fails to reproduce the localized spatial structure and instead tends to generate a less variable and smoother pattern. As expected, the discrepancy between observed and downscaled semivariance values increases with increasing scaling ratio. Overall, the 0.25° experiment yields the best results, but still underestimates the small-scale spatial variability.

A useful quantity to look at when evaluating the small-scale spatial structure of the precipitation fields is the spatial semivariance value at lag 1 (denoted as $\gamma_s(1)$, pentagon symbol in Figure 8), as this value shows (half of) the average squared difference in rainfall intensity for two neighboring grid cells. Spatial semivariance values at larger distance lags are also interesting, but mostly constrained by the data at the coarse resolution and hence not very relevant for assessing the spatial structure of the downscaled fields. Therefore, only $\gamma_s(1)$ will be considered in the following. Figure 9 shows the temporal evolution of $\gamma_s(1)$ over 200 h for different spatial resolutions (0.25°, 0.5°, and 1°) and different RF configurations

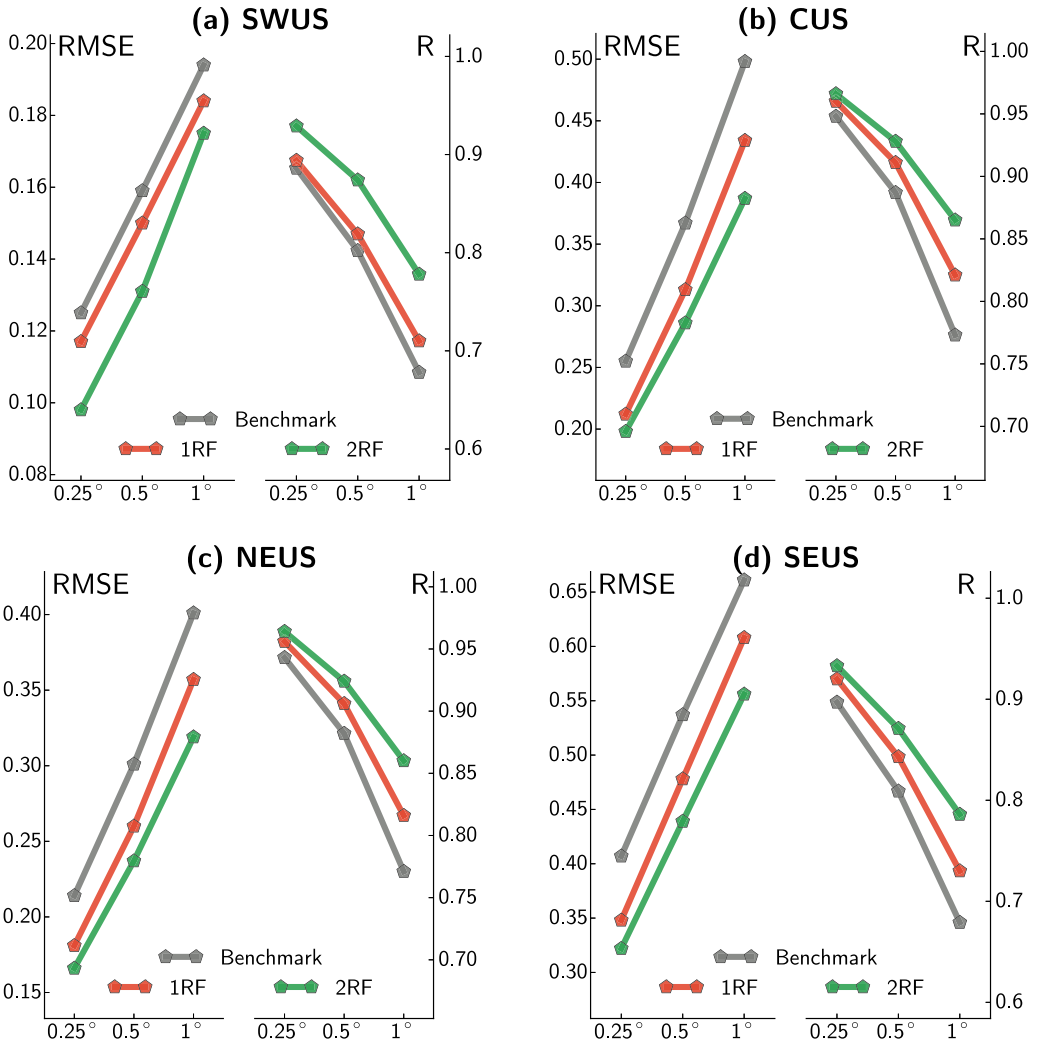


Figure 6. (left axis) Root-mean-square-error (RMSE, unit: mm) and (right axis) correlation coefficient (*R*) showing how Prec-DWARF's downscaling performance (goodness-of-fit) varies with the spatial resolution over (a) SWUS, (b) CUS, (c) NEUS, and (d) SEUS. Gray, red, and green symbols represent results for benchmark, single RF, and double RF experiments.

(single RF or double RF). Results are compared with benchmark experiments. There is a clear diurnal cycle of $\gamma_s(1)$ in this time period, which is associated with the strong diurnal cycle of precipitation in the summer time over SEUS (Figure 9a). As can be seen, the benchmark experiments significantly underestimate the observed semivariance and fail to capture its temporal dynamics for all resolutions. In contrast, Prec-DWARF is able to capture the temporal dynamics of the spatial semivariance at lag 1 relatively well. With higher scaling ratios, the deviation between the downscaled and the observed $\gamma_s(1)$ increases, indicating that the downscaled precipitation in the 1° experiments tends to have a more homogeneous spatial pattern. However, the use of a double RF improves the representation of observed spatial variability compared to the single RF, especially for the 1° experiment. This further demonstrates the advantage of using a double RF rather than a single RF. Figure 10 summarizes the comparison between the observed $\gamma_s(1)$ and the downscaled $\gamma_s(1)$ in the spatial domain for all time steps and for all regions. The benchmark exhibits the strongest deviations from the 1:1 line, indicating its poor performance in capturing the spatial variability at subgrid scale. Prec-DWARF does a better job than bilinear interpolation, even though the downscaled spatial semivariance is still underestimated, especially during times of heavy rain rates and large spatial variability. In addition, the double RF (green colors) appears to be much better at reproducing the observed lag-1 semivariance during periods of heavy rain rates than the single RF (red colors).

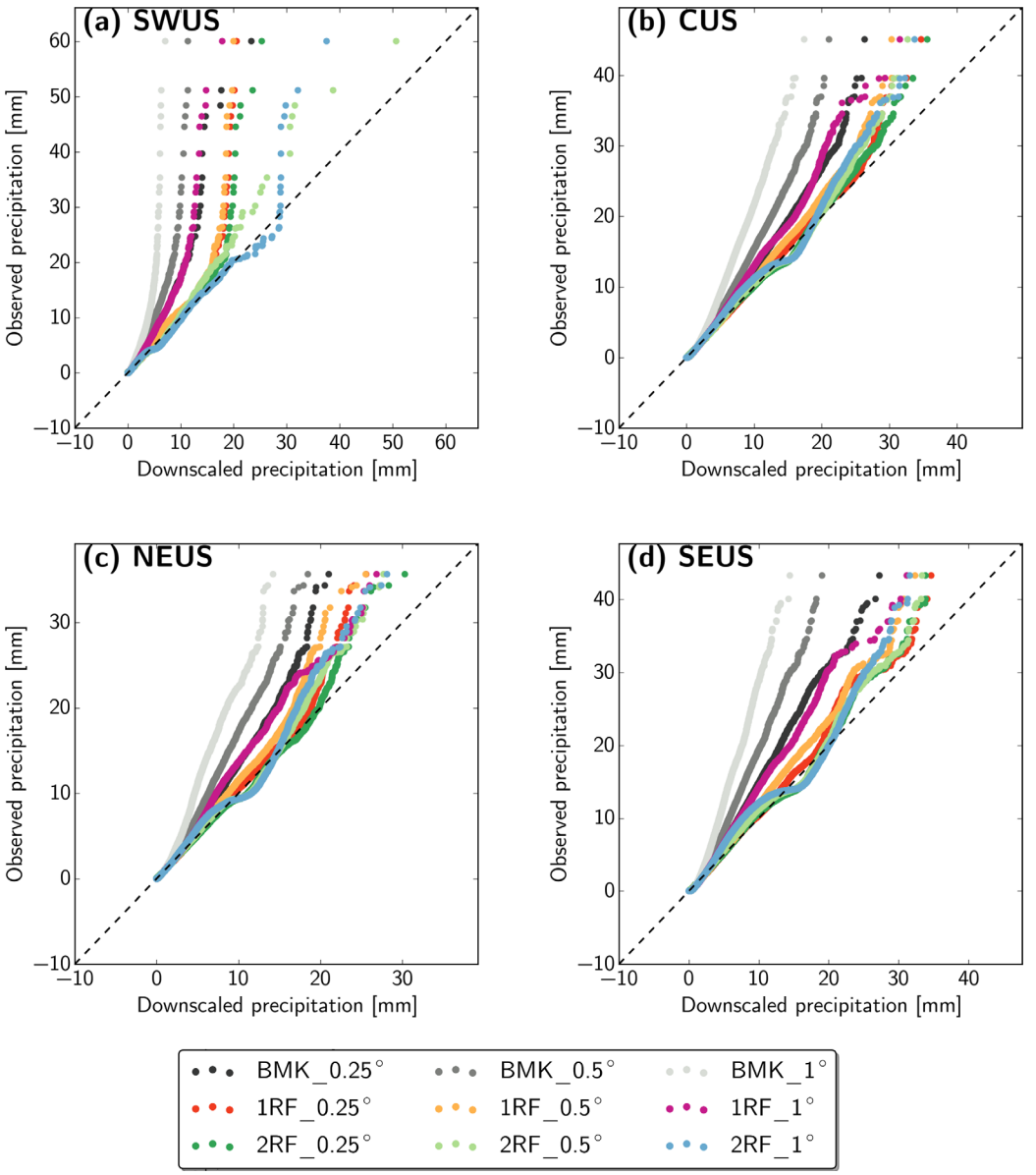


Figure 7. Quantile-Quantile (Q-Q) plot showing the relationship between the observed and downscaled precipitation over (a) SWUS, (b) CUS, (c) NEUS, and (d) SEUS.

Although this study does not focus on the temporal downscaling, we argue that Prec-DWARF can capture the temporal dependence of the observed precipitation to some extent due to the embedded sampling algorithm. To confirm this, we computed the temporal semivariance γ_t for the observed and downscaled precipitation time series (at each grid cell):

$$\gamma_t(h_t) = \frac{1}{2}\langle [I(t+h_t) - I(t)]^2 \rangle \tag{8}$$

This is similar to equation (7), but here t is the time step and h_t specifies the interval between time steps. $\gamma_t(1)$ corresponds to (half of) the average squared difference between rainfall intensities (at a fixed location) separated by a time difference of 1 h. $\gamma_t(1)$ is normalized by the total variance of the whole time series at each grid cell to make the magnitude comparable among different experiments. The percentage describing the relative overestimation of downscaled semivariance to the observed semivariance can be

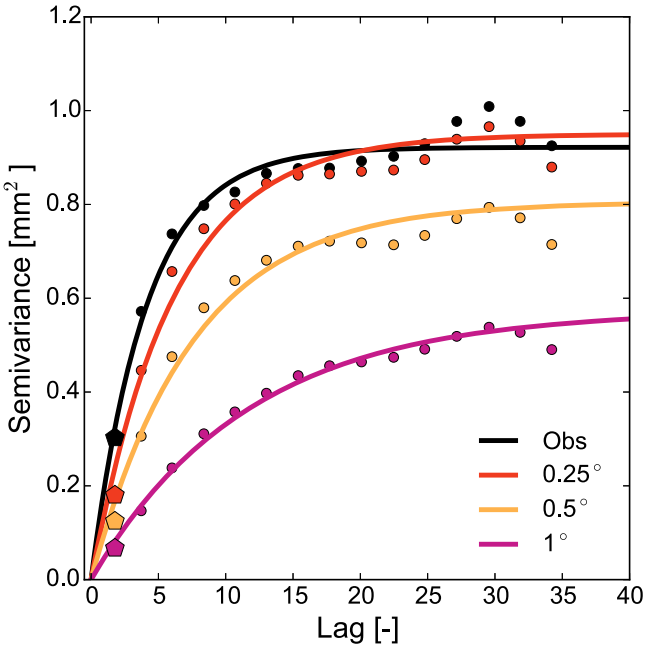


Figure 8. Empirical and fitted exponential variogram models for observed and downscaled precipitation with different coarse resolutions over the SEUS region. The pentagon symbol represents semivariance at lag 1.

expressed as: $\delta = \frac{\widetilde{\gamma_{t,sim}(1)}}{\widetilde{\gamma_{t,obs}(1)}} - 1$, where $\widetilde{\gamma_{t,sim}(1)}$ and $\widetilde{\gamma_{t,obs}(1)}$ are the normalized temporal semivariance at lag 1 for downscaled and observed precipitation separately. Positive/negative values of δ indicate that the algorithm overestimates/underestimates the temporal variability. Figure 11 shows that all nine experiments underestimate the observed lag-1 semivariance at most grid cells and the underestimation gets worse with increasing scaling ratio. The overall underestimation for benchmark experiments is 46.27%, 62.76%, and 76.71% for the 0.25°, 0.5°, and 1° experiments. For single RF, these values reduce to 26.10%, 42.51%, and 60.74%. The double RF slightly performed better with underestimation of 24.37%, 38.53%, and 56.20%. Moreover, the double RF experiments (Figure 11, bottom) show a similar spatial distribution compared to that in the single RF experiments (Figure 11, middle). The underestimated temporal variability at the hourly time scale is not surprising and can be explained by the algorithm's tendency to underestimate the spatial variability of the precipitation.

4.2.4. Dry-Wet Classification

The algorithm's ability to reproduce precipitation occurrences (i.e., dry and wet periods) at an hourly timescale is investigated using receiver operating characteristic (ROC) curves [*Mason and Graham*, 1999]. ROC curves

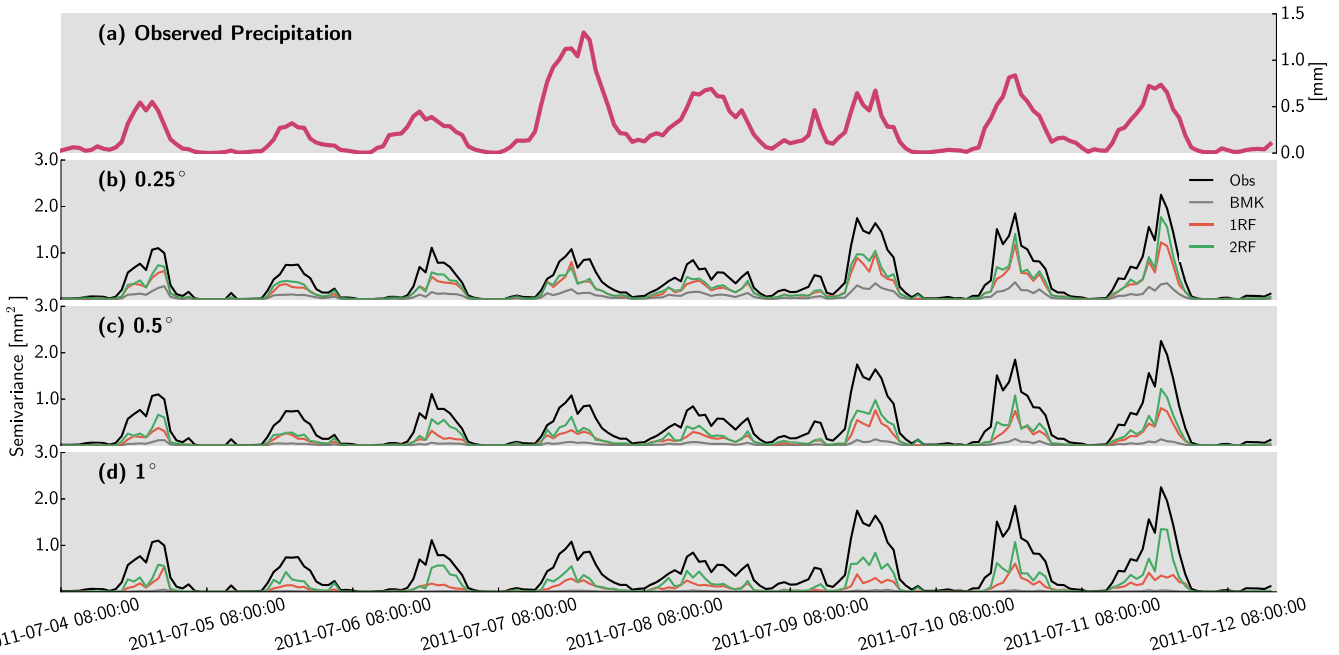


Figure 9. Temporal evolution of (a) domain-averaged NLDAS-2 precipitation and the spatial semivariance at lag 1 (pentagon symbol in Figure 8) for observed (black line) and downscaled precipitation using bilinear interpolation (gray line), a single RF (red line), and double RF (green line) for (b) 0.25°, (c) 0.5°, and (d) 1° experiments over SEUS.

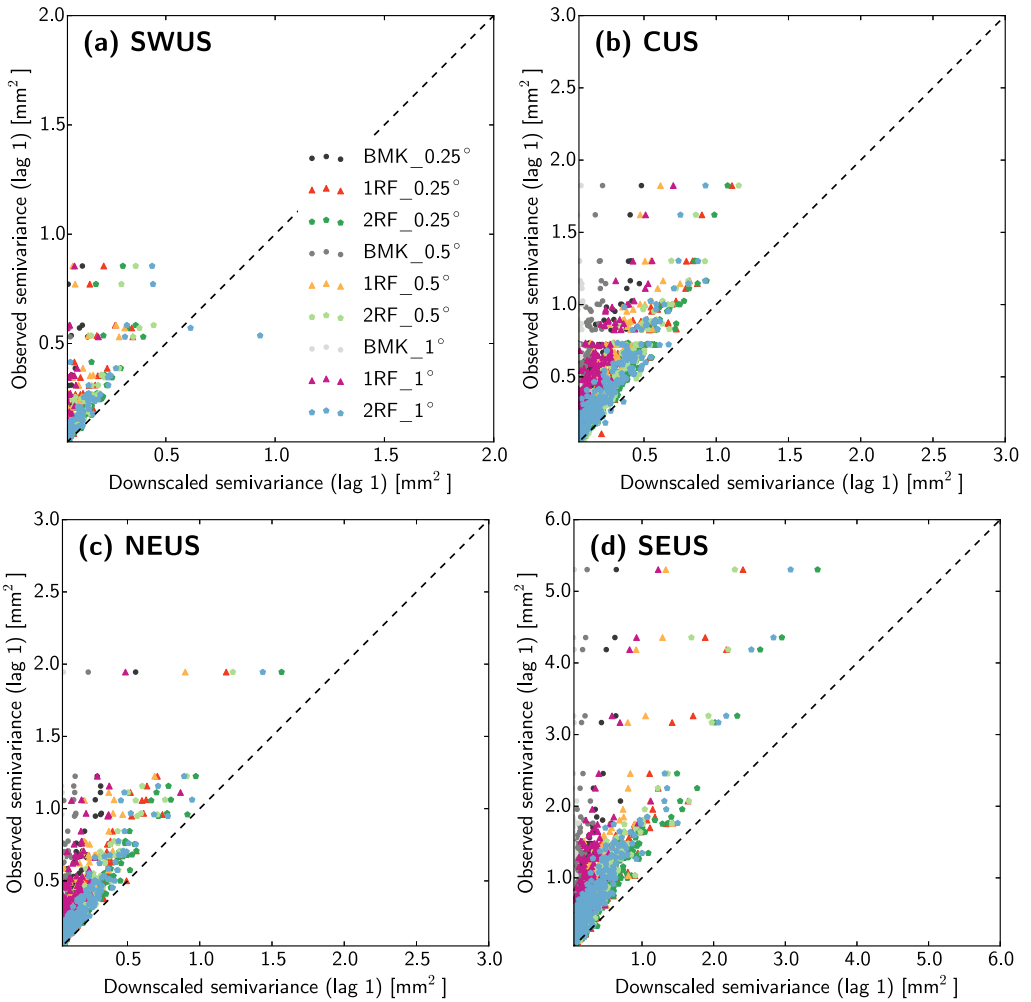


Figure 10. Scatterplot of semivariance at lag 1 between observed precipitation and downscaled precipitation over (a) SWUS, (b) CUS, (c) NEUS, and (d) SEUS.

show the relationship between the false positive ratio (*fpr*, *x* axis) and the true positive ratio (*tpr*, *y* axis). In our case, *fpr* represents the fraction of downscaled pixels that were classified as wet but were actually dry in the observations. *tpr* gives the fraction of downscaled pixels that were correctly classified as wet. Low *fpr* indicates a low false alarm ratio and high *tpr* indicates a high number of correct predictions of precipitation occurrence. In an ROC curve, different *fpr/tpr* pairs are plotted for different thresholds. A point located in the upper left corner ($fpr = 0$ and $tpr = 1$) indicates a perfect prediction. From Figure 12, we can see that Prec-DWARF does a better job than bilinear interpolation method in terms of the dry-wet classification for all climatic regions, as the ROC curves lie toward the upper left corner. The slopes of ROC curves over SWUS, CUS, and NEUS are steeper than those over SEUS, indicating that the algorithm does better in these three regions. For each climatic region, the slope of the ROC curve decreases with lower resolution. In other words, the *tpr* decreases and the *fpr* increases at lower resolution, meaning that the number of correctly downscaled precipitation occurrence is reduced and the false alarm ratio is increased. Again, the double RF beats the single RF, this time in terms of dry-wet classification and this improvement is most distinct over SEUS for the 1° experiments.

4.2.5. Uncertainty Analysis

As RF is an ensemble-based method that consists of multiple decision trees, it can also be used to quantify the uncertainty associated with the downscaled precipitation values. The uncertainty in the downscaled precipitation can be represented by the ensemble spread. Figure 13 compares the domain-averaged observed precipitation to the downscaled precipitation with a single RF (top) and double RF (bottom) for different resolutions. The temporal variability of the downscaled mean value, which is averaged over 50 decision trees, is in good agreement with that of the observations for the 0.25° experiments, both for the single and double RF. As the

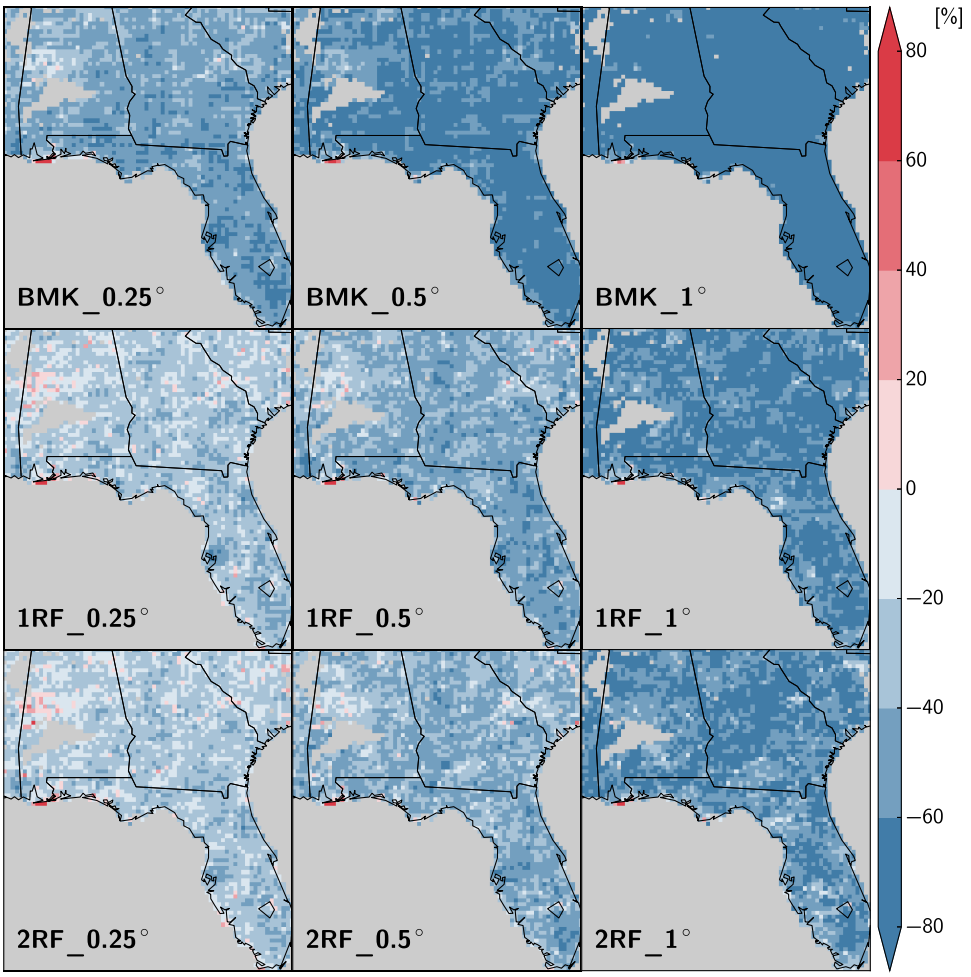


Figure 11. Spatial distribution of δ describing the relative underestimation of the observed temporal semivariance for the (top) benchmark, (middle) single RF, and (bottom) double RF experiments with different scaling ratios over SEUS. Grid cells containing less than 100 pairs of positive precipitation in the time series are masked out when calculating $\gamma_t(1)$ to avoid unreliable semivariable estimates.

resolution becomes coarser (e.g., 1° experiments), the ensemble mean deviates more from the observations. Regardless of the number of RFs, the 0.25° experiments have the smallest ensemble spread, whereas the 1° experiments have the largest. Compared with the single RF, the double RF yields a larger ensemble spread, especially during heavy precipitation events. This makes sense, as extreme rainfall events are often associated with larger spatial and temporal variability and tend to be harder to downscale. The double RF has separate settings for moderate and extreme precipitation and is able to better account for this effect.

4.3. Relative Importance of Individual Covariates

As noted earlier, the predictability of each covariate in Prec-DWARF can be assessed through the covariate importance plot, as shown in Figure 14. Here we show the relative feature importance for all the 21 covariates for single RF and double RF over SEUS. For the double RF, we explicitly output the feature importance for moderate precipitation and extreme precipitation. Precipitation in the central grid ($P_c(\mathbf{X}_{(j)}^m)$) is the most important covariate, followed by precipitation at adjacent grid cells. For the single RF, the relative importance of $P_c(\mathbf{X}_{(j)}^m)$ is larger than 0.8 in the 0.25° experiment. Although this value decreases quickly in the 0.5° and 1° experiments, $P_c(\mathbf{X}_{(j)}^m)$ is still the most important among all the covariates. The ranking is similar as the double RF for moderate precipitation (Figure 14, middle). However, significant differences emerge when RF is applied to downscale the extreme precipitation (Figure 14, right). Once again, $P_c(\mathbf{X}_{(j)}^m)$ has the strongest predictive power for the 0.25° experiment, but its relative importance reduces to less than 0.4. Other covariates start to play a more important role, especially precipitation at adjacent grid cells and atmospheric covariates. For example, the relative importance of temperature is more than 0.15, ranking third in the 0.25° experiment. This strong influence

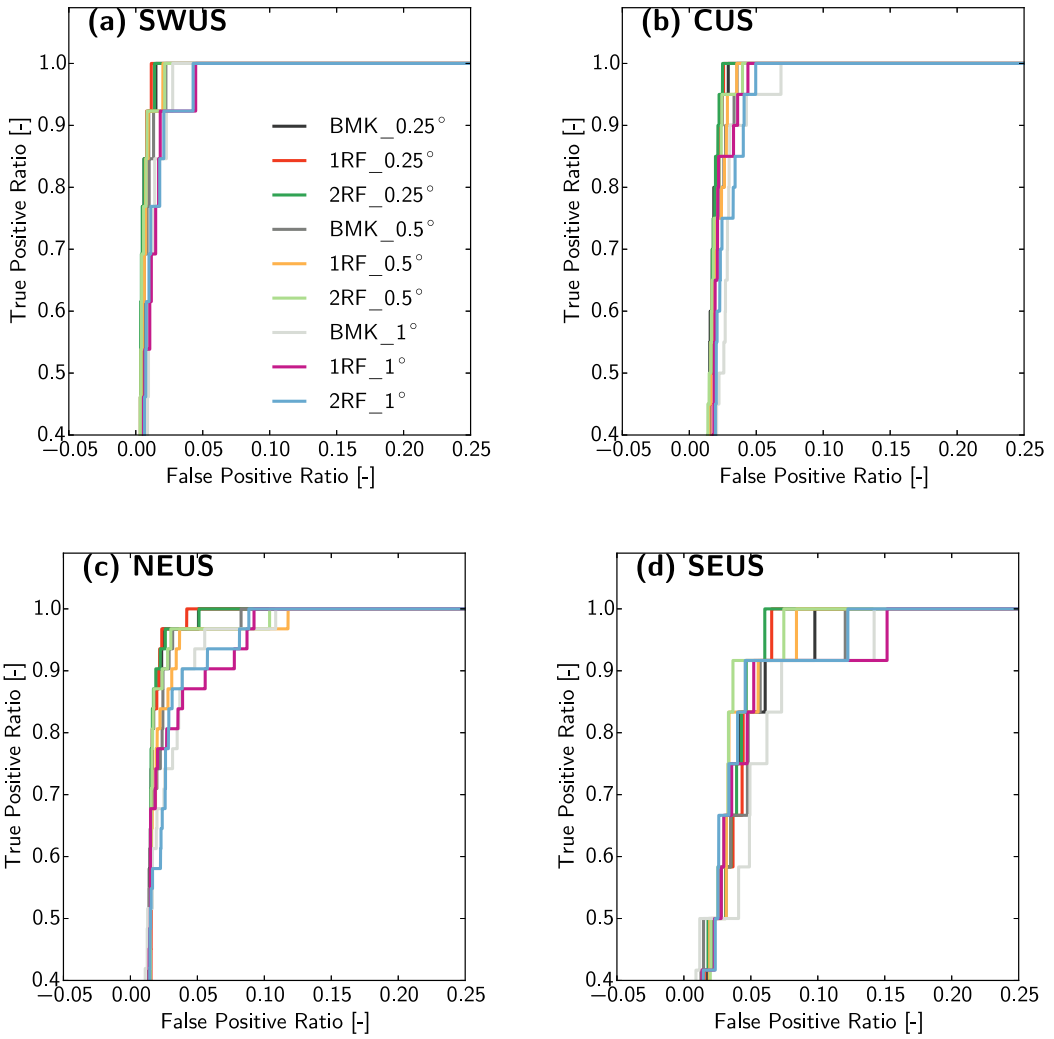


Figure 12. Receiver operating characteristic (ROC) curve for benchmark, single RF, and double RF experiments over (a) SWUS, (b) CUS, (c) NEUS, and (d) SEUS.

of temperature for extreme precipitation downscaling can be explained by the Clausius-Claperon (CC) superscaling [*Lenderink and Van Meijgaard*, 2008; *Haerter et al.*, 2010; *Utsumi et al.*, 2011], which predicts that extreme precipitation intensity increases with temperature beyond the standard CC rate. A recent study by *Berg et al.* [2013] demonstrates that this CC superscaling can be more dominant for convective precipitation, which is consistent with our experiment (such as thunderstorms over SEUS). Besides temperature, other atmospheric covariates also contribute more when used to downscale extreme precipitation. The covariate for distance to the closest dry grid cell ranks even higher than precipitation in the 0.5° and 1° experiments, demonstrating the strong relationship between the precipitation intensity and the area, size, and shape of the wet regions. Including the "dry drift" seems to be particularly valuable for predicting extreme precipitation. Topographic covariates had relatively low importance here. This may be because the regions are large or that the orographic dependence may be missing in the NLDAS-2 data [*Pan et al.*, 2003]. Here we only show the covariate importance spectrum over SEUS, but the findings are similar for the other three regions.

5. Summary and Discussion

5.1. Summary

A novel machine-learning algorithm, Prec-DWARF (**Prec**ipitation **D**ownscaling **W**ith **A**daptable **R**andom **F**orests) for spatial precipitation downscaling has been proposed. Two different implementations are

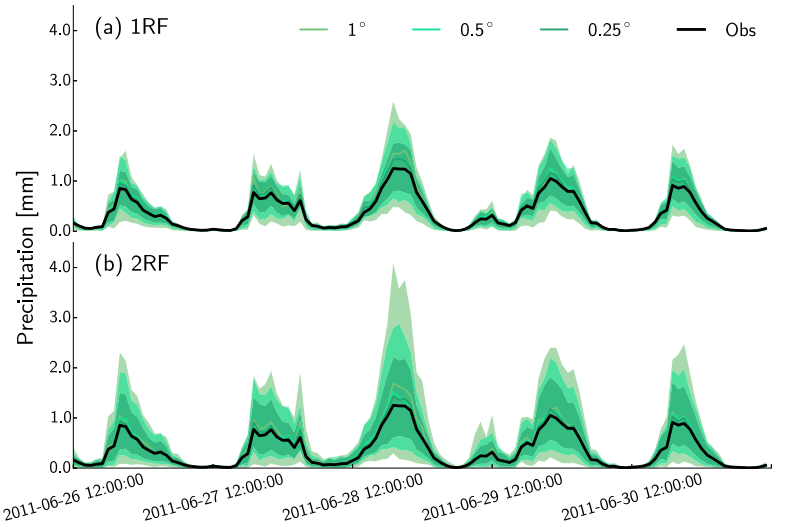


Figure 13. Temporal evolution of the domain averaged precipitation calculated from (top) a single RF and (bottom) a double RF using the ensemble mean (solid line) of 50 decision trees. Shadings with different colors represent the uncertainty spread at different resolutions. The observed precipitation is represented by the black line.

considered: a single RF for low to moderate precipitation values and a double RF for more localized and intense rainfall. Synthetic experiments over four climatic regions in the United States and three different scaling ratios are conducted to assess the performance of the algorithm based on the NLDAS-2 data sets. We present a comprehensive analysis to assess Prec-DWARF's downscaling performance in terms of the spatial and temporal patterns, the overall goodness-of-fit (bias and correlation), distribution, spatial and temporal dependence, wet-dry classification, and the associated uncertainty in the downscaled field. The results demonstrate that Prec-DWARF successfully reproduces the geometrical and statistical characteristics of the NLDAS-2 precipitation, especially for experiments with lower scaling ratios. The double RF consistently performs better than the single RF and the improvement is most significant for experiments with higher scaling ratios. However, the algorithm consistently underestimates the spatial variability and temporal

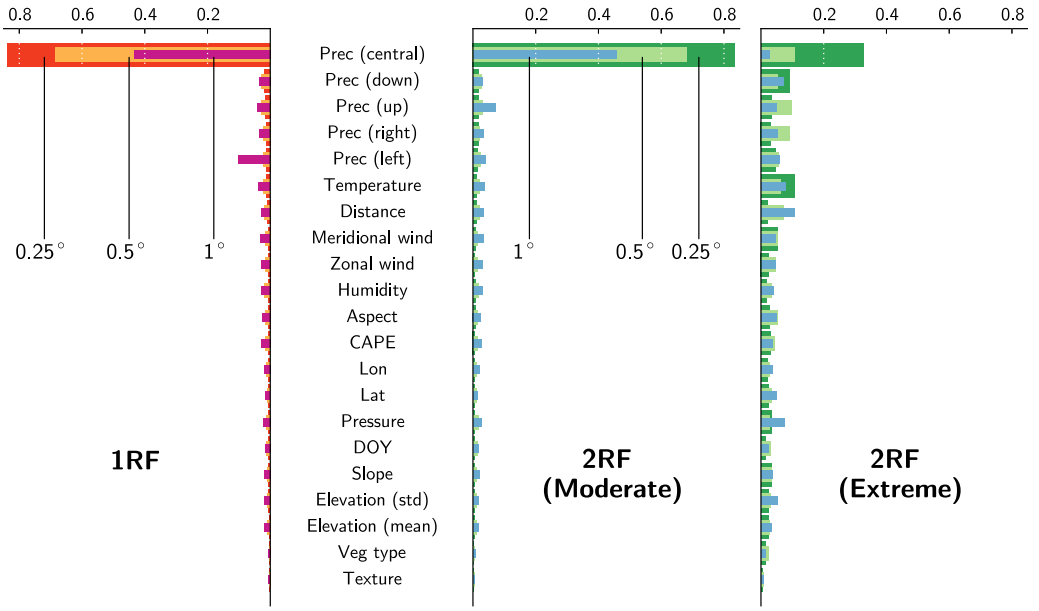


Figure 14. Covariate importance spectrum for (left) the single RF, (middle) double RF for moderate precipitation, and (right) extreme precipitation with different spatial resolutions (0.25°, 0.5°, and 1°) over SEUS. Feature importance (*x* axis) is calculated based on the Gini splitting index.

dependence (as represented by the semi-variance) and frequency of very high rainfall rates (especially over SWUS). In addition, the algorithm overestimates the amount and spatial extent of low intensity rainfall (e.g., drizzle). The poor performance of Prec-DWARF in these aspects can be attributed both to methodological limitations and the difficulty to define adequate covariates for describing and predicting small-scale precipitation variability. Despite the overall promising performance, the following issues need to be addressed.

5.2. Scaling Issues

The current study only focuses on precipitation downscaling in the spatial domain. We do not vary the temporal resolution, but instead only test the performance of Prec-DWARF at hourly scale. Nevertheless, Prec-DWARF is flexible enough to be extended to other temporal resolutions as well. In future studies, it will be important to assess the overall performance of the algorithm for different temporal and space-time scaling ratios. Additional efforts are also needed to determine the performance of the algorithm not only with respect to the scaling ratio but also to the actual spatial scale at which the precipitation is considered. For example, downscaling precipitation from 10 to 2.5 km might be more difficult than to downscale from 100 to 25 km, even though the scaling ratio is identical.

5.3. NLDAS-2 Data and Orographic Precipitation

Data availability and quality play a key role in the machine-learning approach. The constructed relationship highly depends on how representative the training samples are. For precipitation downscaling, how well the training samples represent the full spatial-temporal domain and capture the unique properties of precipitation (highly skewed, intermittent, and spatially correlated) determine how well the algorithm can be used for testing samples. The evaluation of the algorithm also depends on the quality of the data, which in the case of the NLDAS-2 is subject to how the radar and gauge observations are blended, and the spatio-temporal sampling of the rain gauges. In the dry SWUS, due to the scarcity of rain gauges, NLDAS-2 itself may not represent extreme events well [*Guirguis and Avissar*, 2008; *Abatzoglou*, 2013]. This may partially explain the poor performance of Prec-DWARF over SWUS, as shown in the Q-Q plot (Figure 7a). The underestimation of precipitation in the NLDAS-2 at higher elevations has been noted previously [*Pan et al*., 2003] and may contribute to the low importance of elevation as a covariate. *Pan et al*. [2003] found that over the western mountains of the United States, the lack of high elevation gauges in the NLDAS precipitation analysis was the main reason for the underestimation of orographic precipitation.

5.4. Number of RF and Stratified Sampling

In this study, we create a novel subsampling approach to explore the 4-D space (2-D spatial structure + 1-D time evolution + 1-D multiple covariates). We also explicitly sample extreme precipitation and use an additional RF to downscale that part of the data. Despite the relatively good results, the current subsampling method explores the whole spatial domain to extract the training grid cells, but could be improved by sampling preferentially from wet grid cells. In terms of the number of RF, more RF can be used, which will increase the likelihood of maintaining the spatial-temporal dependence as well as reproducing the extreme values, but it also increases the algorithm's complexity as more parameters are needed and it also puts higher requirements on data availability. A single RF would however be preferable, which is more cost-effective for tuning parameters, training, and testing compared to multiple RFs. To use a single RF, the current algorithm could be altered by explicitly sampling the full statistical distribution of rainfall intensity (including its tail). Meanwhile, we also need a better understanding of different physical mechanisms through which precipitation can be formed. The covariates used in this study may be insufficient to distinguish between these different regimes. Moreover, there appears to be a trade-off between the number of covariates, their resolution, and the number of RF needed to predict the precipitation at small scales. Perhaps the only way to achieve better performance using a single RF would be to have access to better or more covariates (see next section).

5.5. Potential Physical and Statistical Covariates

Other physical covariates can be taken into account for further study based on the precipitation type and associated physical processes. For example, optical and microphysical cloud parameters are very useful to distinguish convective and stratiform precipitation. Variables related to vertical cloud base profiles (e.g., cloud-top temperature, cloud fraction) could be utilized from multiple satellite imagery. The distance to the coast can be added to the set of covariates, as precipitation may be influenced by strong land-sea

interaction. Aerosols related to the degree of urbanization may also be useful over urban areas. There is also the possibility to combine different covariates in a nonlinear way, e.g., by multiplying them.

Besides these physical covariates, statistical covariates derived from rainfall characteristics themselves may also provide valuable information similarly to the "dry drift." For example, to better downscale extreme events, we can add covariates from the frequency domain (wavelet) to Prec-DWARF, which may help to reproduce their intensity and gradient at multiple scales. This has great potential and flexibility to reconstruct high-frequency signals and their fluctuations, which are lost or smoothed out when regression-based methods are used [*Kumar and Foufoula-Georgiou*, 1993; *Ebtehaj et al*., 2012; *Foufoula-Georgiou et al*., 2014]. To implement Prec-DWARF for temporal downscaling, the temporal innovation and predictability at different time lags need to be considered.

5.6. Transferability of the Method

The general flexibility of Prec-DWARF that derives from a machine-learning approach enables its application to a variety of situations and data sets. The techniques and ideas described in this paper are relevant for many downscaling algorithms and can be adapted to other situations. For example, the algorithm can be trained on point observations (rain gauges), and be used to downscale different gridded data sets (e.g., satellite, reanalysis, or coupled model output). However, the applicability of the algorithm to downscaling climate output depends on whether the observation-based relationships between the large-scale and small-scale variables still hold in the model-based data sets. Prec-DWARF can also be adapted to different types of precipitation (convective, stratiform, and orographic precipitation) considering different mechanisms that produce precipitation, perhaps based on weather-type classification. Multivariate (e.g., precipitation and temperature) downscaling can also be performed to better constrain their physical relationships. The proposed algorithm can also be used for real-time downscaling. In that case, the training procedure should be slightly modified. Ideally, one would like to train the algorithm using all the available historical data to maximize the learning ability. However, this might not be feasible due to computational constraints. Therefore, it might only be possible to train the algorithm periodically instead of at each time step.

One key application of the method is to downscale coarse resolution satellite or model data in regions without high-resolution observations, such as in most parts of Africa. A possible approach would be to train the model in the United States, where high-quality data are available (e.g., Stage IV data sets at 4 km), and apply it to other regions at similar resolution. This depends on how well the statistical relationships between covariates and fine-scale precipitation transfer to another location and so presumably depends on selecting training data from a climatically similar region. The algorithm could also be applied to downscale precipitation at the continental or global scale, using techniques such as moving window approach and massive parallel computing to overcome the large computational cost of such a task.

Acknowledgments

X. He would like to thank M. Pan, E. F. Wood, and I. Rodríguez-Iturbe at Princeton University for helpful discussions. This study is supported by NOAA grant NA14OAR4310218, and NSF grant 1534544. M. Schleiss acknowledges the financial support of the Swiss National Science Foundation (grant P300P2_158499). The NLDAS-2 data sets are provided by the NASA Goddard Earth Sciences Data and Information Services Center at http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings.

References

Abatzoglou, J. T. (2013), Development of gridded surface meteorological data for ecological applications and modelling, *Int. J. Climatol.*, *33*(1), 121–131.

Bárdossy, A., and G. Pegram (2011), Downscaling precipitation using regional climate models and circulation patterns toward hydrology, *Water Resour. Res.*, *47*, W04505, doi:10.1029/2010WR009689.

Bastola, S., and V. Misra (2014), Evaluation of dynamically downscaled reanalysis precipitation data for hydrological application, *Hydrol. Processes*, *28*(4), 1989–2002.

Berg, P., C. Moseley, and J. O. Haerter (2013), Strong increase in convective precipitation in response to higher temperatures, *Nat. Geosci.*, *6*(3), 181–185.

Beuchat, X., B. Schaefli, M. Soutter, and A. Mermoud (2011), Toward a robust method for subdaily rainfall downscaling from daily data, *Water Resour. Res.*, *47*, W09524, doi:10.1029/2010WR010342.

Breiman, L. (1996), Bagging predictors, *Mach. Learn.*, *24*(2), 123–140.

Breiman, L. (2001), Random forests, *Mach. Learn.*, *45*(1), 5–32.

Carlisle, D. M., J. Falcone, D. M. Wolock, M. R. Meador, and R. H. Norris (2010), Predicting the natural flow regime: Models for assessing hydrological alteration in streams, *River Res. Appl.*, *26*(2), 118–136.

Carter, T. R., M. L. Parry, H. Harasawa, and S. Nishioka (1994), IPCC technical guidelines for assessing climate change impacts and adaptions, in *IPCC SPECIAL report to Working Group II of IPCC*, Cent. for Global Environ. Res., Tsukaba, Japan.

Chaney, N. W., P. Metcalfe, and E. F. Wood (2016a), HydroBlocks: A field-scale resolving land surface model for application over continental extents, *Hydrol. Processes*, *30*, 3543–3559, doi:10.1002/hyp.10891.

Chaney, N. W., E. F. Wood, A. B. McBratney, J. W. Hempel, T. W. Nauman, C. W. Brungard, and N. P. Odgers (2016b), POLARIS: A 30-meter probabilistic soil series map of the contiguous United States, *Geoderma*, *274*, 54–67.

Cosgrove, B. A., et al. (2003), Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project, *J. Geophys. Res.*, *108*(D22), 8842, doi:10.1029/2002JD003118.

Coulibaly, P., Y. B. Dibike, and F. Anctil (2005), Downscaling precipitation and temperature with temporal neural networks, *J. Hydrometeorol.*, *6*(4), 483–496, doi:10.1175/JHM409.1.
Cowpertwait, P. S. (1995), A generalized spatial-temporal model of rainfall based on a clustered point process, *Proc. R. Soc. London Ser. A*, *450*, 163–175.
Cox, D., and V. Isham (1988), A simple spatial-temporal model of rainfall, *Proc. R. Soc. London Ser. A*, *415*, 317–328.
Criminisi, A., J. Shotton, and E. Konukoglu (2011), Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning, *Tech. Rep. MSRTR-2011-114*, vol. 5(6), p. 12, Microsoft Res., Cambridge.
Davy, R. J., M. J. Woods, C. J. Russell, and P. A. Coppin (2010), Statistical downscaling of wind variability from meteorological fields, *Boundary Layer Meteorol.*, *135*(1), 161–175.
Domingos, P. (2012), A few useful things to know about machine learning, *Commun. ACM*, *55*(10), 78–87.
Ebtehaj, A. M., E. Foufoula-Georgiou, and G. Lerman (2012), Sparse regularization for precipitation downscaling, *J. Geophys. Res.*, *117*, D08107, doi:10.1029/2011JD017057.
Eccel, E., L. Ghielmi, P. Granitto, R. Barbiero, F. Grazzini, and D. Cesari (2007), Prediction of minimum temperatures in an alpine region by linear and non-linear post-processing of meteorological models, *Nonlinear Process. Geophys.*, *14*(3), 211–222.
Efron, B. (1979), Bootstrap methods: Another look at the jack knife, *Ann. Stat.*, *7*, 1–26.
Foufoula-Georgiou, E., A. Ebtehaj, S. Zhang, and A. Hou (2014), Downscaling satellite precipitation with emphasis on extremes: A variational ℓ-norm regularization in the derivative domain, *Surv. Geophys.*, *35*(3), 765–783.
Fowler, H. J., and R. L. Wilby (2007), Beyond the downscaling comparison study, *Int. J. Climatol.*, *27*(12), 1543–1545.
Friedman, J., T. Hastie, and R. Tibshirani (2001), *The Elements of Statistical Learning*, vol. 1, *Springer Series in Statistics*, Springer, Berlin.
Guirguis, K. J., and R. Avissar (2008), A precipitation climatology and dataset intercomparison for the western United States, *J. Hydrometeorol.*, *9*(5), 825–841.
Haerter, J., P. Berg, and S. Hagemann (2010), Heavy rain intensity distributions on varying time scales and at different temperatures, *J. Geophys. Res.*, *115*, D17102, doi:10.1029/2009JD013384.
Hautaniemi, S., S. Kharait, A. Iwabu, A. Wells, and D. A. Lauffenburger (2005), Modeling of signal–response cascades using decision tree analysis, *Bioinformatics*, *21*(9), 2027–2035.
He, X., T. Zhao, and D. Yang (2013), Prediction of monthly inflow to the Danjiangkou reservoir by distributed hydrological model and hydro-climatic teleconnections, *J. Hydroelect. Eng.*, *32*(3), 4–9.
He, X., Y. Hong, H. Vergara, K. Zhang, P.-E. Kirstetter, J. J. Gourley, Y. Zhang, G. Qiao, and C. Liu (2016), Development of a Coupled Hydrological-geotechnical Framework for Rainfall-induced Landslides Prediction, *J. Hydrol.*, doi:http://dx.doi.org/10.1016/j.jhydrol.2016.10.016
Hershenhorn, J., and D. Woolhiser (1987), Disaggregation of daily rainfall, *J. Hydrol.*, *95*(3), 299–322.
Higgins, R. W., J. E. Janowiak, and Y.-P. Yao (1996), *A gridded hourly precipitation data base for the United States (1963-1993)*, U.S. Dep. of Commer., Natl. Oceanic and Atmos. Admin., Natl. Weather Serv., Camp Springs, Md.
Hong, Y., R. F. Adler, A. Negri, and G. J. Huffman (2007), Flood and landslide applications of near real-time satellite rainfall products, *Nat. Hazards*, *43*(2), 285–294.
Hostetler, S. (1994), Hydrologic and atmospheric models: The (continuing) problem of discordant scales, *Clim. Change*, *27*(4), 345–350.
Ibarra-Berastegi, G., J. Saénz, A. Ezcurra, A. Elías, J. Diaz Argandoña, and I. Errasti (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, *Hydrol. Earth Syst. Sci.*, *15*(6), 1895–1907, doi:10.5194/hess-15-1895-2011.
Jeong, D. I., A. St-Hilaire, T. B. Ouarda, and P. Gachon (2012), CGCM3 predictors used for daily temperature and precipitation downscaling in Southern Québec, Canada, *Theor. Appl. Climatol.*, *107*(3-4), 389–406.
Jha, S. K., G. Mariethoz, J. Evans, M. F. McCabe, and A. Sharma (2015), A space and time scale-dependent nonlinear geostatistical approach for downscaling daily precipitation and temperature, *Water Resour. Res.*, *51*, 6244–6261, doi:10.1002/2014WR016729.
Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie (2004), CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution, *J. Hydrometeorol.*, *5*(3), 487–503.
Koutsoyiannis, D. (2010), HESS Opinions "A random walk on water?," *Hydrol. Earth Syst. Sci.*, *14*(3), 585–601, doi:10.5194/hess-14-585-2010.
Kühnlein, M., T. Appelhans, B. Thies, and T. Nauß (2014), Precipitation estimates from MSG SEVIRI daytime, nighttime, and Twilight data with random forests, *J. Appl. Meteorol. Climatol.*, *53*(11), 2457–2480.
Kumar, P., and E. Foufoula-Georgiou (1993), A multicomponent decomposition of spatial rainfall fields: 1. segregation of large-and small-scale features using wavelet transforms, *Water Resour. Res.*, *29*(8), 2515–2532.
Lenderink, G., and E. Van Meijgaard (2008), Increase in hourly precipitation extremes beyond expectations from temperature changes, *Nat. Geosci.*, *1*(8), 511–514.
Liaw, A., and M. Wiener (2002), Classification and regression by random forest, *R News*, *2*(3), 18–22.
Lovejoy, S., and B. Mandelbrot (1985), Fractal properties of rain, and a fractal model, *Tellus Ser. A*, *37*, 209–232.
Lovejoy, S., D. Schertzer, and A. Tsonis (1987), Functional box-counting and multiple elliptical dimensions in rain, *Science*, *235*(4792), 1036–1038.
Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers (2011), *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, McKinsey Global Institute.
Maraun, D., et al. (2010), Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Rev. Geophys.*, *48*, RG3003, doi:10.1029/2009RG000314.
Mason, S. J., and N. E. Graham (1999), Conditional probabilities, relative operating characteristics, and relative operating levels, *Weather Forecast.*, *14*(5), 713–725.
Menabde, M., D. Harris, A. Seed, G. Austin, and D. Stow (1997), Multiscaling properties of rainfall and bounded random cascades, *Water Resour. Res.*, *33*(12), 2823–2830.
Olsson, J., C. B. Uvo, and K. Jinno (2001), Statistical atmospheric downscaling of short-term extreme rainfall by networks, *Phys. Chem. Earth Part B*, *26*(9), 695–700.
Onof, C., and H. S. Wheater (1993), Modelling of British rainfall using a random parameter Bartlett-Lewis rectangular pulse model, *J. Hydrol.*, *149*(1), 67–95.
Onof, C., R. Chandler, A. Kakou, P. Northrop, H. Wheater, and V. Isham (2000), Rainfall modelling using Poisson-cluster processes: A review of developments, *Stochastic Environ. Res. Risk Assess.*, *14*(6), 384–411.
Pan, M., et al. (2003), Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent, *J. Geophys. Res.*, *108*(D22), 8850, doi:10.1029/2003JD003994.
Pebesma, E. J. (2004), Multivariable geostatistics in S: The gstat package, *Comput. Geosci.*, *30*, 683–691.
Pedregosa, F., et al. (2011), Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, *12*, 2825–2830.

Pegram, G., and A. Clothier (2001), High resolution space–time modelling of rainfall: The "string of beads" model, *J. Hydrol.*, *241*(1), 26–41.
Peters, J., B. De Baets, N. E. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts (2007), Random forests as a tool for ecohydrological distribution modelling, *Ecol. Modell.*, *207*(2), 304–318.
Rebora, N., L. Ferraris, J. von Hardenberg, and A. Provenzale (2006), RainFARM: Rainfall downscaling by a filtered autoregressive model, *J. Hydrometeorol.*, *7*(4), 724–738.
Rodríguez-Iturbe, I., D. Cox, and V. Isham (1987), Some models for rainfall based on stochastic point processes, *Proc. R. Soc. London Ser. A*, *410*, 269–288.
Rodríguez-Iturbe, I., D. Cox, and V. Isham (1988), A point process model for rainfall: Further developments, *Proc. R. Soc. London Ser. A*, *417*, 283–298.
Roe, G. H. (2005), Orographic precipitation, *Annu. Rev. Earth Planet. Sci.*, *33*, 645–671.
Rummukainen, M. (2010), State-of-the-art with Regional Climate Models, *WIREs Clim. Change*, *1*(1), 82–96.
Schleiss, M., and A. Berne (2012), Stochastic space–time disaggregation of rainfall into DSD fields, *J. Hydrometeorol.*, *13*(6), 1954–1969.
Schleiss, M., S. Chamoun, and A. Berne (2014), Nonstationarity in Intermittent Rainfall: The Dry Drift?, *J. Hydrometeorol.*, *15*(3), 1189–1204.
Sheffield, J., et al. (2014), A drought monitoring and forecasting system for sub-Sahara African water resources and food security, *Bull. Am. Meteorol. Soc.*, *95*(6), 861–882.
Shi, Y., and L. Song (2015), Spatial downscaling of monthly TRMM precipitation based on EVI and other geospatial variables over the Tibetan Plateau From 2001 to 2012, *Mt. Res. Dev.*, *35*(2), 180–194.
Singh, R., K. Ranjan, and H. Verma (2015), Satellite imaging and surveillance of infectious diseases, *J. Trop Dis.*, *S1-004*, 1–6.
Utsumi, N., S. Seto, S. Kanae, E. E. Maeda, and T. Oki (2011), Does higher surface temperature intensify extreme precipitation?, *Geophys. Res. Lett.*, *38*, L16708, doi:10.1029/2011GL048426.
Vrac, M., and P. Naveau (2007), Stochastic downscaling of precipitation: From dry events to heavy rainfalls, *Water Resour. Res.*, *43*, W07402, doi:10.1029/2006WR005308.
Wood, E. F., et al. (2011), Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, *Water Resour. Res.*, *47*, W05301, doi:10.1029/2010WR010090.
Xia, Y., et al. (2012), Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products, *J. Geophys. Res.*, *117*, D03109, doi:10.1029/2011JD016048.
Yano, J.-I. (2010), Downscaling, parameterization, decomposition, compression: A perspective from the multiresolution analysis, *Adv. Geosci.*, *23*(23), 65–71.
Zhan, W., M. Pan, N. Wanders, and E. F. Wood (2015), Correction of real-time satellite precipitation with satellite soil moisture observations, *Hydrol. Earth Syst. Sci.*, *19*(10), 4275–4291, doi:10.5194/hess-19-4275-2015.
Zhang, X.-C., M. Nearing, J. Garbrecht, and J. Steiner (2004), Downscaling monthly forecasts to simulate impacts of climate change on soil erosion and wheat production, *Soil Sci. Soc. Am. J.*, *68*(4), 1376–1385.
Zhao, T., D. Yang, X. Cai, and Y. Cao (2012), Predict seasonal low flows in the upper Yangtze River using random forests model, *J. Hydroelect. Eng.*, *31*(3), 18–38.