




RESEARCH ARTICLE

Comprehensive evaluation of an improved large-scale multi-site weather generator for Germany

Viet Dung Nguyen¹  | Bruno Merz^{1,2}  | Yeshewatesfa Hundecha³ |
Uwe Haberlandt⁴  | Sergiy Vorogushyn¹ 

¹GFZ German Research Centre for Geosciences, Section Hydrology, Potsdam, Germany

²Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany

³Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

⁴Institute of Hydrology and Water Resources Management, Leibniz University of Hannover, Hannover, Germany

Correspondence

Viet Dung Nguyen, GFZ German Research Centre for Geosciences, Section Hydrology, 14473 Potsdam, Germany.
Email: dung@gfz-potsdam.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: FLOOD (project number 01LP1903E); Deutsche Forschungsgemeinschaft, Grant/Award Number: SPATE (project number 278017089)

Abstract

In this work, we present a comprehensive evaluation of a stochastic multi-site, multi-variate weather generator at the scale of entire Germany and parts of the neighbouring countries covering the major German river basins Elbe, Upper Danube, Rhine, Weser and Ems with a total area of approximately 580,000 km². The regional weather generator, which is based on a first-order multi-variate auto-regressive model, is setup using 53-year long daily observational data at 528 locations. The performance is evaluated by investigating the ability of the weather generator to replicate various important statistical properties of the observed variables including precipitation occurrence and dry/wet transition probabilities, mean daily and extreme precipitation, multi-day precipitation sums, spatial correlation structure, areal precipitation, mean daily and extreme temperature and solar radiation. We explore two marginal distributions for daily precipitation amount: mixed Gamma-Generalized Pareto and extended Generalized Pareto. Furthermore, we introduce a new procedure to estimate the spatial correlation matrix and model mean daily temperature and solar radiation. The extensive evaluation reveals that the weather generator is greatly capable of capturing most of the crucial properties of the weather variables, particularly of extreme precipitation at individual locations. Some deficiencies are detected in capturing spatial precipitation correlation structure that leads to an overestimation of areal precipitation extremes. Further improvement of the spatial correlation structure is envisaged for future research. The mixed marginal model found to outperform the extended Generalized Pareto in our case. The use of power transformation in combination with normal distribution significantly improves the performance for non-precipitation variables. The weather generator can be used to generate synthetic event footprints for large-scale trans-basin flood risk assessment.

KEYWORDS

correlation, extreme, flood, large-scale, multi-variate, weather generator

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *International Journal of Climatology* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

1 | INTRODUCTION

Reliable flood risk assessments are challenging, in particular at large spatial scales (de Moel *et al.*, 2015). There is a lack of high-resolution and long-term observational climate data to drive risk assessment models (Ward *et al.*, 2015). On the other hand, process complexity and interactions, for instance between different locations within a river basin, are increasingly difficult to capture with increasing spatial scales (Vorogushyn *et al.*, 2018). In particular, the spatially coherent representation of meteorological fields used to drive hydrological models for flood frequency and risk assessment is a pressing issue.

Derived Flood Frequency Analysis (DFFA), originally introduced by Eagleson (1972) as event-based approach, transforms precipitation distributions into flood peak distributions. Whereas Eagleson (1972) used functional relationship between precipitation and discharge distributions, in the past two decades a continuous simulation approach to DFFA has been increasingly used (Blazkova and Beven, 1997, 2004; Cameron *et al.*, 1999; Grimaldi *et al.*, 2012a, 2012b; Haberlandt and Radtke, 2014). For this, synthetically generated precipitation time series are transformed into discharge series using continuous hydrological modelling. Analogously, Derived Flood Risk Analysis (DFRA) (Falter *et al.*, 2015; Falter *et al.*, 2016; Metin *et al.*, 2018) has emerged to quantify flood damage and risk from continuous model simulations. Both approaches are based on the generation of synthetic, spatially consistent fields of precipitation, temperature and other meteorological variables (e.g., humidity, solar radiation). These long-term continuous time series (in the range of several 1,000 years) are used to drive hydrological models to obtain synthetic discharge series for subsequent flood frequency analysis or inundation modelling and risk assessment. To provide long-term spatially consistent meteorological fields, multi-site stochastic models—weather generators—are typically used. A strong skill of capturing the characteristics of extreme precipitation is indispensable for estimation of flood hazard and risk.

Stochastic multi-variate weather models were originally developed for single sites, where there is no need for considering the spatial correlation structure of weather variables (e.g., Richardson, 1981; Rajagopalan and Lall, 1999). Both multi-variate auto-regressive models (Richardson, 1981) and nearest-neighbour resampling techniques in the multi-variate space (Rajagopalan and Lall, 1999) were developed. To cover larger spatial scales than those characterized by a single location, multi-site weather generators have been proposed in the recent decades. Multi-site stochastic models

are based on (a) *resampling techniques*, such as analogue (Zorita and von Storch, 1999; Chardon *et al.*, 2018; Raynaud *et al.*, 2020) and nearest-neighbour (Beersma and Buishand, 2003; Caraway *et al.*, 2014), (b) *point-process* simulation of rainfall in combination with Markov chain simulation of precipitation occurrence (Fowler *et al.*, 2005; Cowpertwait, 2006) and (c) *time series generation* using the Markov-chain approach in combination with frequency distributions (Wilks, 1998; Breinl *et al.*, 2013; Keller *et al.*, 2015) to simulate occurrence and magnitude, respectively. Furthermore, latent Gaussian variable models in combination with auto-regressive approach considering the spatial correlation structure are employed in Bárdossy and Plate (1992), Hundedcha *et al.* (2009), Kleiber *et al.* (2012), Rasmussen (2013) and Bennett *et al.* (2018). Finally, a combination of Markov chain, generalized additive model and meta-Gaussian random fields is proposed by Serinaldi and Kilsby (2014). An overview of weather generators is provided by Haberlandt *et al.* (2011) and Serinaldi and Kilsby (2014). Here, we briefly revisit major types of weather generators and discuss their merits with focus on DFRA.

Weather generators based on resampling techniques produce a sequence of synthetic weather variables based on reshuffling of previous observations conditioned on certain rules, for example, restricting the occurrence of an extremely cold day in the middle of the heat wave. In a multi-site approach, observations at all sites (climate stations or grid points) are reshuffled for different time steps (e.g., days or hours). Thus, resampling techniques fully preserve spatial correlations between various locations within a single time-step and thus can be easily applied to both small scales as well as to large river basins (Beersma and Buishand, 2003; Raynaud *et al.*, 2020). Standard resampling techniques are generally not able to produce single-day extreme precipitation beyond those observed in the past. Flood-relevant extremes can potentially be generated by multi-day precipitation sums. Beersma and Buishand (2003) demonstrate the ability of a nearest-neighbour resampling model conditioned on weather patterns to generate 10-day extreme precipitation exceeding the observed maxima. Sharif and Burn (2007) propose a perturbation technique in combination with the k-nearest resampling while still retaining the spatial correlation structure and major statistical properties of at-site precipitation. Finally, Raynaud *et al.* (2020) propose a weather generator based on constructing plausible atmospheric trajectories, that is, series of atmospheric states, based on analogues in combination with distribution adjustments to generate unobserved, but plausible series of precipitation and temperature.

Point process precipitation models simulate the space–time distribution of precipitation by applying the rectangular-pulse models for precipitation occurrence, intensity, and duration of spells. Spatial patterns emerge through generating the rain cells in space and aggregating them into storm clusters (Fowler *et al.*, 2000; Fowler *et al.*, 2005). Although the space–time rainfall process has strong physical basis, it comes with a price. The estimation of parameters of these models can be difficult in view of limited observations (Haberlandt *et al.*, 2011). Large-scale applications seem to be difficult using this approach, and it is not often used for flood assessments.

Time series modelling of precipitation at multiple locations can be achieved through Markov-Chain modelling of wetness state (wet/dry; Wilks, 1998; Breinl *et al.*, 2013; Keller *et al.*, 2015) or a multi-variate latent auto-regressive modelling of precipitation series (Bárdossy and Plate, 1992; Hundecha *et al.*, 2009; Rasmussen, 2013; Bennett *et al.*, 2018) in combination with a frequency distribution of precipitation magnitudes. Techniques based on fitting frequency distributions to observed precipitation can extrapolate daily precipitation amounts beyond observations. They may, however, suffer from sampling uncertainties and over-parameterization. Especially, mixed models combining two distributions (Hundecha *et al.*, 2009; Baxevani and Lennartsson, 2015) may face the latter problem given short observational series. On the contrary, these models seem to better capture extreme precipitation, whose distribution may differ from the bulk of records (Hundecha *et al.*, 2009).

Precipitation occurrence and amounts at various locations are correlated. The observed spatial correlation structure should be captured by a weather generator and imposed on the synthetically generated data. This is an additional source of uncertainty, particularly with increasing spatial scale where correlations decay with the inter-site distance. Several approaches are developed to consider the spatial correlation structure in time series models. Breinl *et al.* (2013) proposed a univariate Markov process for multi-site precipitation occurrence, which retains the spatial correlation structure of precipitation occurrence, but can only reproduce occurrence vectors observed in the past. The precipitation amounts sampled from the fitted parametric distributions are reshuffled according to the ranks of the resampled observations in order to sustain the spatio-temporal correlation of observed series. Serinaldi and Kilsby (2014) used discrete-continuous distributions to describe the occurrence and amount of rainfall at a site simultaneously. Furthermore, the spatial correlation structure is considered taking the observed covariance into account, which is used to condition the Gaussian random fields. Finally, the daily

rainfall fields are generated by applying the at-site distributions to the Gaussian random fields. The approach is applied to the Danube basin and parameterized using $0.25^\circ \times 0.25^\circ$ gridded E-OBS precipitation dataset (Haylock *et al.*, 2008). The model performs reasonably well in terms of various criteria of model performance at single sites. However, it tends to overestimate extreme daily precipitation sums at single locations and overestimates the areal rainfall aggregated over sub-basin area. Recently, Sparks *et al.* (2018) proposed a combination of the autoregressive model with lag-1 (AR(1)) with extended Empirical Orthogonal Functions (EOFs), where precipitation is modelled as a censored latent Gaussian process. The method is applied at the scale of Europe based on the coarsely gridded E-OBS precipitation dataset (Haylock *et al.*, 2008) interpolated to the $\sim 1.3^\circ \times 1.6^\circ$ grid. The spatial correlation of observed precipitation seems to be relatively low for most cells (<0.5) in this dataset. The correlation structure of highly correlated locations appears to be strongly underestimated.

The overview of the weather generation approaches shows that point process models are not often used for large-scale flood assessments due to their complexity. The two other types – resampling techniques and time series models—are applied at various scales and can be used to drive hydrological models and risk model chains for risk estimation. However, large-scale applications of weather generators to areas above $150,000 \text{ km}^2$ are rare. Resampling approaches are easily scalable and retain the observed spatial correlation structure, but may suffer from the inherent limitation imposed by the observed time series constraining the generation of extremes. Time series models are not limited by upper bounds of precipitation magnitudes but face challenges of capturing the spatial correlation structure of rainfall fields for large spatial domains (Serinaldi and Kilsby, 2014; Sparks *et al.*, 2018). Further, different approaches to evaluate the spatial performance of weather generators, such as simple pairwise correlation (Wilks, 1998; Sparks *et al.*, 2018), continuity ratio (Wilks, 1998; Breinl *et al.*, 2013), frequency plots of the aggregated domain precipitation at various time scales (Kleiber *et al.*, 2012; Serinaldi and Kilsby, 2014; Baxevani and Lennartsson, 2015; Sparks *et al.*, 2018) and frequency plots of the mean precipitation for n wettest locations (Serinaldi and Kilsby, 2014), are used, which makes the comparison across weather generators extremely difficult. Furthermore, some studies use relatively lax criteria for performance evaluation with regards to extremes (e.g., Sparks *et al.*, 2018), which makes it difficult to gain credibility in risk estimates. Thus, we see an urgent need for comprehensive evaluation studies of large-scale weather generators using comparable performance statistics to identify the most

promising approaches for flood risk assessment. Recently, Bennett *et al.* (2018) proposed a comprehensive and systematic framework to evaluate the performance of weather generators over a range of spatial and temporal scales. The framework includes a systematic categorization of performance results, which enables a clear identification of strengths and weaknesses and makes the performance comparable across various models. Also, Bennett *et al.* (2018) point out to the need of benchmarking of weather generators. They, however, acknowledge a limitation for comparison studies posed by applications in different areas/catchments. Meanwhile, the framework was applied by Evin *et al.* (2018) to assess and compare the performance of the various modifications of the Wilk's weather generator (Wilks, 1998).

In this work, we present a comprehensive evaluation of a multi-site, multi-variate weather generator at the scale of entire Germany and parts of the neighbouring countries covering the major German river basin Elbe, Upper Danube, Rhine, Weser and Ems. We employ the multi-site auto-regressive weather generator developed by Hundsdoerfer *et al.* (2009) to simulate daily precipitation fields and other variables, such as temperature, relative humidity and solar radiation conditioned on the rainfall series. It has been originally developed and evaluated at relatively small spatial scales (Hundsdoerfer *et al.*, 2009) and applied for trend attribution studies (Hundsdoerfer and Merz, 2012), DFFA (Winter *et al.*, 2019) and DFRA (Falter *et al.*, 2015) in a few German and Austrian catchments. We introduce a number of modifications to the original model by Hundsdoerfer *et al.* (2009). (a) A new approach to compute the spatial correlation matrix including the correction procedure to obtain definite positive matrix is implemented. (b) A parsimonious extended Generalized Pareto distribution is implemented additionally to the original Gamma-Generalized Pareto mixture distribution. (c) A power transformation of non-precipitation data is introduced to improve their simulation. The stochastic model is extended here to the Central European domain using observed long-term time series from unprecedented large number of climate stations (528 stations). In this article, we focus on the performance of the model to replicate the statistical properties of observed precipitation, temperature and solar radiation. We employ the framework of Bennett *et al.* (2018) to categorize the performance statistics and make it comparable with other studies using this framework. Specifically, we investigate the ability of the model to capture extremes and the spatial dependence of precipitation over a large spatial domain. Spatial dependence is crucial, when it comes to the generation of realistic event footprints of large-scale trans-basin floods for risk assessment. To our knowledge, this is one of the largest

domains the weather generator of this type is applied to. We, therefore, particularly focus on evaluating the ability of the model to simulate the areal precipitation sums relevant for DFFA and DFRA. The weather generator is envisaged for application to the national-scale flood risk appraisal.

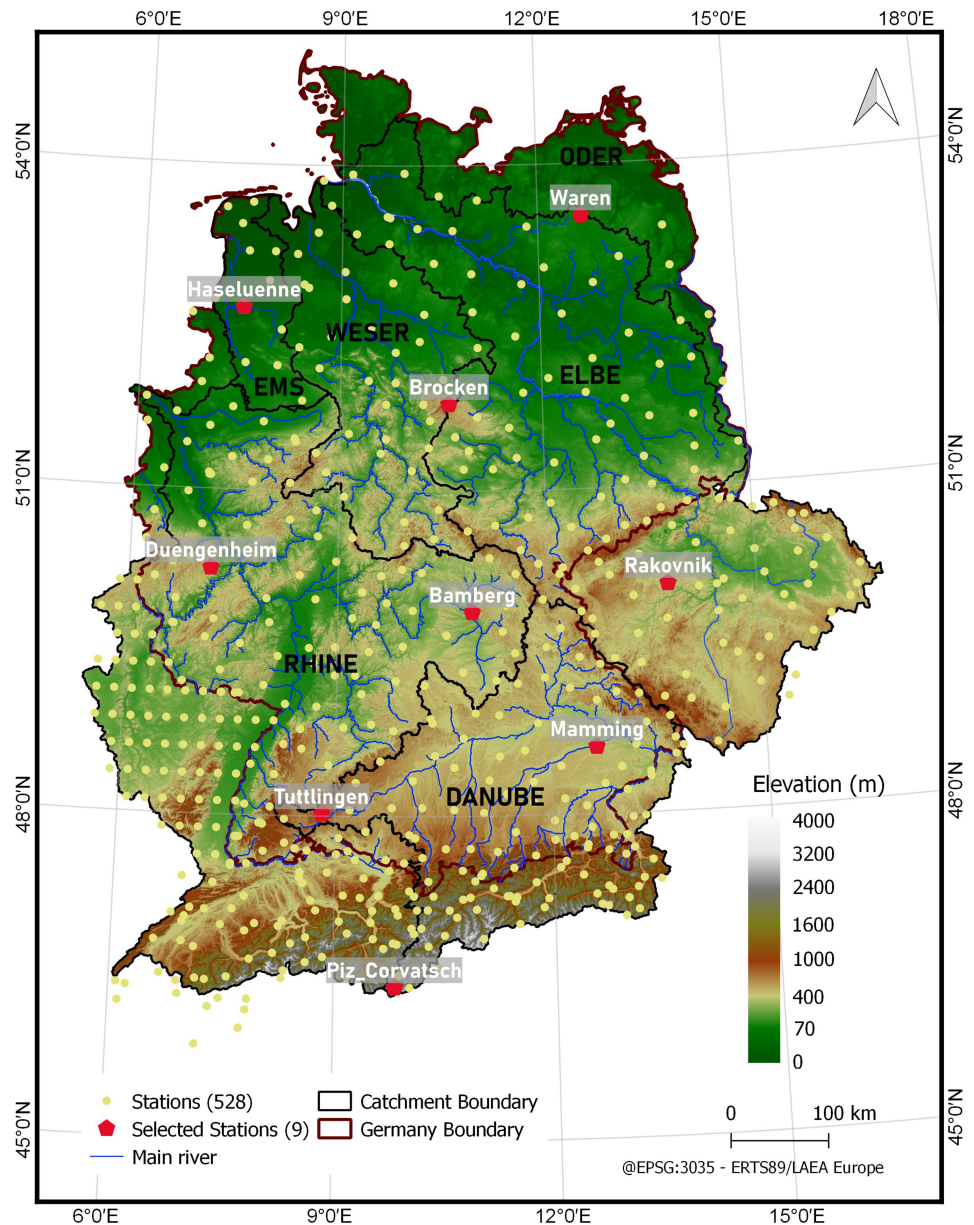
The article is organized as follows. First, an overview of the study region and data is given in Section 2 followed by a brief description of the weather generator. Model setup and evaluation strategy are discussed in Section 4. Results and discussion are detailed in Section 5 followed by the summary of the main findings.

2 | STUDY AREA AND DATA

The study region covers five large river basins, that is, Elbe, Rhine, Upper Danube, Weser and Ems, located in Germany and parts of Austria, Switzerland, Czech Republic and France (Figure 1). The total area comprises about 580,000 km². We use a dataset containing six daily observed variables of precipitation, temperature (max, min, mean), humidity and solar radiation at 528 locations in the 53-years period ranging from January 1951 to December 2003. The dataset is assembled from climate stations in Germany, Austria, Switzerland and Czech Republic, processed and harmonized at the Potsdam Institute for Climate Impact Research (Österle *et al.*, 2006a; Österle *et al.*, 2006b; Österle *et al.*, 2016). The station density is fairly uniform with somewhat higher density in the southern part of the study region characterized by higher elevations. Here a higher variability of climate variables is expected. No direct measurements are accessible for the French part of the Rhine basin. Here, we use the E-OBS gridded dataset (Haylock *et al.*, 2008) at the resolution of 0.25° x 0.25° and thin it out to 0.5° x 0.5° resolution in order to keep a comparable point density as in the remaining domain. Each grid point is treated as climate stations. Both, climate stations (465) and grid points (63) of the E-OBS dataset are referred to as 'stations' hereafter. Inconsistencies may arise from smoothed precipitation in a gridded dataset compared to the station-based records and should be kept in mind. This pragmatic step is taken in view of missing data and can be overturned as soon as gauge records are available.

Out of 528 stations, we chose a subset of nine representative stations for detailed evaluation of the weather generator performance. These stations are spread across the study region and are located along the altitudinal gradient from the Northern Lowlands to the High Alpine region. The climate variables across the 528 stations exhibit a clear seasonal cycle, in particular temperature, relative humidity and radiation (Figure 2). This

FIGURE 1 Study region covered by the weather generator including locations of the 528 stations used to parameterize the weather generator and 9 representative stations chosen for detailed validation [Colour figure can be viewed at wileyonlinelibrary.com]



highlights the need to consider seasonality for the parameterization of the weather generator. The values at the nine selected stations cover a wide range of observations in the entire study region.

The study region is characterized by a temperate climate with continentality increasing eastwards. The western and north-western regions are dominated by Atlantic air masses bringing long-lasting large-scale rainfall in winter and spring. Convective precipitation occurs in the summer half-year. The mean annual precipitation ranges between somewhat above 400 mm in north-east Germany and Czech Republic to more than 2000 mm in the Alpine region (Rauthe *et al.*, 2013). The daily extreme precipitation and temperatures show high variability within each month, whereas precipitation amounts are right-skewed.

3 | MULTI-SITE, MULTI-VARIATE STOCHASTIC WEATHER GENERATOR

The weather generator presented by Hundscha *et al.* (2009) simulates continuous series of daily precipitation at the station locations using the spatial correlation structure of the observations. Daily time series of other variables, such as mean daily temperature, relative humidity and solar radiation, are conditioned on the state of generated precipitation at each site (Hundscha and Merz, 2012). Both steps are based on the first-order multi-variate auto-regressive (MAR-1) model (Bárdossy and Plate, 1992; Wilks, 1999) which is described below in more details.

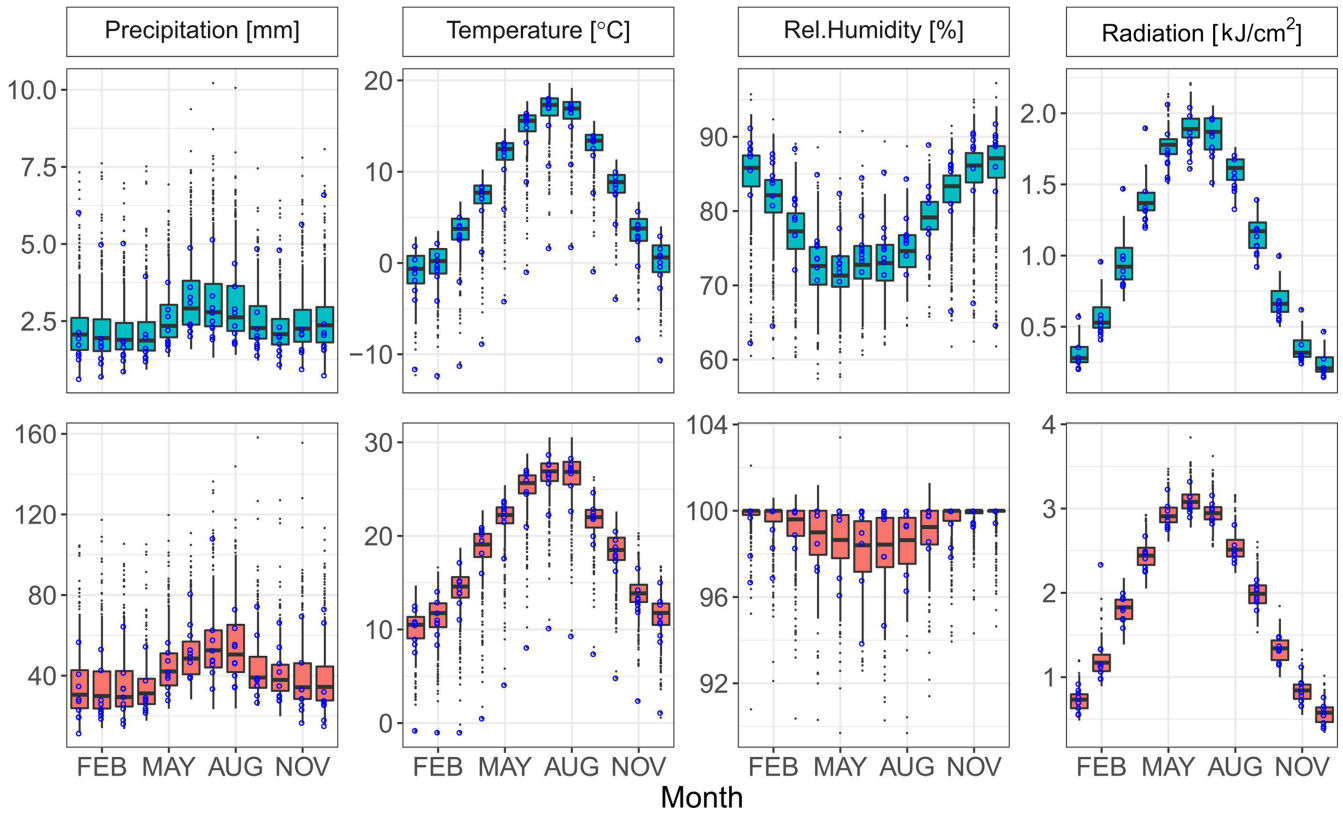


FIGURE 2 Seasonal distribution of mean (upper panel) and 99.9th percentile (lower panel) observed daily precipitation, temperature, relative humidity and solar radiation. Blue circles correspond to the 9 selected stations [Colour figure can be viewed at wileyonlinelibrary.com]

Let $W(t) = (W(t, u_1), \dots, W(t, u_n))$ be a multi-variate normal random vector of n locations $u = (u_1, \dots, u_n)$ at time t with the mean vector whose elements are zero.

$$W(t) = B_m W(t-1) + C_m \Psi(t) \tag{1}$$

$\Psi(t) = (\psi(t, u_1), \dots, \psi(t, u_n))$ is a random vector of independent standard normal variables. The matrices B_m and C_m are related to the lag-0 correlation matrix (M_{m0}) and the lag-1 correlation matrix (M_{m1}) through Equations (2) and (3), where the subscript m represents month from January to December.

$$B_m = M_{m1} M_{m0}^{-1} \tag{2}$$

$$C_m C_m^T = M_{m0} - B_m M_{m1}^T \tag{3}$$

is the transposed matrix of M_{m1} , C_m respectively.

The complete precipitation process (including wet and dry condition) x for month m at an individual station u is simulated using the distribution $H(x)$:

$$H(x) = (1-p) + pF(x) \tag{4}$$

where p represents the wet frequency, conversely $1-p$ stands for the probability of zero rainfall and $F(x)$ is the distribution of the non-zero precipitation amounts (when x is non-positive, $F(x)$ equals 0). The link between the marginal distribution of precipitation expressed in Equation (4) and the MAR-1 model expressed in Equation (1) is:

$$\Phi(W(t, u)) = H(x(t, u)) \tag{5}$$

where Φ stands for the cumulative distribution function of a standard normal distribution.

$F(x)$ is typically assumed to be one of the widely used distributions to describe rainfall amounts: Exponential, Gamma, Weibull or Generalized Pareto (GP). Exponential and Gamma distributions exhibit light tails and are found inferior to describe extreme rainfall amounts (Serinaldi and Kilsby, 2014; Baxevani and Lennartsson, 2015), which typically show heavy-tail behaviour (Papalexiou *et al.*, 2013; Rajah *et al.*, 2014). Whereas heavy-tailed distributions like Weibull (with shape

parameter below 1) and GP are suitable to capture extreme values, they may miss the lower bulk of precipitation amounts. Hence, mixed (two-component) distributions have been used to exploit the advantages of both groups (Frigessi *et al.*, 2002; Vrac and Naveau, 2007; Furrer and Katz, 2008; Li *et al.*, 2012, Baxevani and Lennartsson, 2015). Our weather generator is based on the mixed Gamma-GP distribution (mGGP), shown to outperform the Gamma distribution at stations in Central Germany (Hundecha *et al.*, 2009). The higher flexibility of the mixed distribution comes with the cost of increasing model complexity and number of model parameters. The mixed Gamma-GP distribution has in total 6 parameters of which two come from the gamma distribution, two come from the GP distribution and two from the dynamically mixing function (see Hundecha *et al.*, 2009 for more details).

Fitting this 6-parameter mixed distribution to precipitation data is not a trivial task since the normalization constant should be numerically computed with very high precision (see Li *et al.*, 2012 for more details). We use the global optimization algorithm SCE-UA (Duan *et al.*, 1992) to optimize the log-likelihood function for parameter estimation. The distribution is fitted to each station on a monthly basis to account for seasonality.

Naveau *et al.* (2016) point out to a number of drawbacks of the mixed distribution. In particular, the mixture models may become overparameterized. They propose a more parsimonious extended Generalized Pareto distribution (extGP) which helps move away from the mixed distribution concept but still allows a smooth transition between bulk of the distribution and the heavy tails. Also, Papalexiou and Koutsoyiannis (2013) suggest more parsimonious distributions for precipitation originating from the Generalized Beta distribution family of the second type, that is, three-parameter Generalized Gamma and Burr Type XII distributions. In this study, we explore the 3-parameter extGP distribution (Naveau *et al.*, 2016) (Equation 6) and compare it to the original Gamma-GP mixture distribution. The maximum likelihood method is used to fit the extGP type 1 to station data on a monthly basis.

$$F(x) = \left(GP_{\xi} \left(\frac{x}{\sigma} \right) \right)^{\kappa} \quad (6)$$

where κ controls the shape of the lower tail, σ is a scale parameter and ξ controls the rate of upper tail decay.

To estimate the correlation matrices M_{m0} and M_{m1} in Equations (2) and (3), Hundecha *et al.* (2009) and Hundecha and Merz (2012) used the method described in Bárdossy and Plate (1992) which requires the indicator correlations estimated from the quantiles of $W(t, u)$. In this study, we estimate M_{m0} and M_{m1} through Kendall correlation and then transformed into Pearson's correlation as,

for example, applied by Serinaldi and Kilsby (2014). The two methods give very similar estimation results but the latter is mathematically much more convenient.

In practice, because of potential numerical inaccuracies in computation or missing data, large matrices M_{m0} are sometimes poorly defined (not positive definite) and hence not invertible. Correction is therefore needed to derive B_m and C_m from Equations (2) and (3). One common way to correct the matrices is to apply spatial correlation functions to fit the original correlation matrix (Bárdossy and Plate, 1992; Wilks, 1999; Hundecha *et al.*, 2009; Serinaldi and Kilsby, 2014). However, in many cases, applying this functional approach can result in a different correlation value compared to the original one. In Hundecha *et al.* (2009), we observe pronounced overestimation in the corrected M_{m0} . Therefore, in this study, we use a method of Higham (2002) to find the nearest positive definite correlation matrix of M_{m0} .

Daily temperature is described by a normal distribution in the original model by Hundecha and Merz (2012), whereas for solar radiation a square root transformation is applied prior to fitting a normal distribution. However, we observe that these assumptions are not universally valid for the non-precipitation data. We found often deviations from the normality assumptions at many stations and all months. In this study, we, therefore, attempt to improve the fitting by first applying power transformations in which a positive exponent is selected to minimize the skewness of the non-precipitation data for individual stations and for each month.

Two normal distributions conditioned on the wet/dry state are fitted to each variable and for each month. The variables are simulated also using a multi-variate autoregressive model. The workflow of the employed weather generator is provided in Figure 3 of Hundecha and Merz (2012).

4 | MODEL SETUP AND EVALUATION FOR THE REGIONAL DATASET – REGIONAL WEATHER GENERATOR (RWG)

The weather generator is setup for the presented study area, and this setup termed 'Regional Weather Generator' (RWG) is calibrated and evaluated on a monthly basis using the climate station dataset. We setup three versions of RWG as described below.

RWG0-mGGP is the original model version using the mixed Gamma-GP distribution as the marginal model for precipitation, normal distribution for the non-precipitation variables and additionally the square root transformation for solar radiation (Hundecha and Merz, 2012). In the

RWG0-extGP model version a parsimonious marginal extGP type 1 distribution is used with all other components kept unchanged. Finally, RWG1- extGP incorporates all presented changes including (a) the extGP marginal model, (b) a new procedure to compute and correct the correlation matrices and (c) the use of the power transformation prior to fitting a normal distribution to the non-precipitation variables.

All versions of the RWG are calibrated for all six climate variables at 528 stations and for 12 months. We generate 100 realizations with a time series length of 53 years, that is, the same length as the observations, and compare a range of statistical metrics for synthetic and observed climate. This procedure is typically applied for evaluation of stochastic weather models (Kleiber *et al.*, 2012; Breinl *et al.*, 2013; Serinaldi and Kilsby, 2014; Baxevani and Lennartsson, 2015).

We validate the weather generator with regards to at-site performance at all stations across Germany. However, more detailed results are only shown for the nine representatively selected stations. The simulation of areal precipitation is challenging for stochastic weather models, especially over very large domains, as in this study. This is because when the scale of the study area becomes larger, the number of stations (or grid points) should also increase to maintain the adequate spatial resolution which is necessary to represent the spatial variability at that scale. A coarse network leads to a poor representation of the spatial variability while a dense network often makes the multi-variate problem computationally intractable.

For the at-site evaluation of precipitation, we analyse the intermittence probabilities, mean and extremes at monthly and daily scales as well as multi-day sums. To evaluate the spatial representation of precipitation fields, we look at the monthly plots of pairwise correlation as a function of inter-site distance for both the bulk of the data and extreme precipitation. Additionally, the total precipitation sums at N stations within a certain radius from a selected station is evaluated. With regards to the conditioned variables, we exemplarily present the model performance for mean temperature and solar radiation. These are essential variables used to calculate potential evapotranspiration by hydrological models. In the following, the selected performance criteria are discussed in more details.

We assess the performance of the RWG in reproducing the precipitation intermittence using wet frequencies and (wet/dry) transition probabilities. Wet frequency is defined as the fraction of wet days at a station location. We consider days to be dry if the recorded precipitation is below 0.1 mm.

In the next step, the RWG performance to generate reasonable at-site precipitation amounts is investigated considering daily and monthly time scales. We focus on

daily and monthly mean precipitation and on the 99.9th percentile for daily precipitation and 99th percentile of the monthly sums of precipitation. The percentiles are computed using Weibull plotting positions. The 99th percentile of the monthly sums is extrapolated from 53 years of data using the semi-parametric quantile estimation proposed by Hutson (2002). Daily intensities are important for generation of single runoff events, whereas the monthly amounts control soil moisture and total catchment storage, which influence flood generation as well. Additionally, the n -day maxima of simulated precipitation for $n = 5$ and 10 days are compared to the observed statistics to analyse the plausibility of wet-spell amounts. We consider these durations to be important to generate flood events by single cyclones and flood events resulting from two subsequent storms, with the first one contributing to the catchment wetness. Beersma and Buishand (2003) also used 10-days accumulated precipitation to evaluate a resampling weather generator.

To quantify the ability of the RWG to simulate plausible areal rainfall amounts, we assess the spatial correlation structure of precipitation. For this, we analyse the monthly correlation functions for the bulk and extreme precipitation. First, we calculate the Kendall's tau correlation and transform this into the Pearson correlation for the bulk of the precipitation in each month. This procedure reduces the effect of outliers. To assess the correlation structure of extreme rainfall, we introduce the 80th percentile threshold to daily precipitation at each station. The correlation between the series above this threshold at one station and the corresponding daily precipitation (not necessarily above the threshold) at other stations is computed. Pairwise correlations are evaluated for each station. It should be noted that $\text{corr}(P_{s1}|p(P_{s1} < 0.2), P_{s2}|p(P_{s1} < 0.2))$ is not necessarily equal to $\text{corr}(P_{s2}|p(P_{s2} < 0.2), P_{s1}|p(P_{s2} < 0.2))$, where P_{s1} and P_{s2} are precipitation series at stations $s1$ and $s2$, respectively. Additionally, we analyse the correlation decay functions with inter-site distance between all pairs of stations.

Further, we apply the continuity ratio introduced by Wilks (1998) and used by other authors, for example, Breinl *et al.* (2013), to assess the spatial model performance. It expresses the ratio of precipitation mean at station $s1$ given zero precipitation at station $s2$, to the mean of precipitation at station $s1$ given nonzero precipitation at station $s2$:

$$\text{Continuity ratio} = \frac{E(P_{s1}|P_{s1} > 0 \cap P_{s2} = 0)}{E(P_{s1}|P_{s1} > 0 \cap P_{s2} > 0)} \quad (7)$$

The comparison of the continuity ratios for observed and synthetic series for each station is supposed to reveal the relationship between the mean precipitation at one station to the occurrence of precipitation at the other stations and is related to the model's ability to capture the spatial correlation. We contrast this statistic to the other metrics of the spatial model performance, which have so far not been applied all in one study.

The comparison of observed and simulated precipitation correlations at pairs of stations may be insufficient to reveal deficiencies in the areal rainfall simulation, that is, high matching of pair-wise correlations does not ensure the correct representation of correlation for 3 or more stations. Therefore, we explore the effect of the spatial correlation on the resulting areal precipitation fields. For this, we compute the mean and the 99.9th percentile of daily precipitation sums for each month at stations located within a fixed radius for every station. The exercise is repeated for radii of 100, 200 and 400 km. This analysis reveals a possible over/underestimation of areal precipitation as a function of distance.

Finally, we evaluate the at-site performance of the RWG with regards to the daily mean temperature and solar radiation by comparing the mean and 99.9th percentile for each month and the nine selected stations. Additionally, the comparison for all stations over the entire period is presented.

Additionally, in order to have a more transparent, consistent and comparable assessment of model performance,

we adopt the evaluation and performance framework (CASE) by Bennett *et al.* (2018). This framework is based on the categorization of the model performance into three categories: “good” (G), “fair” (F) and “poor” (P) based on each metric at each location. Good performance is attested if the observed metric is inside the 90% range of model output metric. The model performance is regarded fair if the observations are outside the 90% range, but within the 99.7% limits or absolute relative difference between the observation and simulated mean is 5% or less. Finally, poor performance is regarded otherwise. The evaluation at individual locations provides the percentage of stations with good, fair and poor performance (GFP) for each metric. According to Bennett *et al.* (2018), the overall performance can then be categorized into further 6 categories according to the GFP percentage as follows:

- “overall good” ($G > 50\%$)
- “overall fair” ($F > 50\%$)
- “overall poor” ($P > 50\%$)
- “overall fair-good” ($F + G > 50\%$)
- “overall good-fair” ($F + P > 50\%$)
- “overall variable” ($G + P > 50\%$)

5 | RESULTS AND DISCUSSION

We present the RWG results focusing on (a) at-site and (b) areal performance of precipitation generation.

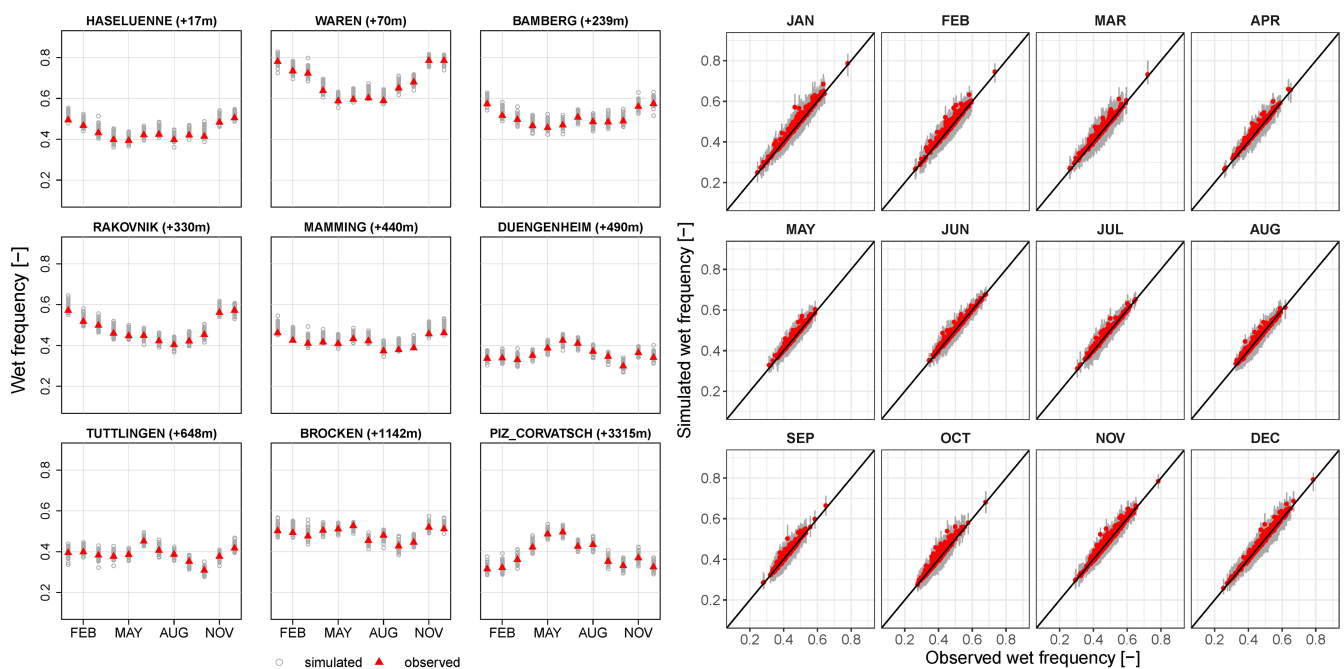


FIGURE 3 Comparison of wet frequency statistics for nine stations (left) and all stations (right) for the model version RWG1-extGP. Red dots (in the right plots) represent the median of the grey range corresponding to 100 model realizations [Colour figure can be viewed at wileyonlinelibrary.com]

Additionally, the results for at-site performance regarding mean daily temperature and solar radiation are shown. We evaluate the at-site performance at 9 selected station locations and compare the range of 100 model realization statistics to the observed values at 528 stations. In the following, we analyse the performance of the different model versions with respect to each metric. A summary of the RWG performance according to the CASE framework is provided in Table 1 for each model version.

All figures below show the results for the RWG1-extGP model version, whereas respective plots for the model version RWG0-mGGP are provided in the Supporting Information. Graphical results for the RWG0-extGP are very similar to the RWG0-mGP version in most cases and are not shown here. The performance is summarized in Table 1. We discuss improvements achieved with the introduced changes and potential causes of poor performance for some indicators.

TABLE 1 Summary of the model performance statistics for three RWG versions and categorization of the model performance according to Bennett *et al.* (2018). GFP [%] indicates the percentage of “good”, “fair”, and “poor” performance for all stations considering all 100 realizations

Metric	MODEL					
	RWG0-mGGP		RWG0-extGP		RWG1-extGP	
	(GFP) [%]	Overall	(GFP) [%]	Overall	(GFP) [%]	Overall
Wet frequency	(96,3,1)	Good	(95,4,1)	Good	(97,3,0)	Good
Transitional probability						
Wet-wet	(90,6,4)	Good	(98,1,1)	Good	(96,2,2)	Good
Dry-dry	(72,13,15)	Good	(78,16,5)	Good	(82,10,8)	Good
Monthly sum						
Mean	(100,0,0)	Good	(100,0,0)	Good	(100,0,0)	Good
99th percentile	(91,0,9)	Good	(90,0,10)	Good	(92,0,8)	Good
Daily intensity						
Mean	(100,0,0)	Good	(100,0,0)	Good	(100,0,0)	Good
99.9th percentile	(92,0,7)	Good	(87,6,13)	Good	(94,6,0)	Good
<i>n</i> -day maxima						
5-day sum	(74,0,26)	Good	(50,1,48)	Fair-good	(51,0,49)	Good
10-day sum	(79,0,21)	Good	(56,0,43)	Good	(60,0,40)	Good
Intersite-correlation						
Entire precipitation range	(53,1,47)	Good	(45,17,37)	Fair-good	(99,1,0)	Good
Precipitation above 80th percentile	(40,0,59)	Poor	(40,0,59)	Poor	(45,1,54)	Poor
Continuity ratio	(53,1,46)	Good	(52,0,48)	Good	(61,0,39)	Good
Areal precipitation						
Mean (for all radii)	(100,0,0)	Good	(100,0,0)	Good	(100,0,0)	Good
99.9th percentile						
<i>r</i> = 100 km	(86,0,14)	Good	(47,0,53)	Poor	(61,0,39)	Good
<i>r</i> = 200 km	(67,0,33)	Good	(28,0,72)	Poor	(37,0,63)	Poor
<i>r</i> = 400 km	(39,0,61)	Poor	(13,0,87)	Poor	(17,0,83)	Poor
Daily temperature						
Mean	(68,21,11)	Good	–	–	(100,0,0)	Good
99.9th percentile	(63,1,36)	Good	–	–	(87,0,13)	Good
Daily solar radiation						
Mean	(67,20,13)	Good	–	–	(98,2,0)	Good
99.9th percentile	(20,36,44)	Fair-poor	–	–	(30,36,34)	Fair-poor

5.1 | At-site RWG performance

The three RWG model versions capture the wet frequency statistics well for individual stations in the entire model domain since they report overall good performance with 95% or more stations showing good performance (Table 1, Figures 3 and Figure S1). Only a very slight overestimation of the observed wet frequencies particularly in winter months is detected, where the mean synthetic values are slightly above the 1:1 line in Figure 3. The analysis of nine selected stations does not point to a specific bias gradient with elevation or geographic location. The performance of all three versions is very similar with a minimal improvement exhibited by RWG1-extGP.

The RWG versions also reproduce the transition probabilities over the entire simulation domain for different months very well (Table 1, Figures 4 and Figure S2). At a few locations, for example, Haseluennen, Waren, Bamberg, Mamming, the models seem to slightly underestimate the dry-dry transition probabilities, that is, dry spells are not perfectly captured by the weather generator. Since the wet frequencies are simulated well (Figure 3), the results point to the more frequent intermittence of dry spells than in the observed data. Also Breinl *et al.* (2013) and Serinaldi and Kilsby (2014) detected an underestimation of the length of dry spells in their models. There is a very slight improvement of the model performance in the RWG1-extGP version

compared to the other two versions (Table 1) and it seems to result from both the marginal model and the improved estimation of the correlation matrices. In the context of flood risk modelling, more frequent intermittence of dry spells may affect antecedent soil moisture prior to some events. We believe, however, that this will not significantly change the magnitude of extreme floods and risk estimates.

The performance of the three RWG models on the monthly scale is overall good as indicated in Table 1, Figures 5 and Figure S3. The RWG0-mGGP model slightly overestimates the 99th percentile values. In the RWG1-extGP model, both mean and the 99th percentiles of the monthly sums are very well resembled in all months. To correctly reproduce the statistics at the monthly time scale is not really challenging and should be a minimum requirement for a weather generator.

Also, the daily mean and extreme (99.9th percentile) precipitation intensities are well captured by the RWG models (Table 1, Figures 6 and Figure S4). The observed daily extreme intensities at the nine selected stations lie within the range of 100 realizations. With regards to the entire model domain (Figures 6 and Figure S4, right), the observed extremes and median of the simulated range are located predominantly on the 1:1 line. Deviations are detected only at a few stations in some months, predominantly overestimation (note the log-scale).

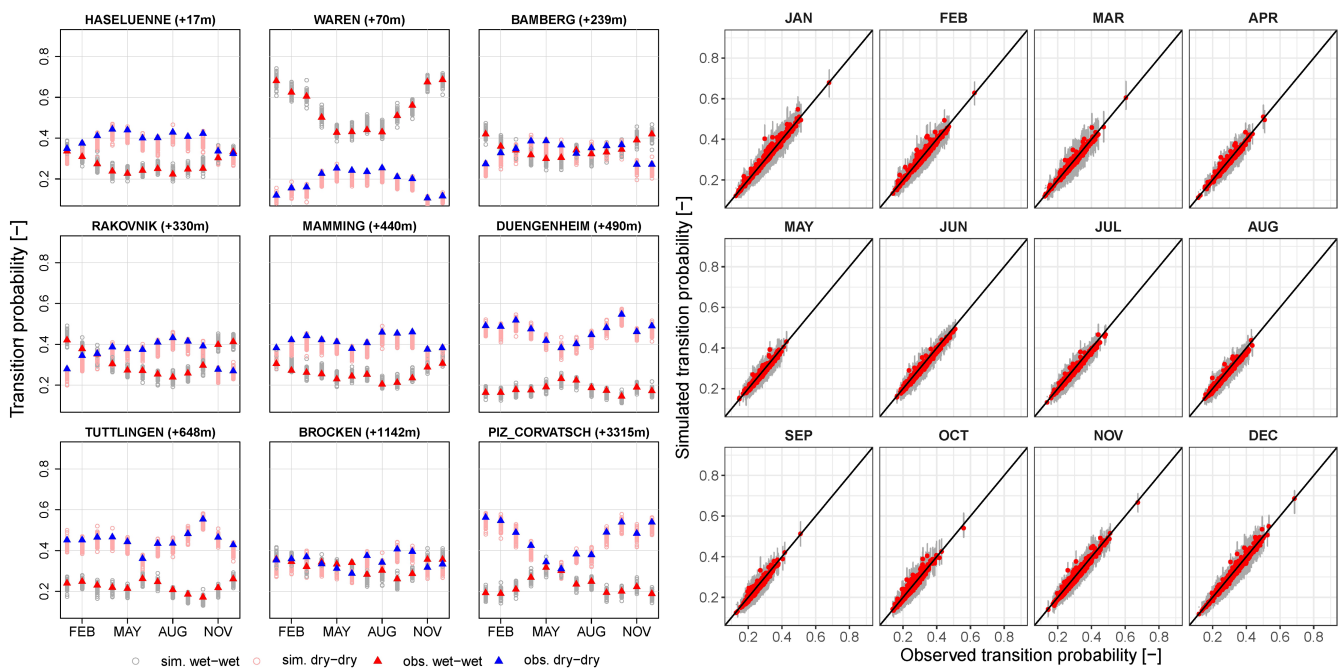


FIGURE 4 Comparison of simulated and observed transition probabilities (wet-wet and dry-dry) at 9 stations (left) and for all stations (wet-wet) (right) for the model version RWG1-extGP. Red dot represents the median of the grey range corresponding to 100 model realizations [Colour figure can be viewed at wileyonlinelibrary.com]

The results indicate an overall good fit of both marginal models to the at-site precipitation series. However, the extGP model delivers a poorer fit at a number of stations compared to mGGP (Table 1). Since marginal distributions

are not bounded, it can be expected that higher extremes are generated in the simulations compared to the limited observations, especially with increasing number of realizations. Hence, this is not a surprise that the grey bars in

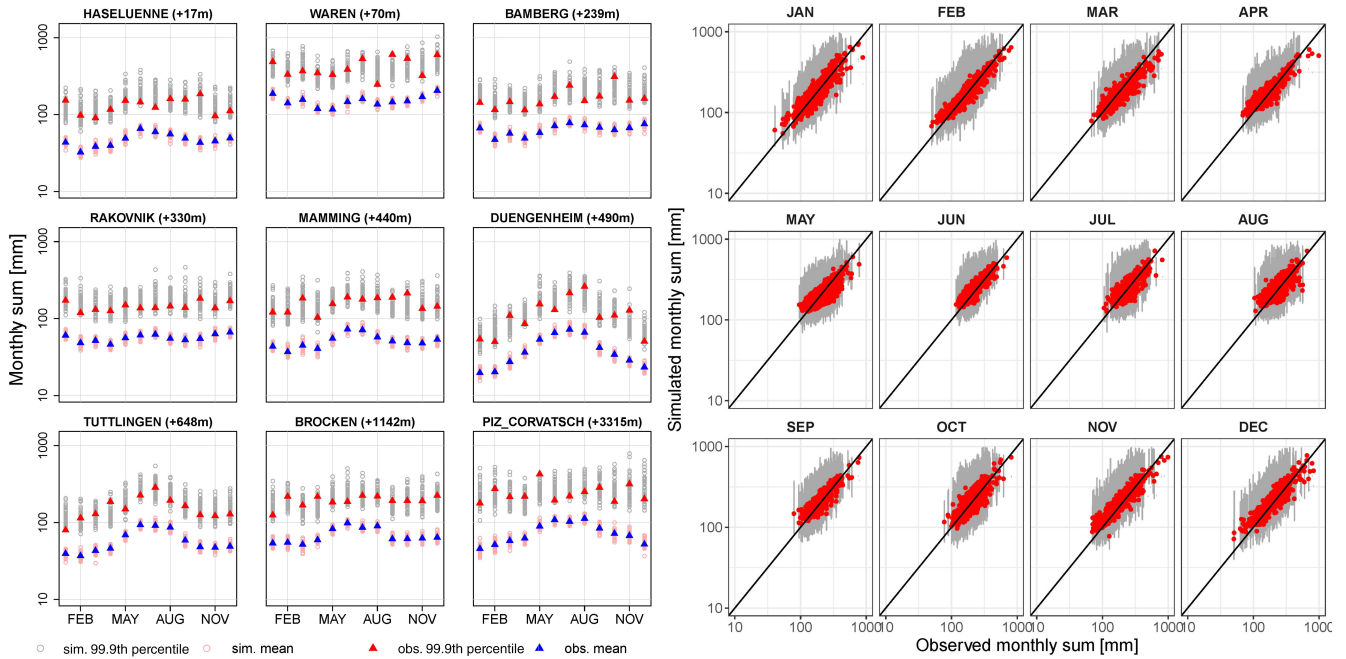


FIGURE 5 Comparison of monthly precipitation sums (mean and 99th percentile) at nine stations (left) and at all stations (99th percentile) (right) for the model version RWG1-extGP. Red dot represents the median of the grey range corresponding to 100 model realizations [Colour figure can be viewed at wileyonlinelibrary.com]

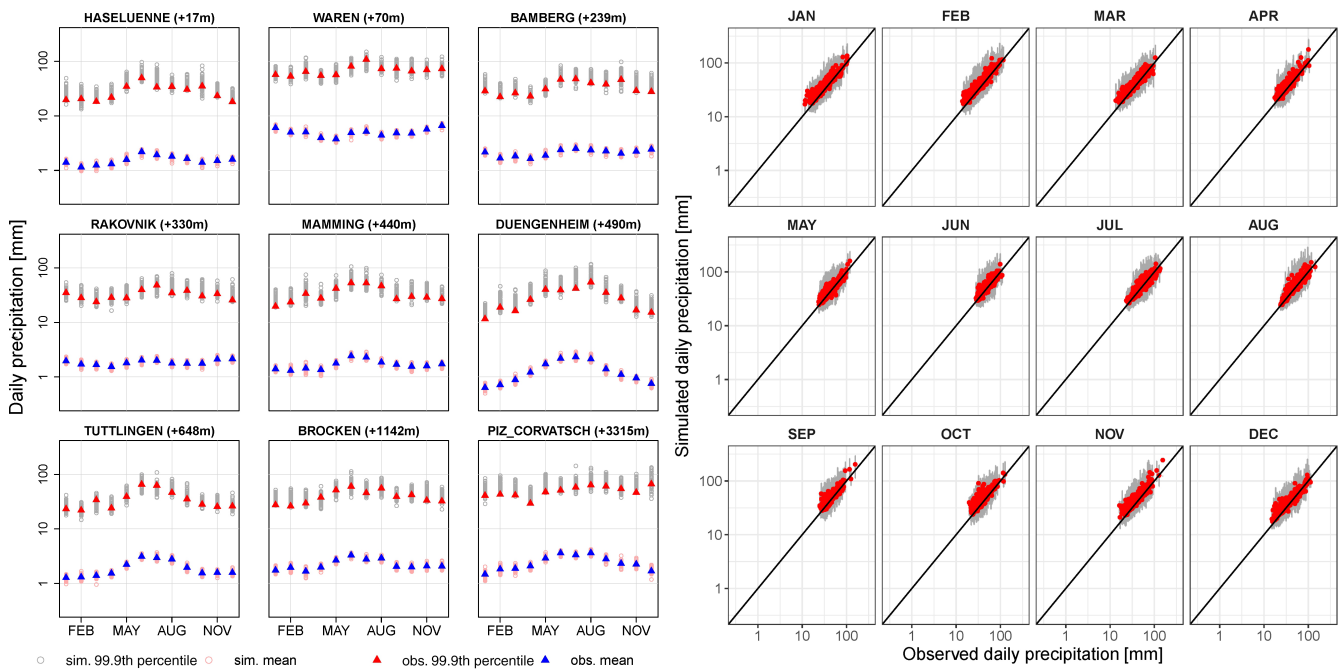


FIGURE 6 Comparison of daily precipitation intensities (mean and 99.9th percentile) for nine stations (left) and at all stations (99.9th percentile) (right) for the model version RWG1-extGP. Red dot represents the median of the grey range corresponding to 100 model realizations [Colour figure can be viewed at wileyonlinelibrary.com]

Figures 6 and Figure S4 (left) tend to overshoot the observed 99.9th percentile precipitation in most cases.

Figure 7 (Figure S5) and Figure 8 (Figure S6) show the ability of the RWG in simulating the multi-day

extreme precipitation sums for 5-day and 10-day accumulation periods, while Table 1 summarizes the overall performance. The observed frequency plots are enclosed by the simulated range of the RWG0-mGGP at all selected

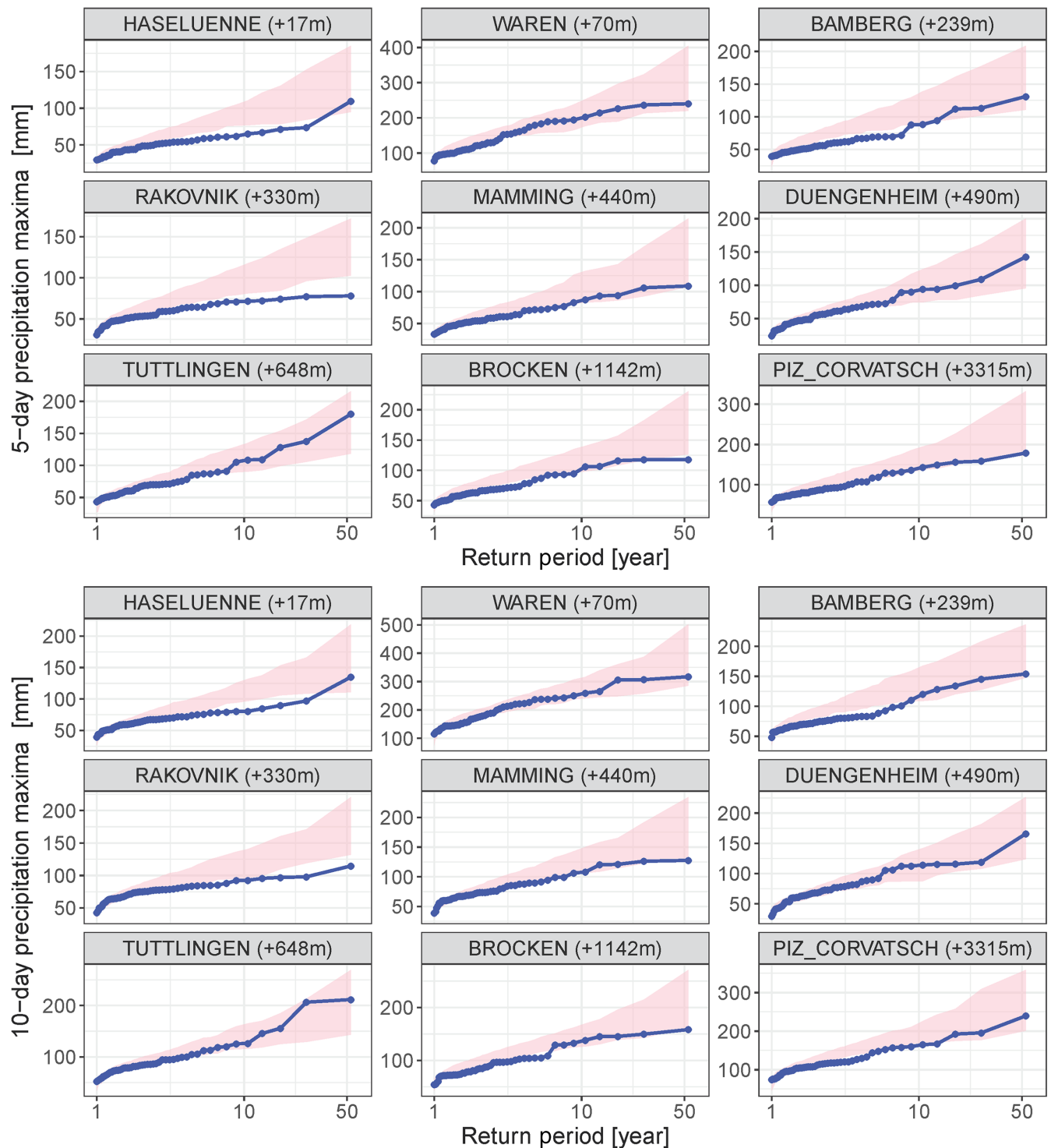


FIGURE 7 Frequency plots of observed (blue) and simulated multi-day extreme precipitation sums (red ranges) accumulated over 5-day (upper panel) and 10-day (lower panel) periods at nine selected stations. Results correspond to the RWG1-extGP model version. Note the log-scale of the x-axis. The Weibull plotting position is used to estimate the return periods [Colour figure can be viewed at wileyonlinelibrary.com]

stations but Rakovnik (Figure S5). The performance of the extGP-based models significantly deteriorates at several stations compared to RWG0-mGGP (Table 1). It results in a number of the observed frequency plots being off the simulated range (e.g., Brocken, Piz Corvatsch; Figure 7). At most of the selected stations, the multi-day extreme precipitation tends to be overestimated in all models. We also observe this tendency when comparing the extreme multi-day sums for all stations (Figures 8 and Figure S6). All RWG models are able to reproduce the 10-day extreme precipitation sums better compared to the 5-day period. Likely, the errors in precipitation modelling at single days average out over the longer period

Given a very similar performance of all versions in terms of daily intensity and monthly sums of precipitation, it may be surprising that RWG0-mGGP performs much better than extGP-based models. First, the fit of the at-site mGGP models is in 62% cases better than extGP according to the Akaike Information criterion (AIC; not shown). extGP seems to have a fatter upper tail compared to mGGP. This difference is not that noticeable for mean daily and extreme values (Table 1), but the error apparently accumulates for the multiday precipitation. To this end, extGP seems not to offer a better performance compared to the mGGP model and tends to overestimate the daily extremes and even more significantly the multi-day sums.

RWG1-extGP is slightly better than RWG0-extGP for both aggregation levels (5-day and 10-day) because of the impact of the dependence structure which will be discussed in Section 5.2.

The model performance with regards to the multi-day extreme precipitation sums can be influenced by both the at-site simulation performance (magnitude and auto-correlation) as well as by the model's ability to capture the

spatial correlation structure. A tendency to overestimate daily precipitation during wet spells at multiple sites might be hardly visible when comparing at-site percentiles of single days, but becomes more pronounced when the precipitation is accumulated over several days. The model performance in spatial terms is therefore analysed in the next section.

5.2 | RWG performance for areal precipitation

For derived flood frequency and risk analysis, both event precipitation intensity and total precipitation input over a large spatial domain are of interest. The literature review suggests that most state-of-the-art stochastic weather models are capable of simulating at-site precipitation characteristics down to the daily time scale, even for high quantiles. What remains challenging is the correct representation of the spatial correlation structure in the multi-site models and over large spatial domains (Serinaldi and Kilsby, 2014).

Although the density cloud of the pairwise correlation coefficients of the RWG1-extGP model in Figure 9 lies very close around the 1:1 line, indicating that the RWG1-extGP estimates the correlation structure of the full range of precipitation well in all months with only slight over-/underestimation. Figure 10 illustrates the observed and simulated correlations of precipitation for station pairs plotted against their distance. For this plot, the entire precipitation data are analysed. As expected, correlations decay from about 0.9 for nearby stations to below 0.2–0.4 for inter-site distances above ~500 km. The decay is much stronger for summer months compared to autumn and winter months. More localized summer precipitation events of convective origin result

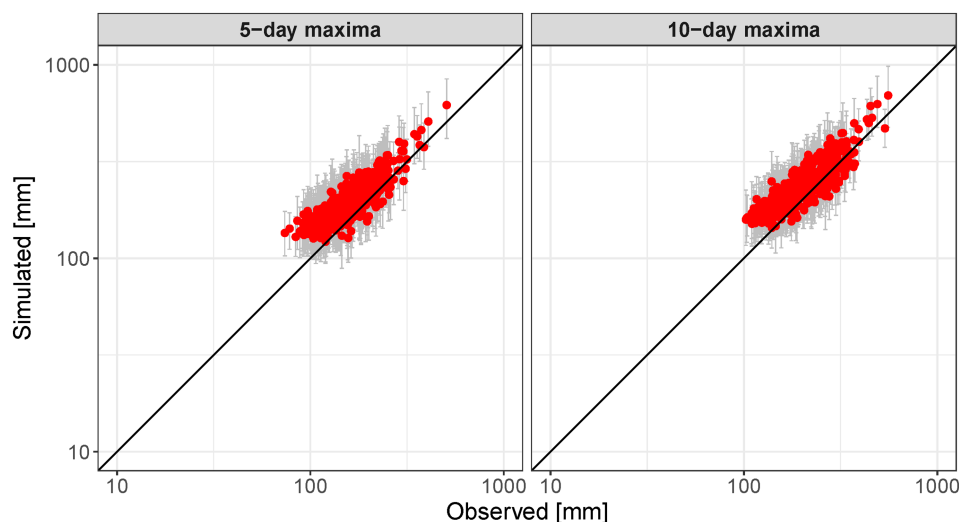


FIGURE 8 Observed vs. simulated multi-day extreme precipitation sums accumulated over 5-day (left panel) and 10-day (right panel) periods at all stations. Red dot represents the median of the grey range corresponding to 100 model realizations. Results correspond to the RWG1-extGP model version [Colour figure can be viewed at wileyonlinelibrary.com]

in shorter correlation distances (or lower correlations over larger distances) compared to the large-scale stratiform autumn/winter storms.

Table 1 and comparison between Figures 9 versus S7 and Figures 10 versus S8 reveal a significant improvement of the inter-site correlations for the entire data range when a new procedure for the estimation of the correlation matrix is implemented (RWG1-extGP). Whereas the use of the extGP marginal model slightly deteriorates model performance from overall good to overall fair-good, the RWG1-extGP performs good at 99% of the stations (Table 1).

The analysis of dependence in extreme precipitation above the 80th percentile threshold reveals a considerable overestimation of the correlation for all model versions (Table 1, Figures 11 and Figure S9). The overestimation is particularly striking from April till September, when the at-site precipitation is higher compared to the winter-half year, and is usually produced by localized convective storms. Figures 12 and Figure S10

indicate that extreme rainfalls are much more strongly correlated over large distances that would be expected from observations. The deviation is strongest for stations located between 200 and 600 km apart. The implementation of the new correlation matrix estimation procedure seems to have very limited impact on the correlation of high quantiles. The percentage of GFP narrowly shifts from (40,0,59) for the RWG0-models to (45,1,54) for RWG1-extGP resulting in the overall poor model performance (Table 1).

This result implies that extreme precipitation is more likely to occur in the simulated data at multiple locations than would be expected from observations. This leads to an overall higher areal rainfall volume and may result in flood discharge overestimation for large catchments. This overestimation of correlation between station pairs may explain the slight overestimation of the multi-day extreme precipitation sums. If precipitation were overestimated during wet spells because of more probable

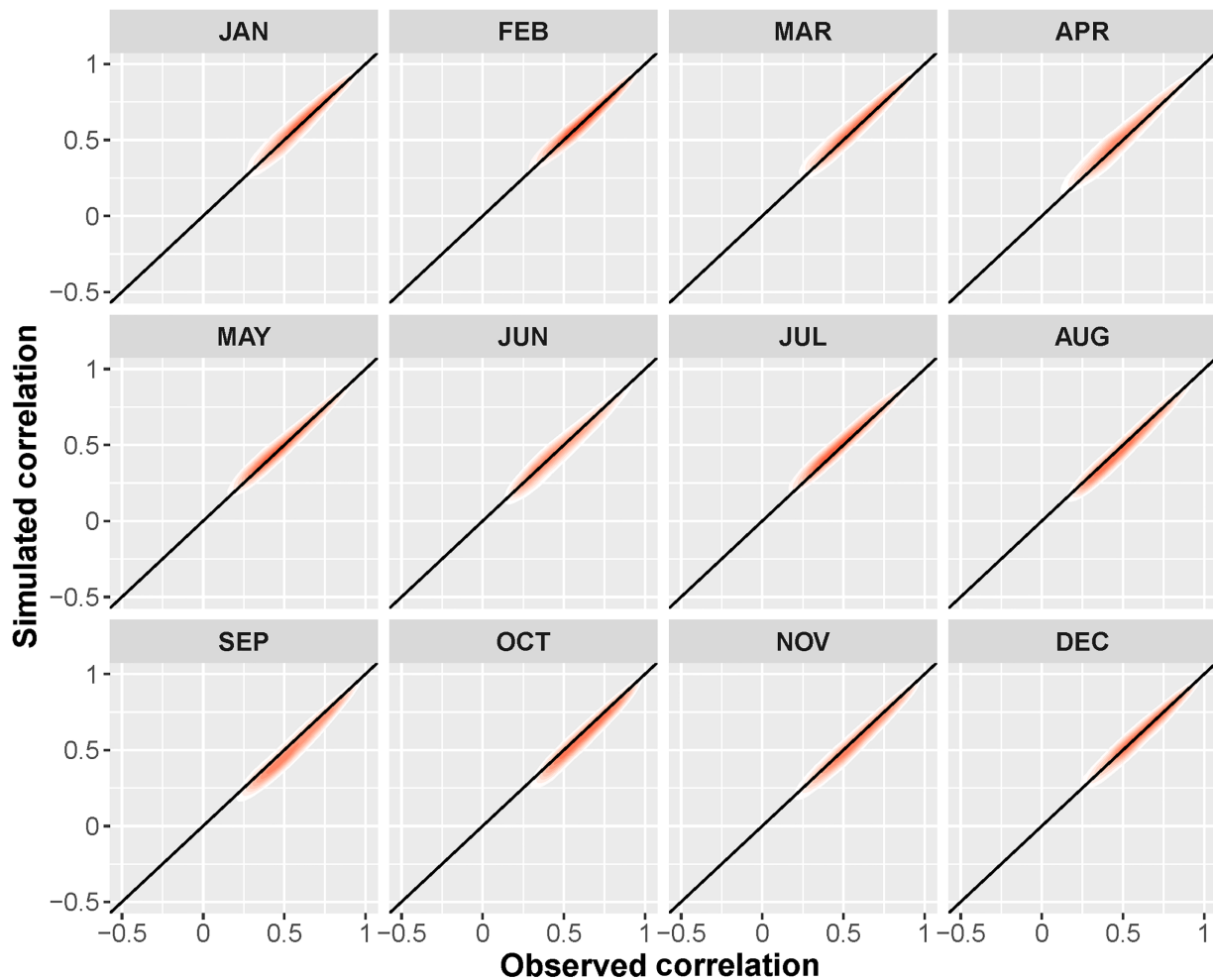


FIGURE 9 Observed and simulated correlation coefficients for all precipitation data in different months for the RWG1-extGP model. Increasing density of the points is indicated by the shaded colour from white to red [Colour figure can be viewed at wileyonlinelibrary.com]

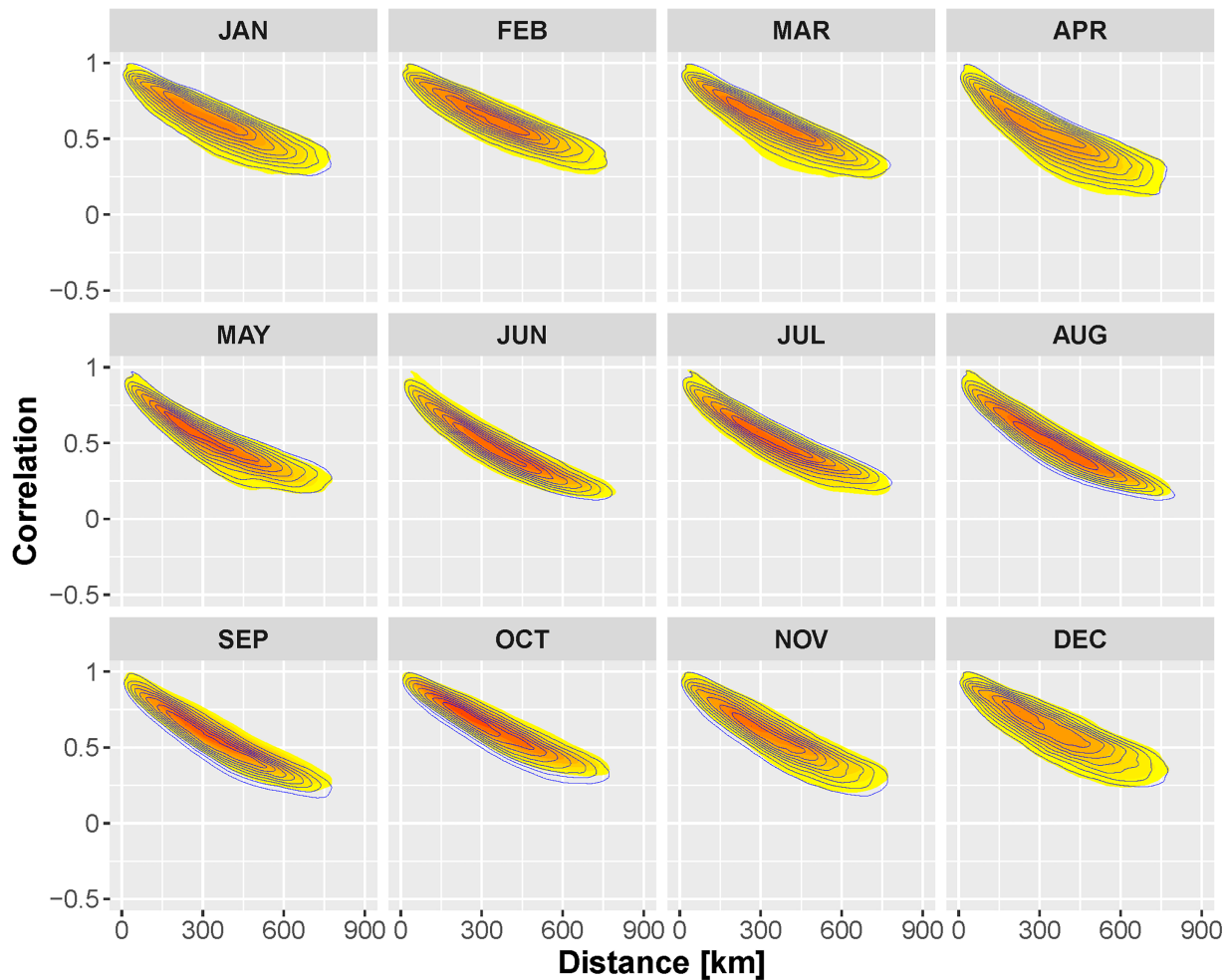


FIGURE 10 Correlation vs. distance between station pairs for observed and simulated precipitation with the RWG1-extGP model version. Increasing density of points for observed series is indicated in shaded colours from yellow to red. The density of points for the simulated series is indicated by the contour lines [Colour figure can be viewed at wileyonlinelibrary.com]

occurrence of extremes at multiple sites, single-site multi-day sums would also be greater than expected.

Also, the comparison of continuity ratios for observed and simulated precipitation reveals the overestimation of the spatial correlation in the summer half-year in all models (Table 1, Figures 13 and Figure S11). The observed continuity ratio is higher than that of the simulated series, that is, the nominator in Equation (11) is larger, which could mean that simulated events are more localized in space. However, interpretation of the continuity ratio is not straightforward. Contrary to the correlation plots in Figures 9–12 (Figures S7–S10), it conditions the mean values on the occurrence of rainfall at other stations and does not reveal the threshold behaviour that is immediately visible in Figures 11–12 (Figures S9–S10). Although continuity ratio was applied in some studies (Wilks, 1998; Breinl *et al.*, 2013), we would rather recommend the use of monthly correlation plots and correlation decay functions with inter-site distance.

Table 1 shows an excellent performance of the three RWG model versions in reproducing the mean areal precipitation at three selected radii. However, the extreme (99.9th percentile) areal precipitation is clearly overestimated as demonstrated in Figures 14, Figure S12 and Table 1. The tendency to overestimation already visible for the radius of 100 km, although the simulated envelope mostly encloses observed values. With increasing radius, the deviation between observed and simulated extreme areal precipitation increases, particularly in the summer months. Because both mean and extreme daily precipitation intensity at individual station are well captured, the marked overestimation of inter-site correlation of extreme daily precipitation discussed above is apparently the main cause of the overestimation of the extreme areal precipitation.

Comparing the performance for three RWG model versions, we see a combined impact of the selected marginal distribution and the dependence structure on the

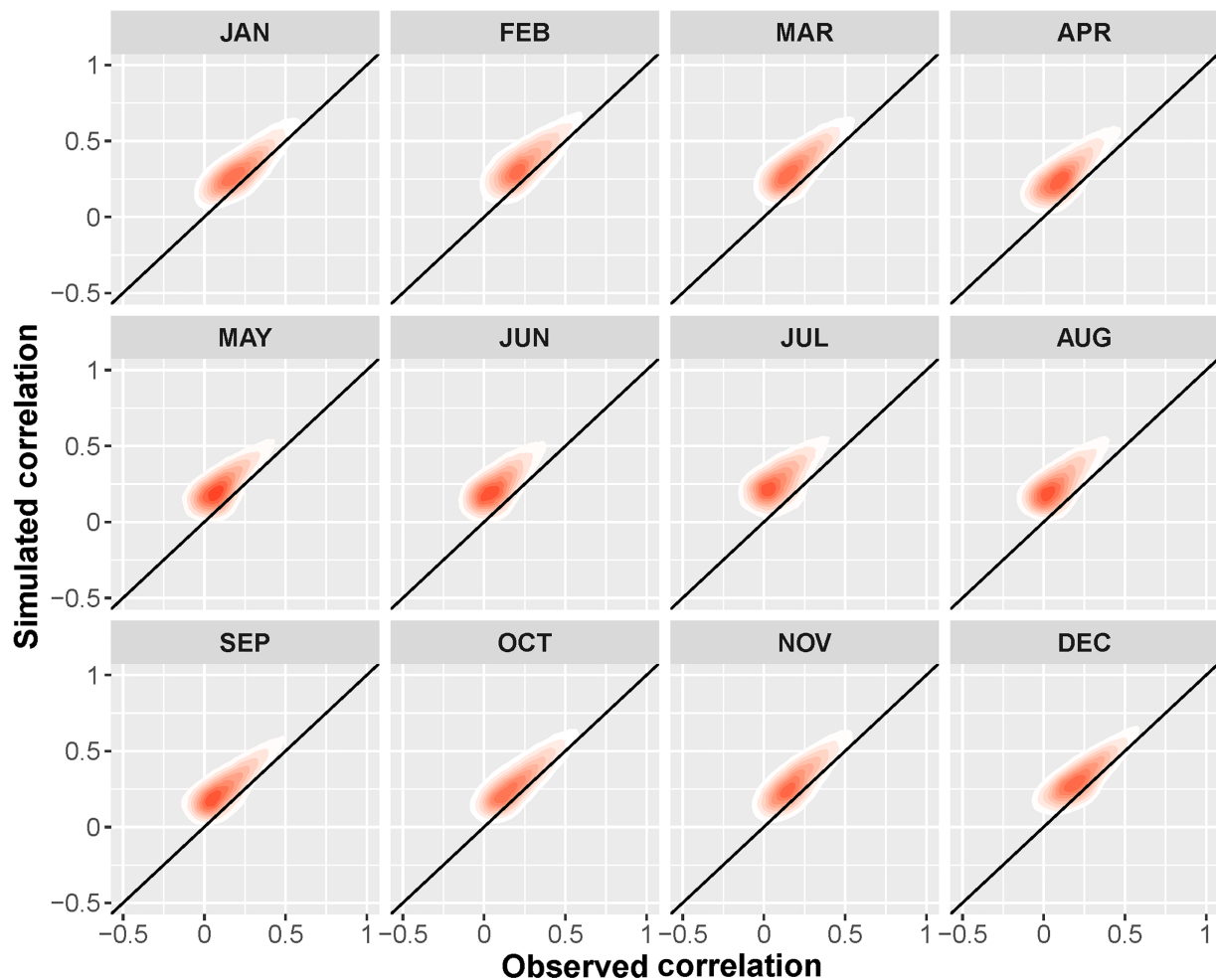


FIGURE 11 Observed and simulated correlation coefficients for precipitation data above the 80th percentile threshold in different months for the RWG1-extGP model. Increasing density of the points is indicated by the shaded colour from white to red [Colour figure can be viewed at wileyonlinelibrary.com]

extreme areal precipitation. RWG0-extGP is worse than RWG0-mGGP although they share the same dependence structure. This may indicate the impact of the overestimation of the upper quantiles of the at-site daily precipitation since extGP may have a fatter tail compared to mGGP. RWG1-extGP performs better than RWG0-extGP because of the improvement in the correlation structure (Table 1, Figures 9-12 and Figures S7-S10).

Analysis of the spatial correlation structure and areal precipitation reveals some difficulties of the RWG in capturing summer precipitation events. Overestimation of the areal rainfall increases with increasing spatial scale. Simple correlation plots for the entire precipitation series (Figures 9-10 and Figures S7-S8) do not reveal the problem. Neither is the continuity ratio instrumental in identifying the reason for the poor model performance. Hence, we recommend the use of threshold correlation plots in combination with the analysis of areal precipitation sums.

Baxevani and Lennartsson (2015) also face the problem of overestimation of correlations in their latent Gaussian field model, but rather in winter months. Serinaldi and Kilsby (2014) indicate a considerable overestimation of daily areal precipitation in their model in all months. As suggested by Serinaldi and Kilsby (2014), we assume the covariance function derived from the bulk of the precipitation data to be responsible for the overestimation of areal extreme rainfall. Also, Breinl *et al.* (2020) indicate a faster decrease of the areal reduction factor (ratio of catchment to point rainfall) with catchment area for convective precipitation compared to stratiform precipitation. This is related to the typical footprints of both precipitation types and hence to inter-site correlation. The bulk of the precipitation data, which might be dominated by large-scale stratiform events with stronger inter-site correlations dominate the parameterization. Our analysis reveals that correlation distances for extreme rainfall are shorter,

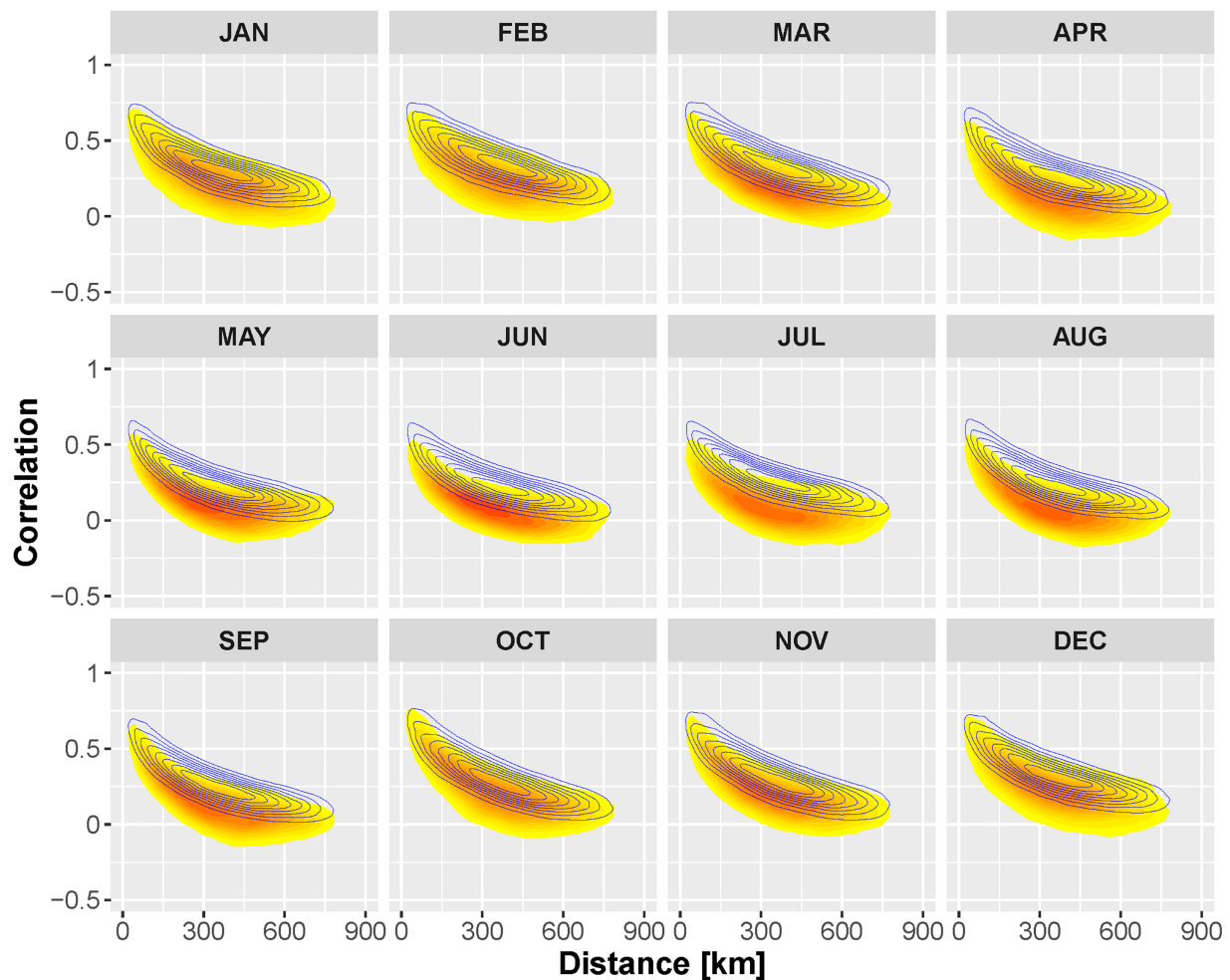


FIGURE 12 Correlation versus distance between station pairs for observed and simulated precipitation above the 80th percentile threshold with the RWG1-extGP model version. Increasing density of points for observed series is indicated in shaded colours from yellow to red. The density of points for the simulated series is indicated by the contour lines [Colour figure can be viewed at wileyonlinelibrary.com]

particularly in summer months, which is consistent with physical reasoning. In summer, convective rainfall events are more frequent and can be expected to have shorter correlation distances. Müller and Haberlandt (2015) also detected lower correlations and much stronger variability of correlations with distance for daily precipitation above the 4 mm threshold in a catchment in northern Germany.

A potential avenue to improve the simulated correlation of extreme precipitation could be the use of threshold approach to separately model the bulk and extreme precipitation as suggested by Müller and Haberlandt (2015). We tested this approach adopting the 80th percentile threshold. This results in a better performance with respect to the correlation of extreme precipitation, but at the cost of a strong performance deterioration for the bulk precipitation (not shown). Another approach we explored is to change the

current meta-Gaussian based dependence structure to a so-called meta-skew-normal (Azzalini and Dalla Valle, 1996) because the latter is more flexible in its dependence strength at different quantiles. Testing this approach revealed a similar trade-off between the performance of RWG with regards to the correlation of the entire precipitation range compared to the extreme precipitation as in the above-mentioned threshold approach. Although the performance deterioration was not that strong (not shown), We think more effort is required to test different approaches and find optimal way to estimate the spatial correlation structure and this should be focus of future research. For instance, Evin *et al.* (2018) apply a copula-based approach following Bárdossy and Pegram (2009) to simulate tail dependence of precipitation at multiple sites. This approach seems, however, to notably under estimate spatial correlations (Evin *et al.*, 2018).

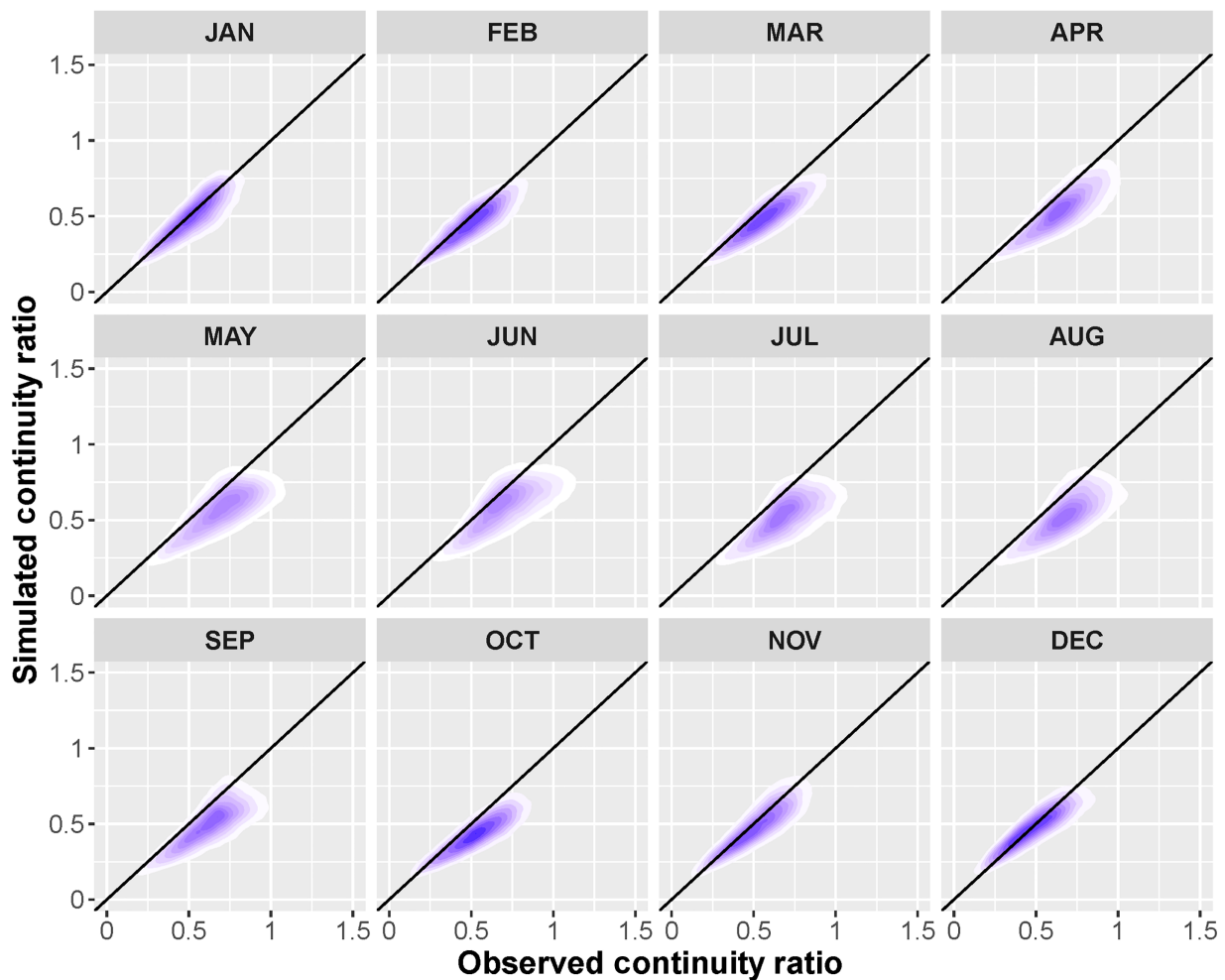


FIGURE 13 Monthly observed and simulated spatial continuity ratio for the RWG1-extGP model version. Increasing density of points is indicated by shaded colours from white to blue [Colour figure can be viewed at wileyonlinelibrary.com]

5.3 | RWG performance with regards to mean temperature and solar radiation

Two climate variables conditioned on the wet-dry state, mean temperature and solar radiation are evaluated in a similar way as precipitation. Figures 15 and Figure S13 demonstrate the ability of two model versions in reproducing the mean daily temperature statistics. The introduced power transformation in the RWG1 model improved the performance with regards to daily mean temperature significantly resulting in good performance at all stations. The 99.9th percentile values of mean daily temperatures are more difficult to simulate particularly in the winter months, when we detect a positive bias of 1–3° in the RWG0-mGGP. In RWG1-extGP, the performance improves significantly with the percentage of GFP shifting from (63,1,36) to (87,0,13). The overestimation is, however, rather small and we consider it to be not decisive for the purpose of flood risk assessment, although it might have some control on snowmelt events during the winter season.

Finally, we evaluate the RWG performance in reproducing the observed values of solar radiation (Figures 16 and Figure S14). The mean daily values are generally well captured by the RWG0-mGGP model version with GFP percentage values of (67,20,13; Table 1). The power transformation substantially improved the overall performance with regards to the daily mean values in the RWG1-extGP model with GFP of (98,2,0). However, the 99.9th percentiles remain strongly overestimated particularly in the summer months at all stations (Figure 16). The RWG1-extGP model performs slightly better than RWG0-mGGP, but the overall performance of both versions with respect to extreme values remains “fair-poor”. The dispersion of points along the 1:1 line for all stations is high from May to September (Figures 16 and Figure S14). It seems the power transformation offers some relief, but does not completely solve the problem. It seems that normal distribution is not the best choice also for the power-transformed radiation data. Further investigations are needed to better characterize

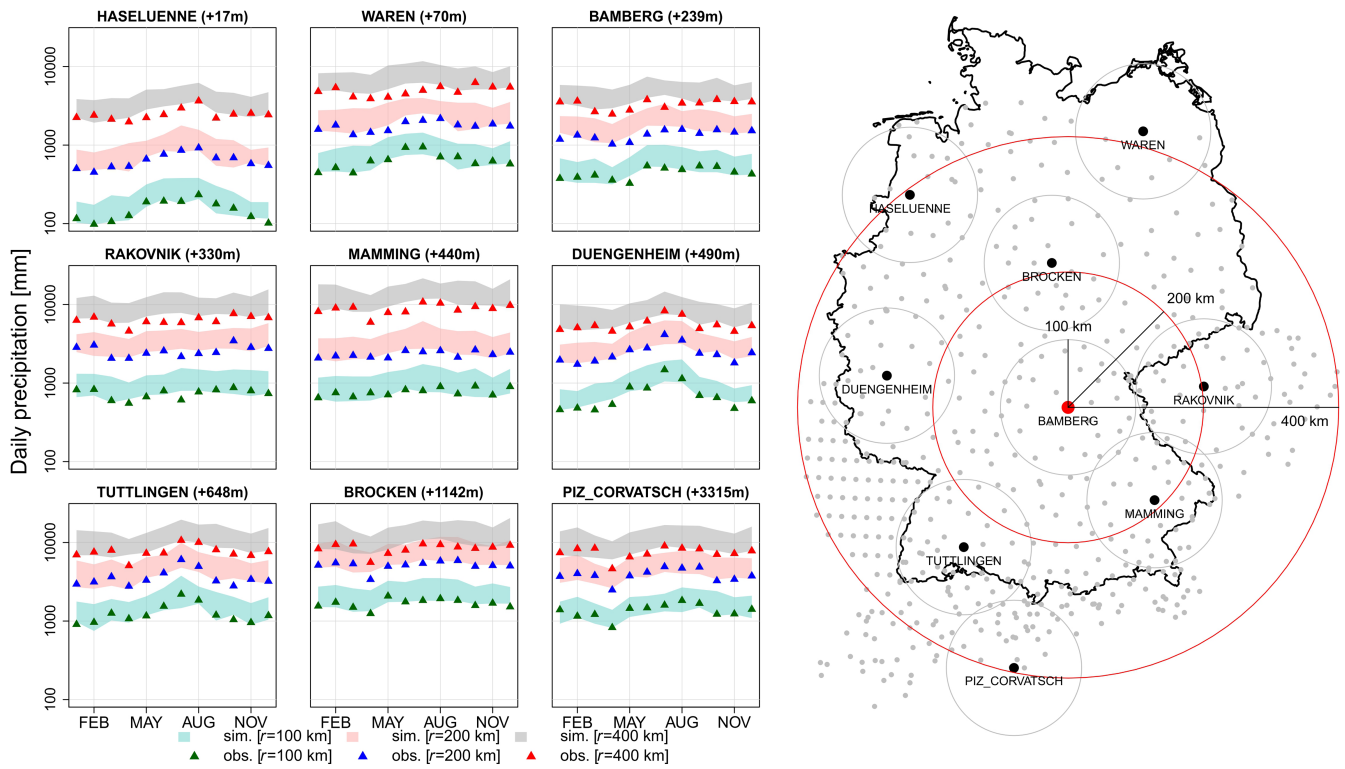


FIGURE 14 99th percentile of daily observed and simulated precipitation accumulated over all stations within various radii from nine selected stations. Shaded ranges represent the range of 100 model realizations with the RWG1-extGP model version. Note the log-scale of the y-axis [Colour figure can be viewed at wileyonlinelibrary.com]

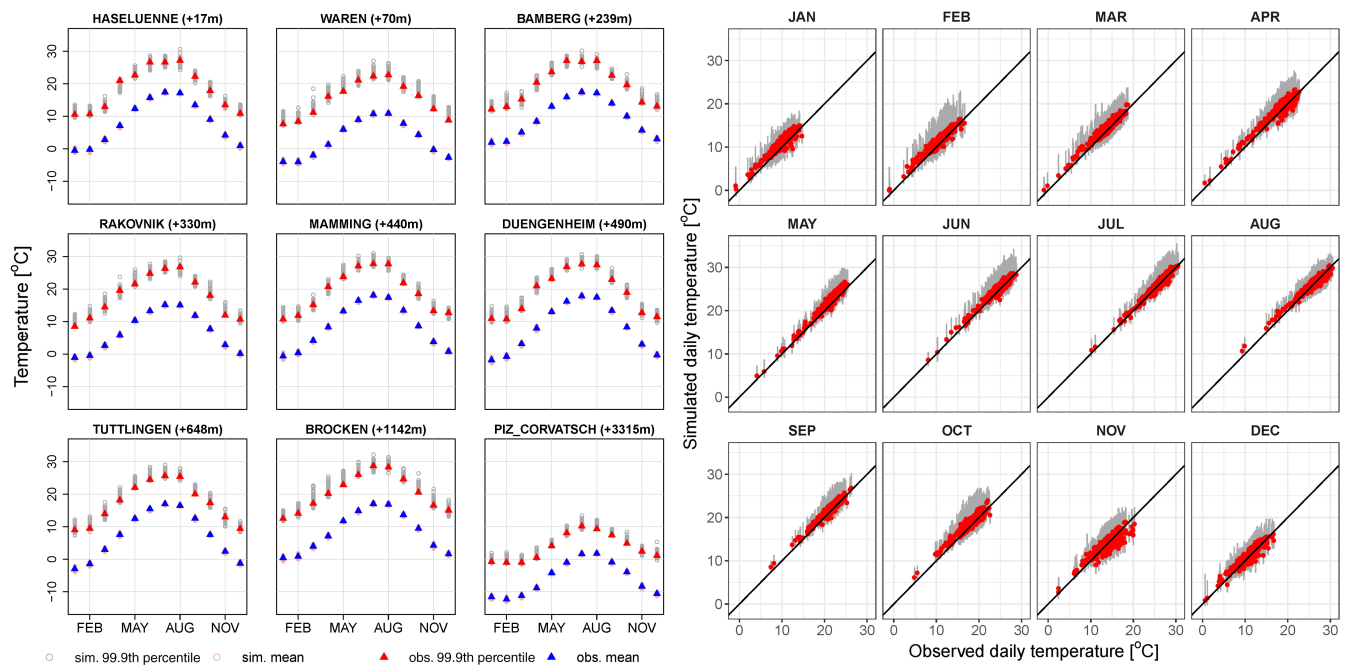


FIGURE 15 Comparison of observed and simulated mean daily temperature (mean and 99.9th percentile) at nine stations (left) and for all stations (99.9th percentile). Red dot represents the median of the grey range generated with the RWG1-extGP model version [Colour figure can be viewed at wileyonlinelibrary.com]

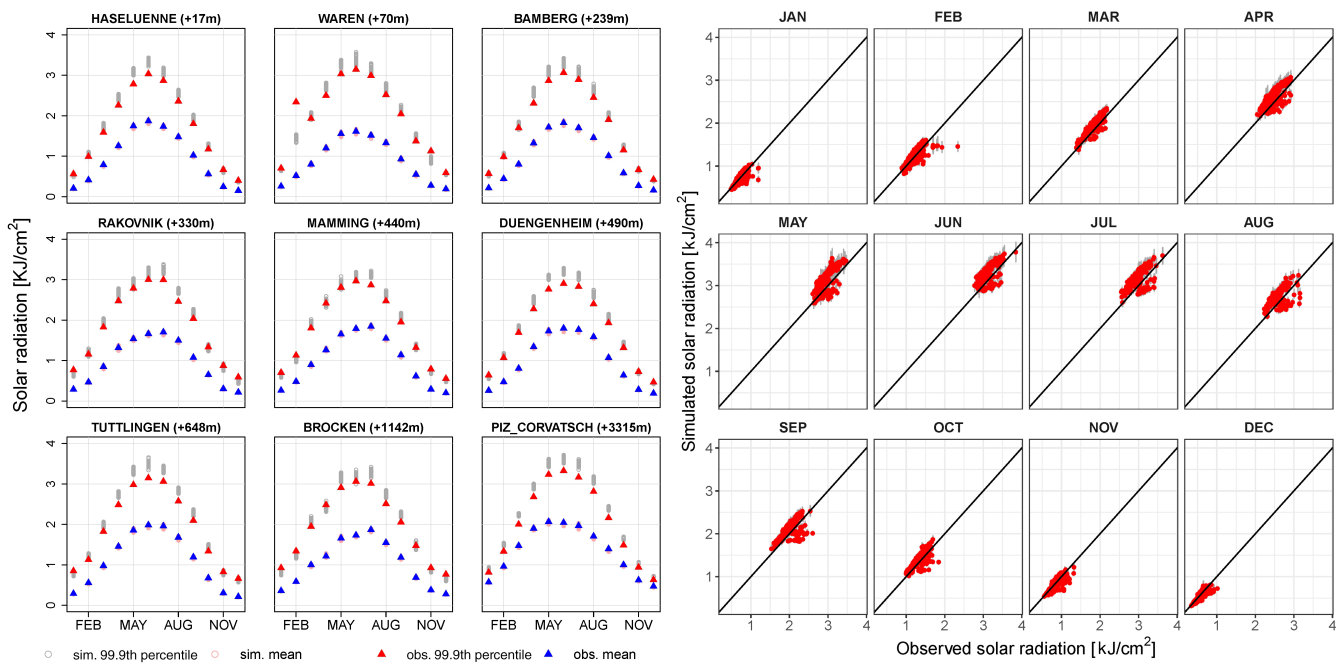


FIGURE 16 Comparison of observed and simulated mean daily solar radiation (mean and 99.9th percentile) at nine stations (left) and for all stations (99.9th percentile). Red dot represents the median of the grey range generated with the RWG1-extGP model version [Colour figure can be viewed at wileyonlinelibrary.com]

the distribution of solar radiation and should include other types of transformations and marginal distribution families to capture extreme values. Overall, we believe this dispersion to be not decisive for shaping flood peaks and will likely not significantly influence regional flood risk estimates.

6 | CONCLUSIONS

This article presents a comprehensive evaluation of a multivariate auto-regressive weather generator setup at regional scale of about 580,000 km² covering major river basins in Central Europe. The regional setup of the weather generator (RWG) is calibrated and evaluated at 528 climate stations for the purpose of derived flood frequency and risk analysis. We validated the performance of the three RWG model versions in simulating precipitation occurrence and dry/wet transition probabilities, mean daily and extreme (99.9th percentile) precipitation, multi-day precipitation sums, spatial correlation structure and areal precipitation. Finally, the performance with regard to daily mean temperature and daily mean solar radiation is analysed. These variables are conditioned on the wet/dry state simulated by the weather generator. The original model version RWG0-mGGP uses a six-parameter mixed Gamma-Generalized Pareto (mGGP) as a marginal distribution for precipitation, normal

distribution for mean daily temperature and normal distribution in combination with root square transformation for solar radiation. With the second model version RWG0-extGP, we explore the effect of using a parsimonious 3-parameter extended Generalized Pareto (extGP) distribution with other components kept the same. Finally, in the third version RWG1-extGP, a new procedure to estimate and correct the spatial correlation matrices is introduced along with the power transformation applied to all non-precipitation variables in combination with normal distribution. We evaluate and categorize the model performance using the CASE framework by Bennett *et al.* (2018).

All RWG model versions are overall good in representing the observed wet day frequencies. The wet-wet transition probabilities are very well reproduced by the RWG, whereas dry-dry transition probabilities are somewhat underestimated. This suggests that the models tend to simulate more frequent intermittence of dry spells compared to observations. Monthly mean and extreme precipitation are well captured by all model versions. This also applies to daily mean precipitation, whereas the daily extremes seem to be somewhat overestimated. The RWG tends to generate higher at-site extreme precipitation, particularly in summer months. All versions perform very similarly for the above-mentioned statistics except for the extreme daily precipitation. Here, the extGP marginal distribution performs

slightly worse at a number of stations as it tends to generate heavier tails than mGGP. Hence, the performance for 5-day and 10-day accumulated precipitation strongly deteriorates for the extGP-based versions as the error at the individual days is likely to accumulate. The RWG0-mGGP captures the pair-wise correlations and the correlation decay with distance between stations for the entire precipitation range with good and poor performance at 53% and 47% of the stations, respectively. However, the use of a new procedure to estimate spatial correlations and correct the correlation matrix to positive definite boosts the performance to good fit at all of the stations. Zooming into the correlation performance for extreme precipitation above the 80th percentile threshold revealed considerable overestimation of the pair-wise correlations and decay function for all versions. The overestimation is particularly pronounced in the summer half-year. The new procedure for correlation estimation does not noticeably affect this problem. The overestimation of spatial correlation leads to the simulation of more severe flood events in spatial terms compared to what has been observed. The use of the power transformation for the non-precipitation variables improves the model performance considerably. However, RWG1-extGP tends to slightly overestimate extreme solar radiation particularly in summer. We consider this bias to be negligible in the context of applications of the weather generator for derived flood frequency and flood risk analysis.

Overall, we conclude that all versions of the weather generator are very skilful in capturing precipitation intermittence and weather extremes at individual locations. The mixed Gamma-Generalized Pareto model effectively captures both the bulk and extremes of at-site precipitation. As also found by Evin *et al.* (2018), the use of extended Generalized Pareto distribution does not result in improved model performance, even some worsening is observed. As a limitation, we acknowledge that our models are not cross-validated given a very high computational burden for this large-scale application. The new procedure for estimation of spatial correlation and correction of the correlation matrix as well as power transformation for non-precipitation variables significantly improve model performance. The current representation of the spatial precipitation correlation structure without differentiating between extreme and average precipitation should, however, be improved in the future. Despite the above-mentioned limitations, the RWG is able to computationally handle a large dataset of a several hundred stations and can be used for large-scale derived flood frequency and trans-basin flood risk assessment considering the above-mentioned limitations.

ACKNOWLEDGEMENT

This research has been funded by the Federal Ministry of Education and Research of Germany in the framework of the project FLOOD (project number 01LP1903E) as a part of the ClimXtreme Research Network on Climate Change and Extreme Events within framework programme Research for Sustainable Development (FONA3). Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the research group FOR 2416 “Space-Time Dynamics of Extreme Floods (SPATE)” (project number 278017089) is gratefully acknowledged. We thank Dr. Francesco Serinaldi, Dr. Korbinian Breinl and an anonymous reviewer for constructive comments that helped to improve the manuscript.

Open Access funding enabled and organized by ProjektDEAL.

ORCID

Viet Dung Nguyen  <https://orcid.org/0000-0002-2649-2520>

Bruno Merz  <https://orcid.org/0000-0002-5992-1440>

Uwe Haberlandt  <https://orcid.org/0000-0002-3650-4249>

Sergiy Vorogushyn  <https://orcid.org/0000-0003-4639-7982>

REFERENCES

- Azzalini, A. and Dalla Valle, A. (1996) The multivariate skew-normal distribution. *Biometrika*, 83, 715–726.
- Bárdossy, A. and Plate, E. (1992) Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research*, 28, 1247–1259. <https://doi.org/10.1029/91WR02589>.
- Bárdossy, A. and Pegram, G.G.S. (2009) Copula based multisite model for daily precipitation simulation. *Hydrology and Earth System Sciences*, 13, 2299–2314.
- Baxevani, A. and Lennartsson, J. (2015) A spatiotemporal precipitation generator based on a censored latent Gaussian field. *Water Resources Research*, 51, 4338–4358. <https://doi.org/10.1002/2014WR016455>.
- Beersma, J.J. and Buishand, T.A. (2003) Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. *Climate Research*, 125, 121–133.
- Bennett, B., Thyer, M., Leonard, M., Lambert, M. and Bates, B. (2018) A comprehensive and systematic evaluation framework for a parsimonious daily rainfall field model. *Journal of Hydrology*, 556, 1123–1138.
- Blazkova, S. and Beven, K.J. (1997) Flood frequency prediction for data limited catchments in The Czech Republic using a stochastic rainfall model and TOPMODEL. *Journal of Hydrology*, 195, 256–278.
- Blazkova, S. and Beven, K.J. (2004) Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in The Czech Republic. *Journal of Hydrology*, 292, 153–172.
- Breinl, K., Turkington, T. and Stowasser, M. (2013) Stochastic generation of multi-site daily precipitation for applications in risk

- management. *Journal of Hydrology*, 498, 23–35. <https://doi.org/10.1016/j.jhydrol.2013.06.015>.
- Breini, K., Müller-Thomy, H. and Blöschl, G. (2020) Space–time characteristics of areal reduction factors and rainfall processes. *Journal of Hydrometeorology*, 21(4), 671–689. <https://doi.org/10.1175/JHM-D-19-0228.1>.
- Caraway, N.M., McCreight, J.L. and Rajagopalan, B. (2014) Multi-site stochastic weather generation using cluster analysis and k-nearest neighbor time series resampling. *Journal of Hydrology*, 508, 197–213. <https://doi.org/10.1016/j.jhydrol.2013.10.054>.
- Cameron, D.S., Beven, K.J., Tawn, J., Blazkova, S. and Naden, P. (1999) Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty). *Journal of Hydrology*, 219, 169–187. [https://doi.org/10.1016/S0022-1694\(99\)00057-8](https://doi.org/10.1016/S0022-1694(99)00057-8).
- Chardon, J., Hingray, B. and Favre, A.-C. (2018) An adaptive two-stage analog/regression model for probabilistic prediction of small-scale precipitation in France. *Hydrology and Earth System Sciences*, 22, 265–286.
- Cowpertwait, P.S. (2006) A spatiotemporal point process model of rainfall for the Thames catchment, UK. *Journal of Hydrology*, 330(3–4), 1779–1794.
- de Moel, H., Jongman, B., Kreibich, H., Merz, B., Penning-Rowsell, E. and Ward, P.J. (2015) Flood risk assessments at different spatial scales. *Mitigation and Adaptation Strategies for Global Change*, 20(6), 865–890. <https://doi.org/10.1007/s11027-015-9654-z>.
- Duan, Q., Sorooshian, S. and Gupta, V. (1992) Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, 28, 1015–1031.
- Eagleson, P.S. (1972) Dynamics of flood frequency. *Water Resources Research*, 8, 878–898.
- Evin, G., Favre, A.-C. and Hingray, B. (2018) Stochastic generation of multi-site daily precipitation focusing on extreme events. *Hydrology and Earth System Sciences*, 22, 655–672. <https://doi.org/10.5194/hess-22-655-2018>.
- Falter, D., Schröter, K., Dung, N.V., Vorogushyn, S., Kreibich, H., Hundecha, Y., Apel, H. and Merz, B. (2015) Spatially coherent flood risk assessment based on long-term continuous simulation with a coupled model chain. *Journal of Hydrology*, 524, 182–193. <https://doi.org/10.1016/j.jhydrol.2015.02.021>.
- Falter, D., Dung, N.V., Vorogushyn, S., Schröter, K., Hundecha, Y., Kreibich, H., Apel, H., Theisselmann, F. and Merz, B. (2016) Continuous, large-scale simulation model for flood risk assessments: proof-of-concept. *Journal of Flood Risk and Management*, 9, 3–21.
- Fowler, H.J., Kilsby, C.G. and O'Connell, P.E. (2000) A stochastic rainfall model for the assessment of regional water resource systems under changed climatic conditions. *Hydrology and Earth System Sciences*, 4(2), 263–282.
- Fowler, H.J., Kilsby, C.G., O'Connell, P.E. and Burton, A. (2005) A weather-type conditioned multi-site stochastic rainfall model for the generation of scenarios of climatic variability and change. *Journal of Hydrology*, 308, 50–66.
- Frigessi, A., Haug, O. and Rue, H. (2002) A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5, 219–235. <https://doi.org/10.1023/A:1024072610684>.
- Furrer, E.M. and Katz, R.W. (2008) Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, 44(12), W12439. <https://doi.org/10.1029/2008WR007316>.
- Grimaldi, S., Petroselli, A. and Serinaldi, F. (2012a) Design hydrograph estimation in small and ungauged watersheds: continuous simulation method versus event-based approach. *Hydrological Processes*, 26, 3124–3134. <https://doi.org/10.1002/hyp.8384>.
- Grimaldi, S., Petroselli, A. and Serinaldi, F. (2012b) A continuous simulation model for design-hydrograph estimation in small and ungauged watersheds. *Hydrological Sciences Journal*, 57, 1035–1051. <https://doi.org/10.1080/02626667.2012.702214>.
- Haberlandt, U., Hundecha, Y., Pahlow, M. and Schumann, A. (2011) Rainfall generators for application in flood studies. In: Schumann, A.H. (Ed.) *Flood Risk Assessment and Management*. Springer: Dordrecht, pp. 117–147 DOI: 10.301007/978-90-481-9917_47.
- Haberlandt, U. and Radtke, I. (2014) Hydrological model calibration for derived flood frequency analysis using stochastic rainfall and probability distributions of peak flows. *Hydrological Earth System Science*, 18, 353–365. <https://doi.org/10.5194/hess-18-353-2014>.
- Haylock, M. R., Hofstra, N., Klein Tank, A. M. G, Klok, E. J., Jones, P. D. and New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research*, 113, D20119. <https://doi.org/10.1029/2008JD010201>.
- Higham, N. J. (2002): Computing the nearest correlation matrix—a problem from finance, *IMA Journal of Numerical Analysis*, 22(3), 329–343, doi:<https://doi.org/10.1093/imanum/22.3.329>, 2002.
- Hundecha, Y., Pahlow, M. and Schumann, A. (2009) Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes. *Water Resources Research*, 45(12), W12412. <https://doi.org/10.1029/2008WR007453>.
- Hundecha, Y. and Merz, B. (2012) Exploring the relationship between changes in climate and floods using a model-based analysis. *Water Resources Research*, 48(4), W04512. <https://doi.org/10.1029/2011WR010527>.
- Hutson, A.D. (2002) A semi-parametric quantile function estimator for use in bootstrap estimation procedures. *Statistics and Computing*, 12, 331–338. <https://doi.org/10.1023/A:1020783911574>.
- Keller, D.E., Fischer, A.M., Frei, C., Liniger, M.A., Appenzeller, C. and Knutti, R. (2015) Implementation and validation of a Wilks-type multi-site daily precipitation generator over a typical alpine river catchment. *Hydrology and Earth System Sciences*, 19, 2163–2177.
- Kleiber, W., Katz, R.W. and Rajagopalan, B. (2012) Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resources Research*, 48, W01523. <https://doi.org/10.1029/2011WR011105>.
- Li, C., Singh, V.P. and Mishra, A.K. (2012) Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water Resources Research*, 48. <https://doi.org/10.1029/2011WR011446>.
- Metin, A.D., Dung, N.V., Schröter, K., Guse, B., Apel, H., Kreibich, H., Vorogushyn, S. and Merz, B. (2018) How do changes along the risk chain affect flood risk? *Natural Hazards and Earth System Sciences*, 18, 3089–3108. <https://doi.org/10.5194/nhess-18-3089-2018>.

- Müller, H. and Haberlandt, U. (2015) Temporal rainfall disaggregation with a Cascade model: from single-station disaggregation to spatial rainfall. *Journal of Hydrologic Engineering*, 20, 4015026. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001195](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001195).
- Naveau, P., Huser, R., Ribereau, P. and Hannart, A. (2016) Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4), 2753–2769.
- Österle, H., Gerstengarbe, F. and Werner, P. (2006a) *Ein neuer meteorologischer Datensatz für Deutschland, 1951–2003*. Potsdam: Potsdam Institute for Climate Impact Research.
- Österle, H., Werner P.C., Gerstengarbe F.W. (2006b): Qualitätsprüfung, Ergänzung und Homogenisierung der täglichen Datenreihen in Deutschland, 1951–2003: Ein neuer Datensatz, *Deutsche Klimatagung. Klimatrends: Vergangenheit und Zukunft*. München (Munich).
- Österle, H., Gerstengarbe, F.W. and Werner, P.C. (2016) *Die Elbe im Globalen Wandel, chap. 2.2 Ein Meteorologischer Datensatz für Deutschland, 1951–2003*. Stuttgart, Germany: Schweizerbart Science Publishers, pp. 81–84.
- Papalexiou, S.M. and Koutsoyiannis, D. (2013) Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, 49(1), 187–201. <https://doi.org/10.1029/2012WR012557>.
- Papalexiou, S.M., Koutsoyiannis, D. and Makropoulos, C. (2013) How extreme is extreme? An assessment of daily rainfall distribution tails. *Hydrology and Earth System Sciences*, 17(2), 851–862.
- Rajagopalan, B. and Lall, U. (1999) A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, 35(19), 3089–3101.
- Rajah, K., O’Leary, T., Turner, A., Petrakis, G., Leonard, M. and Westra, S. (2014) Changes to the temporal distribution of daily precipitation. *Geophysical Research Letters*, 41(24), 8887–8894. <https://doi.org/10.1002/2014GL062156>.
- Rasmussen, P.F. (2013) Multisite precipitation generation using a latent autoregressive model. *Water Resources Research*, 49(4), 1845–1857.
- Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A. and Gratzki, A. (2013) A central European precipitation climatology – part I: generation and validation of a high-resolution gridded daily data set (HYRAS). *Meteorologische Zeitschrift*, 22(3), 235–256. <https://doi.org/10.1127/0941-2948/2013/0436>.
- Raynaud, D., Hingray, B., Evin, G., Favre, A.-C. and Chardon, J. (2020) Assessment of meteorological extremes using a synoptic weather generator and a downscaling model based on analogues. *Hydrology and Earth System Sciences*, 24, 4339–4352.
- Richardson, C.W. (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17, 182–190. <https://doi.org/10.1029/WR017i001p00182>.
- Serinaldi, F. and Kilsby, C.G. (2014) Simulating daily rainfall fields over large areas for collective risk estimation. *Journal of Hydrology*, 512, 285–302. <https://doi.org/10.1016/j.jhydrol.2014.02.043>.
- Sharif, M. and Burn, D.H. (2007) Improved k-nearest neighbor weather generating model. *Journal of Hydrologic Engineering*, 12(1), 42–51. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2007\)12:1\(42\)](https://doi.org/10.1061/(ASCE)1084-0699(2007)12:1(42)).
- Sparks, N., Hardwick, R., Schmid, M. and Toumi, R. (2018) IMAGE: a multivariate multi-site stochastic weather generator for European weather and climate. *Stochastic Environmental Research and Risk Assessment*, 32, 771–784. <https://doi.org/10.1111/jfr3.12105>.
- Vorogushyn, S., Bates, P.D., de Bruijn, K., Castellarin, A., Kreibich, H., Priest, S., Schröter, K., Bagli, S., Blöschl, G., Domeneghetti, A., Gouldby, B., Klijn, F., Lammersen, R., Neal, J.C., Ridder, N., Terink, W., Viavattene, C., Viglione, A., Zanardo, S. and Merz, B. (2018) Evolutionary leap in large-scale flood risk assessment needed. *WREs Water*, 5, e1266. <https://doi.org/10.1002/wat2.1266>.
- Vrac, M. and Naveau, P. (2007) Stochastic downscaling of precipitation: from dry events to heavy rainfalls. *Water Resources Research*, 43, W07402. <https://doi.org/10.1029/2006WR005308>.
- Ward, P.J., Jongman, B., Salamon, P., Simpson, A., Bates, P., De Groot, T., Muis, S., Coughlan de Perez, E., Rudari, R., Trigg, M.A. and Winsemius, H.C. (2015) Usefulness and limitations of global flood risk models. *Nature Climate Change*, 5, 712–715. <https://doi.org/10.1038/nclimate2742>.
- Wilks, D.S. (1998) Multi-site generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210, 178–191.
- Wilks, D.S. (1999) Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain. *Agricultural and Forest Meteorology*, 96, 85–101. [https://doi.org/10.1016/S0168-1923\(99\)00037-4](https://doi.org/10.1016/S0168-1923(99)00037-4).
- Winter, B., Schneeberger, K., Nguyen, D., Huttenlau, M., Achleitner, S., Stötter, J., Merz, B. and Vorogushyn, S. (2019) A continuous modelling approach for design flood estimation on sub-daily time scale. *Hydrological Sciences Journal*, 64(5), 539–554.
- Zorita, E. and von Storch, H. (1999) The analog method as a simple statistical downscaling technique: comparison with more complicated models. *Journal of Climate*, 12, 2474–2488.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Nguyen VD, Merz B, Hundecha Y, Haberlandt U, Vorogushyn S. Comprehensive evaluation of an improved large-scale multi-site weather generator for Germany. *Int J Climatol*. 2021;41:4933–4956. <https://doi.org/10.1002/joc.7107>