Stochastic Model Output Statistics for Bias Correcting and Downscaling Precipitation Including Extremes

GERALDINE WONG AND DOUGLAS MARAUN

GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

MATHIEU VRAC

Laboratoire des Sciences du Climat et de l'Environnement, CEA Saclay, Gif-sur-Yvette, France

MARTIN WIDMANN AND JONATHAN M. EDEN

School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, United Kingdom

THOMAS KENT

School of Mathematics, University of Leeds, Leeds, United Kingdom

(Manuscript received 4 October 2013, in final form 14 April 2014)

ABSTRACT

Precipitation is highly variable in space and time; hence, rain gauge time series generally exhibit additional random small-scale variability compared to area averages. Therefore, differences between daily precipitation statistics simulated by climate models and gauge observations are generally not only caused by model biases, but also by the corresponding scale gap. Classical bias correction methods, in general, cannot bridge this gap; they do not account for small-scale random variability and may produce artifacts. Here, stochastic model output statistics is proposed as a bias correction framework to explicitly account for random small-scale variability. Daily precipitation simulated by a regional climate model (RCM) is employed to predict the probability distribution of local precipitation. The pairwise correspondence between predictor and predictand required for calibration is ensured by driving the RCM with perfect boundary conditions. Wet day probabilities are described by a logistic regression, and precipitation intensities are described by a mixture model consisting of a gamma distribution for moderate precipitation and a generalized Pareto distribution for extremes. The dependence of the model parameters on simulated precipitation is modeled by a vector generalized linear model. The proposed model effectively corrects systematic biases and correctly represents local-scale random variability for most gauges. Additionally, a simplified model is considered that disregards the separate tail model. This computationally efficient model proves to be a feasible alternative for precipitation up to moderately extreme intensities. The approach sets a new framework for bias correction that combines the advantages of weather generators and RCMs.

1. Introduction

Precipitation is the main source of freshwater strongly affecting river runoff, groundwater recharge, and the water level of lakes and reservoirs. As such, it is an indispensable resource for ecosystems, agriculture, and almost all human activities (Bates et al. 2008). Extreme precipitation is a major hazard; according to the Munich Re Group, global overall losses due to hydrological events in 2011 amounted to some $64 billion (https://www.munichre.com/touch/naturalhazards/en/homepage/default.aspx). Anthropogenic climate change is expected to considerably influence the hydrological cycle, leading to shifts in global precipitation patterns and an increasing magnitude of extreme precipitation over many regions (Meehl et al. 2007; Seneviratne et al. 2012).

To quantify the often localized impacts of changing precipitation, numerical models, such as hydrological models, are often employed (Xu 1999). These models

Corresponding author address: Douglas Maraun, GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany.
E-mail: dmaraun@geomar.de

DOI: 10.1175/JCLI-D-13-00604.1

require realistic high-resolution scenarios of precipitation change as input, which often includes a realistic representation of intensities ranging from dry days to extreme events; temporal variability from daily to decadal variability; and spatial variability (Maraun et al. 2010b). Yet our knowledge of future climate change mainly stems from global coupled atmosphere–ocean general circulation models (AOGCMs). These models have a rather coarse resolution of typically more than 100 km, and the actual scale on which they provide reasonable skill is, in general, even larger (Grotch and MacCracken 1991). Therefore, AOGCMs do not directly provide information on regional scales and do not correctly represent localized extreme precipitation.

To bridge the scale gap between global climate simulations and the required local-scale input, different downscaling approaches have been proposed. Dynamical downscaling nests a high-resolution regional climate model (RCM) over a limited area into the coarse global model (Rummukainen 2010). Perfect prog (PP) statistical downscaling infers a statistical link (usually as some form of regression model) between large-scale and local-scale weather observations and transfers this relationship to a global climate simulation for future simulations (Maraun et al. 2010b). Model output statistics (MOS) approaches, originally developed to correct systematic biases in weather forecasts (Glahn and Lowry 1972), statistically ''correct'' climate model biases (Maraun et al. 2010b). For a chosen climate model, MOS infers a correction function between a simulated and the corresponding observed variable in the present-day climate and applies this correction function to a future simulation with the same model. Weather generators (WGs) are statistical models that explicitly model the temporal structure (and often intervariable relationships) on short time scales up to several days (Maraun et al. 2010b). To represent longer-term variability and the climate change signal, WGs can be employed in a PP setting; that is, their parameters can be conditioned on the large-scale circulation (Wilby and Wigley 2000; Vrac and Naveau 2007). All statistical approaches implicitly assume that the inferred statistical relationships are valid under climate change.

Here, we propose a new stochastic framework for MOS and implement a specific statistical model. The framework combines the advantages of MOS and PP weather generators. The key idea is to simultaneously correct for systematic biases and stochastically downscale to station scale.

Over the last years, large projects such as the Prediction of Regional Scenarios and Uncertainties for Defining European Climate Change Risks and Effects (PRUDENCE; Christensen and Christensen 2007) and the Ensemble-Based Predictions of Climate Changes and Their Impacts (ENSEMBLES; van der Linden and Mitchell 2009) have stimulated the use of RCMs. A key advantage of RCMs compared to purely statistical downscaling approaches is that they explicitly resolve mesoscale atmospheric processes and, therefore by construction, provide spatially coherent and—to a certain degree—physically consistent output. However, compared to observed climate, RCM simulations are, in general, considerably biased (Christensen et al. 2008) and therefore often cannot directly be used as input for impact models. MOS approaches are therefore used to postprocess model output. Most currently used MOS approaches are calibrated in a distribution-wise setting; that is, only long-term observational distributions (climatologies) are compared with long-term simulated distributions to derive the correction function (Maraun et al. 2010b).

Many impact models, such as hydrological and agricultural yield models, are calibrated against point data. RCMs, however, simulate gridbox values that represent area averages. In particular, precipitation is highly variable in space and time; gridbox averages, in general, cannot explain all subgrid variability, such as localized high intensity events or even wet day probabilities. This representativeness problem between grid and point scale cannot be overcome by traditional bias correction methods, because they are deterministic (i.e., they only correct systematic biases but do not add random small-scale variability). Any deterministic MOS approach aiming to correct the simulated variability, such as variance correction or quantile mapping (Piani et al. 2010), if used for downscaling below gridbox scale may consequently produce wrong variability and trends (Maraun 2013).

For PP downscaling, von Storch (1999) suggested randomization to add the necessary subgrid small-scale variability to a downscaled time series. This approach, however, requires a separation of the local-scale variability into variance explained by the predictor and random noise. Such a separation in turn requires a regression model (i.e., ultimately pairwise correspondence between the predictor and predictand). For RCMs with boundary conditions from a free-running AOGCM, this precondition is not fulfilled. As a simple way to utilize RCM output but produce local random variability, change factor–adjusted WGs have been proposed (e.g., Kilsby et al. 2007). Here, the WG parameters, calibrated to observed present-day conditions, are adjusted by the (long-term average) climate change signal of the corresponding parameters derived from an RCM simulation. Such WGs, however, are not downscaling in a strict sense, as the day-to-day weather

sequence is not conditioned on the RCM but only on the future RCM climatology. Therefore, although this approach can provide potentially useful information, these WGs are not consistent with the RCM-simulated weather and do not produce any random variability beyond a few days; in particular, they produce no interannual or decadal variability.

A solution to ensure pairwise correspondence between observed and simulated time series is to use perfect boundary conditions (i.e., reanalysis data) for the RCM and additionally nudge the RCM to the reanalysis fields at large scales (von Storch et al. 2000). In such a setting, the simulated large-scale circulation is in strong agreement with the observed weather, and thus the local simulated and observed day-to-day weather sequences are correlated. Therefore, one is able to formulate regression models between simulated predictors and observed predictands (e.g., Widmann et al. 2003; Themeßl et al. 2011). A pairwise MOS approach provides the basis for keeping much of the explained variability from the RCM, in particular interannual and decadal variability, but also correcting systematic biases and adding the required small-scale variability.

Here, we propose a pairwise stochastic MOS approach for correcting and downscaling climate model output. Conceptually similar approaches have recently been developed in weather forecasting to use MOS for predicting continuous probability distributions (e.g., Gneiting et al. 2005; Friederichs 2010; Berrocal et al. 2010; Thorarinsdottir and Johnson 2012). In addition to correcting systematic biases, this approach also downscales to local scales and represents the full local-scale intensity distribution ranging from dry days to extreme events. To model wet day probabilities, a logistic regression is used. Precipitation intensities are described by a mixture model (Frigessi et al. 2002; Vrac and Naveau 2007); that is, moderate precipitation is modeled by a gamma distribution and the extreme tail is modeled by a generalized Pareto (GP) distribution. The dependence of the model parameters on the simulated RCM precipitation is modeled by a vector generalized linear model (Yee and Wild 1996; Yee and Stephenson 2007; Maraun et al. 2010a, 2011). In its current version, our model does not explicitly account for spatial dependencies (beyond the dependency imprinted by the RCM); that is, it is a single site model. By drawing random numbers from the predictand time-varying distribution, our model can be used as a precipitation generator. It therefore sets the stage for a new framework: MOS weather generators that are consistent with the large-scale circulation simulated by the chosen numerical model (be it RCM or GCM).

Section 2 gives a general overview of the approach, and the data used in this study are described in section 3. In section 4, the statistical model used for downscaling precipitation occurrence and precipitation intensities is described, along with the model selection procedure. Finally, the goodness of fit and performance of our stochastic MOS approach are evaluated in section 5, and an example application is shown.

2. General approach

Gridbox values represent area averages and do not provide information about local subgrid variability. Part of the subgrid variability is systematic: for example, because of elevation or rain shadow effects (i.e., in statistical terms, predicted or explained by gridbox values). But in particular for precipitation, a considerable fraction of subgrid variability is, in general, random (in the sense that it cannot be predicted by gridbox values). For instance, the occurrence of convective precipitation might be well predictable by the gridbox value, but not the exact position, let alone the exact amount at a particular location. This issue is one aspect of the representativeness problem between gridbox and point values (Zwiers et al. 2013).

Current variance-correcting MOS approaches, such as quantile mapping, cannot overcome the representativeness problem for two reasons: First, they are deterministic and do not add unexplained random variability. Second, they are calibrated in a distribution-wise setting (i.e., calibrated on long-term distributions) and therefore cannot even separate the local variability into a systematically varying fraction and an unexplained small-scale variability that has to be modeled as random noise. Instead, deterministic variance corrections merely inflate the systematic variability to match the total local variability, while also inflating other systematic features, such as long-term trends. Furthermore, if applied to more than one gauge within a grid box, the predicted subgrid spatial structure is completely deterministic. As a consequence, the spatial extent of dry areas and extreme events might be heavily overestimated as well (Maraun 2013).

The aim of this study, then, is to separate the explained variance from the total local-scale variance and explicitly model the unexplained small-scale variance. One way to achieve this is by building a regression model. A prerequisite for any regression model is the pairwise correspondence between predictors and predictands (i.e., in our case, temporal correspondence between simulated and observed precipitation at the daily scale). To distinguish such regression-based MOS approaches from simple distribution-wise methods, we term these approaches "pairwise MOS."

To ensure the required temporal correspondence between simulated and observed individual events, we use an RCM driven with perfect boundary conditions from reanalysis data and additionally spectrally nudged within the domain to the large-scale reanalysis fields (von Storch et al. 2000). An initial setup without spectral nudging resulted in too much freedom of the RCM to develop its own small-scale variability, such that the resulting model showed predictive skill only across specific regions. Driving the RCM with boundary conditions from reanalysis data allows one to correct only for RCM biases. If transferred to AOGCM-driven RCM simulations (e.g., to bias correct and downscale future simulations) AOGCM biases will be preserved.

With the above setting in place, we can build a regression model in a MOS context, where simulated precipitation as predictor is bias corrected and downscaled. The deterministic part of the regression model (explained variance) would correct systematic biases, and the chosen noise model would describe the (unexplained) small-scale variability. In this study, the regression model comprises a logistic regression for wet day probabilities and a vector generalized linear model predicting the parameters of a mixture probability distribution for precipitation intensities. Simulating from the logistic model and the mixture model creates sequences of local-scale precipitation that explicitly include random small-scale variability.

3. Data

To test and illustrate our method, we bias correct and downscale daily precipitation simulated by the Consortium for Small-Scale Modeling in Climate Mode (COSMO-CLIM) version 4.8 RCM (Rockel et al. 2008) to a set of rain gauges across the United Kingdom separately for winter [December–February (DJF)] and summer [June–August (JJA)]. The simulation has been carried out by the Helmholtz Centre Geesthacht for the period 1961–2000 at a horizontal resolution of 0.22° (~25 km) over a rotated grid covering the European domain (van der Linden and Mitchell 2009). At the boundaries, the RCM is driven with the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis (NCEP1) data (Kalnay et al. 1996) (a so-called perfect boundary setting). Spectral nudging (von Storch et al. 2000) was applied for large-scale wind speed components in the upper levels.

For local-scale observations, against which the RCM is bias corrected, we used daily precipitation data from the Met Office Integrated Data Archive System (MIDAS) available from the British Atmospheric Data Centre (http://badc.nerc.ac.uk/view/badc.nerc.ac.uk__ATOM__dataent_ukmo-midas). We selected nine gauges to represent the nine precipitation regions across the United Kingdom (Wigley et al. 1984; Gregory et al. 1991). Within a region, the selection was arbitrary from a quality-checked set of gauges covering the simulation period (Maraun et al. 2008). Figure 1 shows the locations of these nine rain gauges.

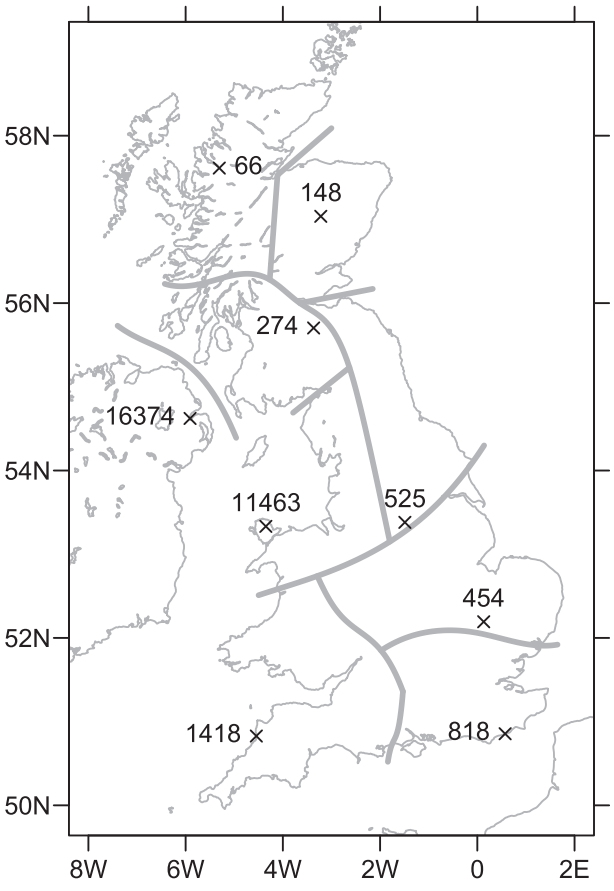

FIG. 1. Example gauges: Kinlochewe (source ID 66), Balmoral (source ID 148), Blyth Bridge (source ID 274), Belfast (source ID 16374), Anglesey (source ID 11463), Sheffield (source ID 525), Bude (source ID 1418), Cambridge (source ID 454), and Hastings (source ID 818). Gray lines represent borders between precipitation regions.

Figures 2 and 3 show quantile–quantile (QQ) plots of uncorrected RCM gridbox simulations against point observations for the nine selected gauges. The discrepancy between simulation and observation represents the overall effect of model biases and the representativeness problem. That is, in particular at the lower and upper tails, a considerable fraction of the discrepancy is caused by the scale mismatch between gridbox and point scale. As an example gauge for our detailed discussions, we

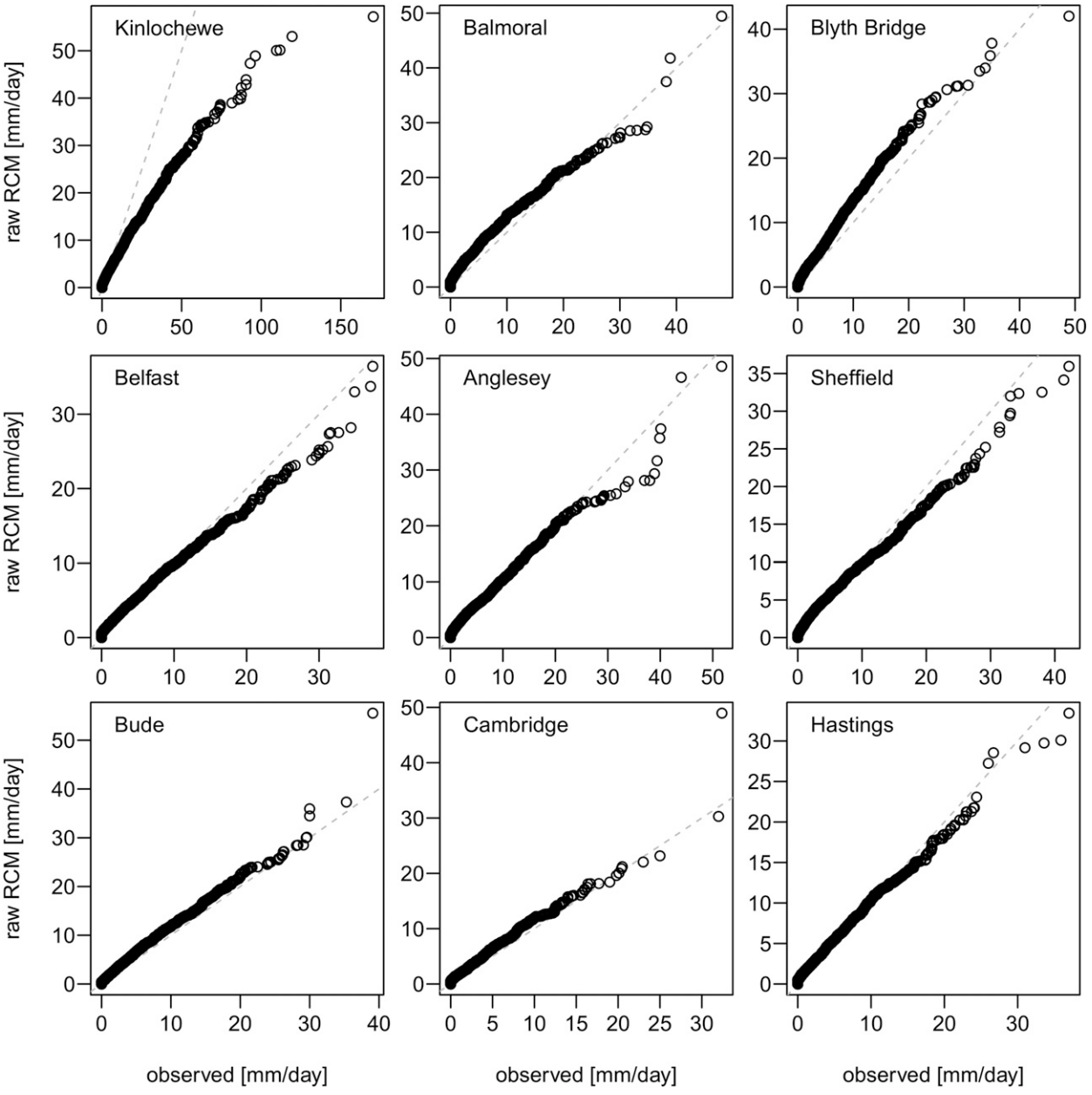

FIG. 2. Empirical QQ plots of RCM gridbox precipitation against observed precipitation for the nine example gauges for DJF.

consider Cambridge Botanic Garden (Met Office source identifier 454) throughout the whole manuscript. Extreme precipitation in Cambridge has its strongest intensities in summer with a slightly heavy-tailed distribution (Maraun et al. 2009).

Even in a perfect boundary setting, the trajectories of simulated weather systems might—randomly and systematically—slightly diverge from the observed trajectories. As precipitation additionally exhibits high spatial and temporal variability, the temporal correspondence between gridbox-simulated and observed local-scale daily precipitation is, in general, relatively weak. To increase the agreement, we therefore average the simulated precipitation (i.e., the predictor) across the square of nine neighboring grid boxes centered on the grid box containing the actual gauge.

For Cambridge, Fig. 4 demonstrates the overall good temporal correspondence of simulated (gray) and observed (black) precipitation. The wet–dry day sequences correspond very well on a day-to-day basis. The relative average intensities of individual wet periods also correspond well, but systematic biases stand out. In

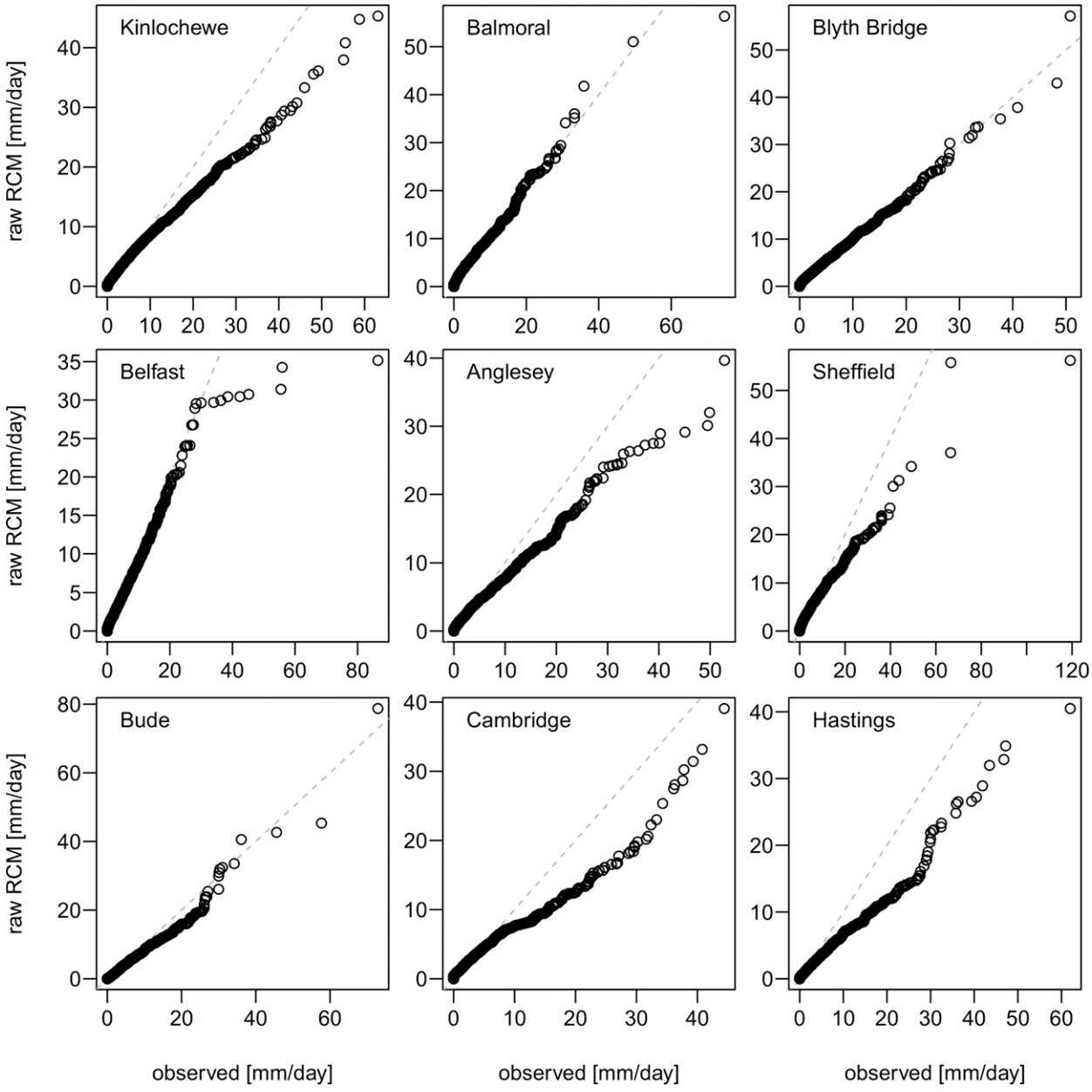

FIG. 3. As in Fig. 2, but for JJA.

particular, during summer, model values are on average much lower than observed values. Individual intensities show considerable random differences.

4. Statistical model

The implementation of our pairwise stochastic MOS approach is discussed in the following sections. First, we introduce the mixture probability distribution for modeling precipitation intensities in a stationary context. Second, we present the downscaling approach: a logistic regression model to predict wet day probabilities and the vector generalized linear model to predict precipitation intensities based on simulated precipitation.

a. Stationary model

Classical continuous distributions like the gamma distribution are commonly used to model precipitation intensities (Katz 1977). These distributions are able to model the bulk of the precipitation distribution but do not perform as well in modeling the extreme

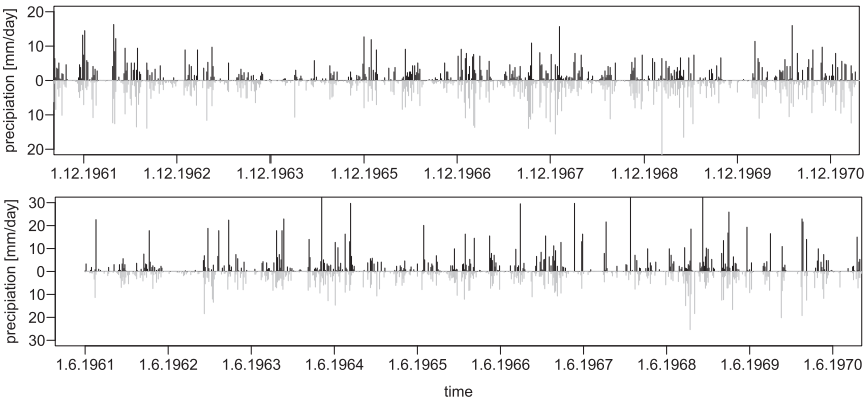

FIG. 4. Daily precipitation time series (section) for Cambridge for (top) DJF and (bottom) JJA. Observed (black) and raw RCM (gray) averaged across 3 × 3 grid boxes.

precipitation. The tail of the gamma distribution is, in general, too light to model high rainfall intensities and underestimates the extremes (e.g., Vrac and Naveau 2007; Maraun et al. 2010b). Hence, an extreme value distribution, such as the generalized Pareto distribution, might be required when modeling the extreme tails of the precipitation distribution above a certain threshold.

To consider both bulk and extreme tails of the precipitation distribution, Vrac and Naveau (2007) proposed the following stationary model of combining both gamma and GP distributions. This model is a variant of that of Frigessi et al. (2002). The distribution $l_\theta(r)$ of observed precipitation r on wet days is modeled as

$$l_\theta(r) = c(\boldsymbol{\theta})(\{[1 - w_{m,\tau}(r)]f_{\lambda,\gamma}(r)\} + [w_{m,\tau}(r)g_{\xi,\sigma}(r)]),$$
$$\boldsymbol{\theta} = (\lambda, \gamma, \xi, \sigma, m, \tau), \tag{1}$$

where $f_{\lambda,\gamma}$ is the probability density function (pdf) of the gamma distribution with rate parameter λ and shape parameter γ,

$$f_{\lambda,\gamma}(r) = \frac{\lambda^\gamma}{\Gamma(\gamma)} r^{\gamma-1} e^{-\lambda r}, \quad \lambda, \gamma > 0, \tag{2}$$

and $g_{\xi,\sigma}$ is the pdf of the GP distribution:

$$g_{\xi,\sigma}(r) = \frac{1}{\sigma}\left[1 + \frac{\xi(r-u)}{\sigma}\right]^{[-(1/\xi)-1]} \quad \text{when} \quad x \geq u. \tag{3}$$

Here, $\sigma > 0$ is the scale parameter and ξ is the shape parameter that influences the different tail behavior of the GP distribution: 1) for $\xi < 0$, the upper tail is bounded; 2) for $\xi = 0$, an (light tailed) exponential distribution is obtained; and 3) for $\xi > 0$, the upper tail is unbounded and is heavy tailed. Here, ξ was constrained to be strictly positive to ensure identifiability of the mixture model parameters. For the United Kingdom, this assumption is valid for most regions in general and all selected gauges in particular (Maraun et al. 2009).

The function $w_{m,\tau}$ is a weight function that represents the transition between the gamma and GP pdfs. It is expressed as

$$w_{m,\tau}(r) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{r-m}{\tau}\right), \quad m, \tau > 0, \tag{4}$$

with location parameter m, which denotes the location of the center of this transition, while τ affects the rapidity of transition between the two distributions. The weight function takes values in (0, 1) and is a nondecreasing function converging to 1 as rainfall r goes to ∞. At $w = 0.5$, there is an equal weight for the gamma and GP pdfs in the mixture model Eq. (1). This corresponds to the condition $r = m$. For small values of w, there is a greater emphasis placed on the gamma distribution; this corresponds to the case where $r < m$. Consequently, small rainfall values are captured predominantly by the gamma distribution. Conversely, for high values of w, there is more emphasis on the GP distribution; thus, heavy rainfalls are captured by the GP distribution. To create the mixture pdf, the mixture function must be normalized, and this is achieved by multiplying the mixture function by a constant $c(\boldsymbol{\theta})$.

In the mixture pdf in Eq. (1), the threshold u in the GP distribution is set to zero, as the location parameter m in the weight function fulfills the purpose of a threshold in Eq. (1). Vrac and Naveau (2007) attribute the advantages of having a weight function and fixing the threshold to zero to 1) solving the difficult threshold selection problem with an unsupervised estimation procedure and 2) avoiding a discontinuity in the pdf $l_\theta(r)$, which may occur when nonzero thresholds are allowed.

As wet days, we consider days with precipitation above $1\,\mathrm{mm\,day^{-1}}$. As the support of the gamma distribution comprises $(0, \infty)$, a nonzero wet day threshold would assign zero probability density to all intensities ranging from zero to that threshold and lead to a serious misfit of the gamma distribution for small intensities. We therefore account for the ignored values in $(0, 1]$ by shifting all precipitation events on wet days by $-1\,\mathrm{mm\,day^{-1}}$ for calibration and shifting the estimated distribution back by $+1\,\mathrm{mm\,day^{-1}}$. The convincing QQ plots for low precipitation intensities above the wet day threshold (see section 5) demonstrate that this procedure is justified.

b. Nonstationary model for downscaling

In our context, downscaling refers to predicting the distribution of local-scale precipitation r_i at time step $i = 1, \ldots, n_t$ at a certain rain gauge by using the precipitation x_i, simulated at the gridbox level by an RCM. Precipitation downscaling on a daily time scale, in general, consists of two steps. Given a predictor value, first the precipitation occurrence is downscaled. Conditional on a wet day, the precipitation intensity is downscaled in a second step.

1) Downscaling precipitation occurrence

A logistic regression is often used to model the changing probability of rainfall occurrence (Chandler and Wheater 2002). The logistic regression model belongs to the class of generalized linear models (GLMs), which are a generalization of simple linear regression: the time-dependent expectation μ_i of a random variable is linked via a monotonic link function $g(\cdot)$ to a linear combination of n_p predictors $x_{1,i}, \ldots, x_{p,i}$ (Dobson 2001),

$$g(\mu_i) = \alpha_0 + \sum_{j=1}^{n_p} \alpha_j x_{j,i}, \tag{5}$$

where $\alpha_0, \ldots, \alpha_{n_p}$ are regression coefficients. In our case, the probability p_i that a day i is wet is modeled as a function of simulated RCM precipitation x_i,

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \alpha_1 x_i, \tag{6}$$

where $g(\cdot)$ is the so-called logit link function and α_0 and α_1 are coefficients to be estimated. Hence, the probability p_i of a day i being wet can be expressed as

$$p_i = \frac{\exp(\alpha_0 + \alpha_1 x_i)}{1 + \exp(\alpha_0 + \alpha_1 x_i)}. \tag{7}$$

2) Downscaling precipitation intensities

To model the influence of our predictor, RCM-simulated precipitation x_i, on the parameters of the mixture model, we employ a vector generalized linear model (VGLM) as regression model (Yee and Wild 1996; Yee and Stephenson 2007). The general idea of VGLMs is to predict n_θ distribution parameters $\theta_{k,i}$, $k = 1, \ldots, n_\theta$, for each time step i,

$$g_k(\theta_{k,i}) = \beta_{k,0} + \sum_{j=1}^{n_p} \beta_{k,j} x_{j,i}, \tag{8}$$

where $g_k(\cdot)$ represents the link functions for each distribution parameter $x_{j,I}$; $i = 1, \ldots, n_p$ represents the predictors; and $\beta_{k,j}$ represents the VGLM coefficients. In our particular case, the VGLM reads as follows:

$$\begin{aligned} \lambda_i &= \lambda_0 + \lambda_1 x_i \\ \gamma_i &= \gamma_0 + \gamma_1 x_i \\ \sigma_i &= \sigma_0 + \sigma_1 x_i \\ \xi_i &= \xi_0 \\ m_i &= m_0 + m_1 x_i \\ \tau_i &= \tau_0 . \end{aligned} \tag{9}$$

It is calibrated separately for each rain gauge. No link function was chosen, first because predictor and predictand are the same physical variables. Furthermore, an initial formulation with an exponential link function, in several cases, resulted in unrealistically high predictions for high predictor values. In principle, this could lead to negative parameter values and therefore a failure of the calibration. This case, however, did not occur (even for moderate extrapolation; see section 5b). Estimates of the shape parameter ξ are often rather uncertain because they are dominated by the most extreme and therefore rare values. If the shape parameter is modeled dependent on predictors, the problem is exacerbated; the estimated regression coefficients are highly uncertain and often indistinguishable from zero, but they still might produce unphysical predictor–predictand relationships. Therefore, in line with general practice, the shape parameter is kept constant (but different for each gauge). Initial analyses showed that the downscaling results were insensitive to an influence of the predictor on τ. Therefore, τ was also kept constant.

To obtain the distribution function of precipitation R_i, we combine the probability of wet day occurrence p_i from Eq. (7) and the mixture model distribution, which defines our precipitation intensities, to give

$$\Pr(R_i \le r) = \Pr(R_i \le r \,|\, W) p_i + (1 - p_i), \tag{10}$$

where $\Pr(R_i \le r \,|\, W)$ is the corresponding cumulative distribution function of the mixture model distribution.

The parameters in Eqs. (7) and (9) are estimated using maximum likelihood estimation (MLE).

c. Model selection

The full model given by the logistic regression Eq. (7) and the mixture model Eq. (9) is quite complex and, given the limited amount of calibration data, susceptible to overfitting. Therefore, we carried out a systematic statistical model selection to reduce the complexity of the model to a justified degree. As potential candidate models, we consider all simplifications of the full model that remove the influence of the predictor on a model parameter, including the stationary mixture. Additionally we consider a simplified version of the model that does not include the Pareto distribution for the tail and thus does not explicitly account for extremes. This model employs only the gamma distribution and thus simplifies to the following:

$$\lambda_i = \lambda_0 + \lambda_1 x_i$$
$$\gamma_i = \gamma_0 + \gamma_1 x_i. \tag{11}$$

In the following, we refer to this model as the VGLM gamma model. The candidate models are compared by means of statistical model selection criteria that assess whether the improvement in likelihood by increasing the model complexity is justified by the increased number of parameters. To select the optimal model structure, different information criteria exist for different settings. In our case, where the set of candidate models most likely does not include the hypothetical true model, the Akaike information criterion (AIC; Akaike 1973) is a suitable choice: in this case, it asymptotically (for an infinite number of data points) selects the model that minimizes the mean squared error of prediction (Shao 1997). The AIC is defined as $-2\log(L) + 2k$, where L is the likelihood corresponding to the maximum likelihood estimate of the k model parameters. To avoid calibrating all possible candidate models, we carry out a backward elimination (Davison 2003). Starting from the full VGLM mixture model Eq. (9), in each step the parameter is omitted that minimizes the AIC until no omission further improves the AIC. In addition, we explicitly calculate the AIC for the stationary mixture model and the VGLM gamma.

5. Results

In this section, we first present the final selected model structure for the nine selected gauges and assess its goodness of fit. Second, we assess the downscaling performance of the selected model. Within the assessment, we evaluate the downscaling performance by comparing the model's predictive power with the climatology. Moreover, we compare the VGLM mixture model to the simpler VGLM gamma model. The latter comparison is relevant for two reasons: 1) it shows whether explicitly including an extreme value model improves the representation of extreme precipitation events but 2) it also shows for what range of values the simple and computationally efficient VGLM gamma model provides a feasible alternative to the complex VGLM mixture model. Finally, we present an example application for Cambridge. Our stochastic MOS method aims to predict local precipitation and is thus conceptually closely related to weather forecasting. Both goodness of fit and model performance can thus be assessed from a forecast verification perspective. In the following sections, we will therefore also discuss the quality of our model with respect to the forecast verification attributes of reliability, resolution, and sharpness (Wilks 2006; Jolliffe and Stephenson 2003).

TABLE 1. Akaike information criterion values of VGLM mixture model with different set of parameters fixed for Cambridge (source ID 454) for DJF, 1961–2000. The AIC of the finally selected model structure is highlighted in bold.

Parameters fixed	AIC
Eq. (2)	4843
$\xi_1, \tau_1, \gamma_1 = 0$	4847
$\xi_1, \tau_1, \lambda_1 = 0$	4837
$\xi_1, \tau_1, m_1 = 0$	4835
$\xi_1, \tau_1, m_1, \gamma_1 = 0$	**4825**
$\xi_1, \tau_1, m_1, \gamma_1, \lambda_1 = 0$	4848
$\xi_1, \tau_1, m_1, \gamma_1, \sigma_1 = 0$	4843
$\xi_1, \tau_1, m_1, \lambda_1 = 0$	4833
$\xi_1, \tau_1, m_1, \sigma_1 = 0$	4836
$\xi_1, \tau_1, \sigma_1 = 0$	4845
$\xi_1, \tau_1, \sigma_1, \gamma_1 = 0$	4848
$\xi_1, \tau_1, \sigma_1, \lambda_1 = 0$	4842
$\xi_1, \tau_1, \sigma_1, m_1 = 0$	4836

a. Model selection results and goodness of fit

We carry out the model selection separately for winter and summer across all nine example gauges. The selection procedure is illustrated for the rain gauge at Cambridge in Tables 1 and 2. The tables list AIC values for all considered candidate models; the AIC of the finally selected model structure is shown in boldface.

For each of the nine rain gauges, a different model structure is chosen based on the systematic model selection approach. Tables 3 and 4 show the AIC values for the stationary mixture model Eq. (1), the VGLM gamma model Eq. (11), and the selected VGLM mixture model.

TABLE 2. As in Table 1, but for JJA.

Parameters fixed	AIC
Eq. (2)	4851
$\xi_1, \tau_1, \gamma_1 = 0$	4868
$\xi_1, \tau_1, \lambda_1 = 0$	4849
$\xi_1, \tau_1, \lambda_1, \gamma_1 = 0$	4866
$\xi_1, \tau_1, \lambda_1 m_1 = 0$	**4847**
$\xi_1, \tau_1, \lambda_1, m_1, \gamma_1 = 0$	4864
$\xi_1, \tau_1, \lambda_1, m_1, \sigma_1 = 0$	4852
$\xi_1, \tau_1, \lambda_1, \sigma_1 = 0$	4853
$\xi_1, \tau_1, m_1 = 0$	4849
$\xi_1, \tau_1, \sigma_1 = 0$	4856
$\xi_1, \tau_1, \sigma_1, \gamma_1 = 0$	4874
$\xi_1, \tau_1, \sigma_1, \lambda_1 = 0$	4853
$\xi_1, \tau_1, \sigma_1, m_1 = 0$	4854

Except for Balmoral in summer (where the AIC of the stationary is slightly lower than that of the VGLM mixture model), the stationary mixture model has been selected for none of the gauges.[1] That is, our model has predictive power and, thus, the stochastic MOS approach—predicting local precipitation from gridbox precipitation—is, in principle, feasible. The simpler VGLM gamma model seems to suffice (according to the AIC) for all but one gauge in winter, whereas the VGLM mixture model is required for most gauges in summer. The reason for this difference is likely the different processes governing extreme precipitation in different seasons; In winter, extreme precipitation is often associated with large-scale weather systems that are well simulated by climate models. In summer, precipitation extremes are often caused by subgrid convective events.

We assess the (absolute) goodness of fit of our stochastic MOS downscaling model using residual QQ plots, where standardized empirical quantiles are plotted against standardized theoretical quantiles. For comparison, we also consider the VGLM gamma model Eq. (11). A QQ plot should only be used for quantiles of an unconditional distribution. As the predicted distribution varies from day to day with simulated precipitation, observations and model distributions have therefore been standardized to the stationary gamma distribution (see, e.g., Coles 2001). As the standardization shifts both model and observation according to the strength of the predictor for any particular event, the absolute values in the QQ plots can only be seen as a rough guide. Figures 5 and 6 show the QQ plots of all wet days for the VGLM gamma (blue circles) and VGLM mixture (black circles)

[1] Note that AIC values are calculated for the intensity distributions conditional on wet days; that is, the AIC does not assess the occurrence model. Thus, stationarity here refers to "constant intensity distribution, given a wet day."

TABLE 3. Akaike information criterion for the nine example gauges for DJF, 1961–2000. For the VGLM mixture model, for each gauge the finally selected model structure is chosen. The optimal model based on the AIC value is highlighted in bold. Note that AIC values are given, conditional on wet days, also for the stationary model.

Station	Stationary mixture	VGLM gamma	VGLM mixture
Kinlochewe	15 650	**14 855**	14 905
Balmoral	7786	7704	**7701**
Blyth Bridge	8129	**8040**	8054
Belfast	7963	**7792**	7803
Anglesey	9024	**8861**	8875
Sheffield	7718	**7638**	7640
Bude	8958	**8865**	8889
Cambridge	4845	**4821**	4825
Hastings	6712	**6605**	6626

models for all nine rain gauges, for winter and summer seasons, respectively. For all rain gauges in winter and summer, both the VGLM mixture and VGLM gamma models are capable of reproducing the observed quantiles for the bulk of the distribution. In other words, both models effectively correct systematic biases for low and medium precipitation. Consistent with the AIC results (Table 3), the VGLM gamma suffices to describe the tail for most gauges during winter but diverges considerably from the observations at higher quantile ranges. A particularly strong divergence of the VGLM gamma model occurs for Kinlochewe in winter but, because of the relatively poor performance of the VGLM mixture model for medium intensity precipitation (see Fig. 5), the gamma model produces a better AIC value. As the standardization procedure implicitly accounts for the effect of the predictor, the QQ plots also indicate that both models produce reliable and well-calibrated predictions (i.e., they exhibit low conditional bias). For the mixture model, this is also valid for heavy precipitation.

For our example gauge at the Cambridge Botanic Garden, both models effectively describe the bulk of the

TABLE 4. As in Table 3, but for JJA.

Station	Stationary mixture	VGLM gamma	VGLM mixture
Kinlochewe	10 391	**10 167**	10 170
Balmoral	**5691**	5754	5691
Blyth Bridge	7084	**7036**	7038
Belfast	6211	6201	**6163**
Anglesey	6493	6428	**6428**
Sheffield	5589	5581	**5549**
Bude	6084	6044	**6016**
Cambridge	4871	4902	**4847**
Hastings	4700	4700	**4673**

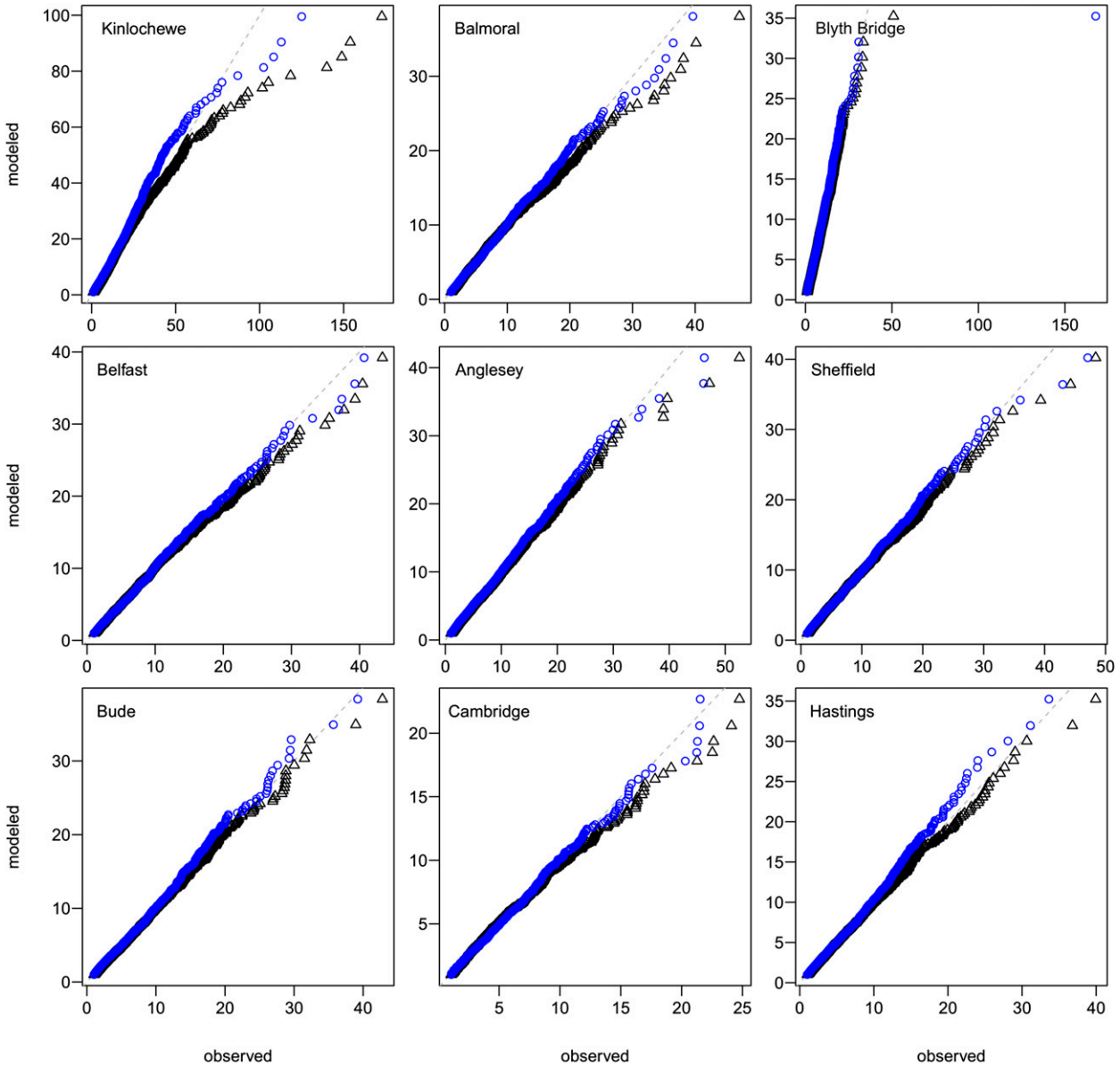

FIG. 5. QQ plots (mm day^{-1}) for the nine example gauges for DJF. Standardized to stationary gamma distribution fitted to observed wet day intensities. VGLM gamma model (black triangles) and VGLM mixture model (blue circles).

observations in winter. While the mixture model describes almost all observations very well, the gamma model diverges more strongly from the empirical distribution for high quantiles. During summer, the VGLM mixture model almost perfectly describes the observed precipitation distribution from low intensities toward extremes, whereas the VGLM gamma model overestimates low intensities but considerably underestimates moderate to high intensities.

b. Performance of the stochastic MOS approach

In the context of our stochastic MOS, downscaling refers to the prediction of local-scale precipitation from RCM-simulated gridbox precipitation. Beyond model selection and goodness of fit, mainly two questions are of interest: How is the predictive power of our model compared to the climatology? How do predictions with the VGLM mixture model differ from predictions based on the simpler VGLM gamma model? We assess both questions visually by plotting the predicted distribution conditional on simulated precipitation and quantitatively by means of skill scores in a cross validation.

Figures 7a and 7b show the dependence of the predicted distribution on simulated precipitation for winter and summer, respectively, at Cambridge. Depicted is

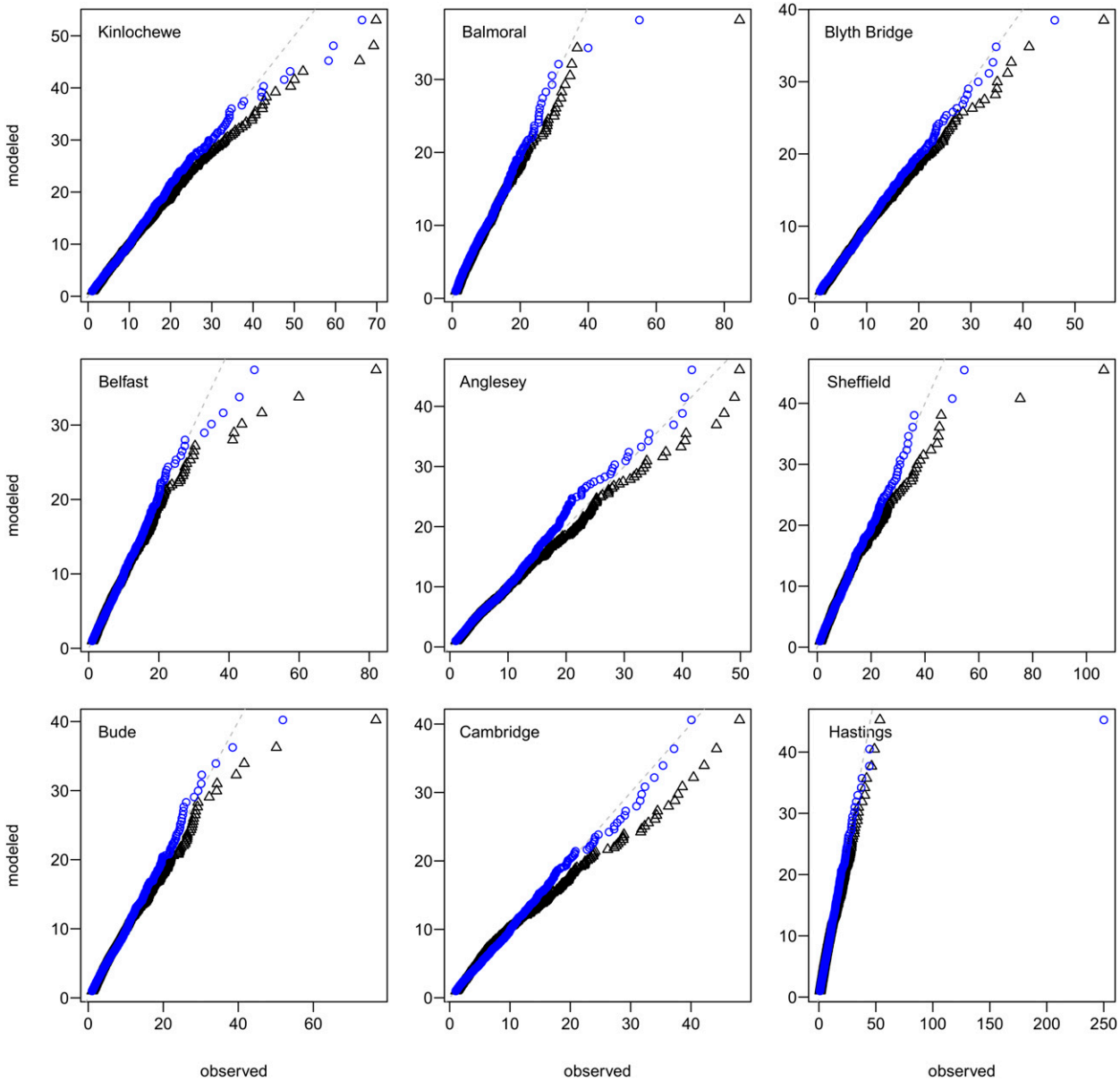

FIG. 6. As in Fig. 5, but for JJA.

a selection of quantiles from the VGLM mixture model (solid lines) and the VGLM gamma model (dashed lines); for comparison, the corresponding quantiles of the stationary model (climatology; dotted lines) are also shown. Given a simulated precipitation value, local precipitation will fall below a certain quantile with the corresponding (color coded) probability. The quantiles get smaller with smaller simulated precipitation (predictor) values and are cut off below 1 mm day^{-1} by the chosen wet day threshold.

The distribution of RCM-simulated predictor values is indicated by selected sample quantiles (vertical dashed–dotted lines). Obviously, the predicted distribution for both winter and summer depends strongly on simulated precipitation. For instance, in winter, the dry day probability decreases from more than 75% for zero simulated precipitation (blue line) to 25% for about 9 mm day^{-1} simulated precipitation (magenta line).[2] Similarly, the probability of exceeding 10 mm day^{-1} changes from roughly 10% for 10 mm day^{-1} simulated

[2] The 75% quantile is zero for zero simulated precipitation; the 25% quantile is just zero for 9 mm day^{-1} simulated precipitation.

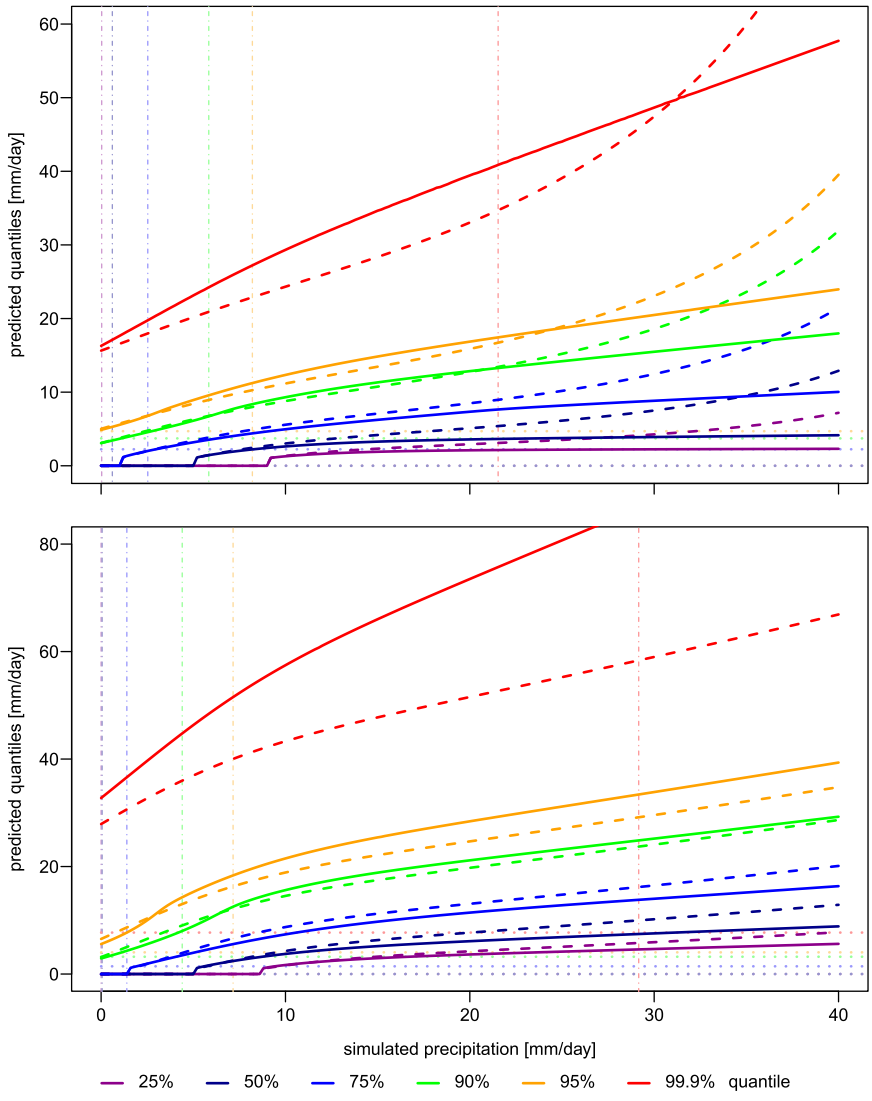

FIG. 7. Predicted quantiles as function of simulated precipitation at Cambridge for (top) DJF and (bottom) JJA: VGLM mixture model (solid), VGLM gamma model (dashed), stationary mixture model (dotted), and corresponding sample quantiles of the RCM-simulated predictor (dashed–dotted).

precipitation (green line) to roughly 25% for 40 mm day^{-1} simulated precipitation (blue line).[3]

In terms of forecast attributes, our prediction has a considerable resolution, as the predicted distribution changes strongly with different simulated precipitation values. In principle, the behavior with respect to resolution lies between two extreme cases. The climatology (dotted lines) is independent of any predictor. It therefore has only little predictive power and no resolution at all. The higher the predictive power of the model, the more strongly the quantiles would depend on simulated precipitation (i.e., the steeper they would slope toward high simulated values). Also, the fraction of unexplained variance, which is basically given by the width of the distribution, would shrink. With higher predictive power, the individual quantiles would grow closer together and finally collapse to a single deterministic prediction in the case of perfect local predictability. The relative slope of and distance between the individual quantiles should thus give a rough idea of the predictive power of our model for these quantiles, in terms of

[3] The 10 mm day^{-1} corresponds to the 90% quantile for 10 mm day^{-1} simulated precipitation and to the 75% quantile for 40 mm day^{-1} simulated precipitation.

resolution and sharpness. Notably, the relative slopes of the winter and summer quantiles are very similar (e.g., the change of the 95th percentile from 10 to 40 mm day^{-1} simulated precipitation is roughly a factor of 2 for winter and summer), suggesting a similar performance.

For winter, the VGLM gamma model (dashed lines) agrees well with the VGLM mixture model for low values of predicted precipitation, confirming the conclusions from the QQ plots. Divergence from the mixture model occurs in two different ways. For the bulk of simulated precipitation values, the gamma model predicts lower extremes than the mixture model: see the 99.9th percentile up to a range of about 30 mm day^{-1} of simulated precipitation. This deviation is expected because of the light tail of the gamma distribution. Considering the deficiencies of the gamma model for high quantiles revealed by the QQ plots, this deviation further indicates that the gamma model should be used with care for extrapolations beyond the range of observed values. For high values of simulated precipitation, however, the gamma model shows a completely different behavior. In this range, the model is not well constrained by data (see the vertical lines indicating quantiles of simulated precipitation) and is—as a result of its stiff parameterization—mostly determined by data in the lower range of simulated precipitation. In these rare cases, the conditional gamma distribution is very broad and predicts even higher values than the mixture distribution for basically all quantiles (for very high quantiles, of course, the light tail of the gamma distribution dominates again; not shown). The fact that the model selection procedure favored the more flexible VGLM mixture model indicates that this behavior of the VGLM gamma model is an artifact of the too-rigid model structure. For summer, the behavior of the VGLM gamma model is qualitatively similar, but the divergence from the mixture model for high simulated precipitation sets in much later. In fact, the divergence toward too-high predicted intensities is stronger for lower quantiles than for high quantiles. The effect of the light tail is clearly visible already in the 95th percentile. This underestimation is in accordance with the results from the QQ plots (see Fig. 6).

To quantify the predictive performance of our model, we use skill scores developed in the context of forecast verification (Wilks 2006; Jolliffe and Stephenson 2003). Skill scores measure the performance of a forecast relative to a reference forecast. They are designed to range from 1 for a perfect forecast, through 0 for one that does not provide any improvement over the reference, and to negative values for forecasts performing worse than the reference. The actual evaluation is carried out as cross validation. For this purpose, our dataset of 40 seasons is separated into training periods of 30 seasons and testing periods of 10 seasons. Four nonoverlapping testing periods are chosen, starting from the first 10-season period, while the training periods are chosen accordingly. The skill scores are calculated for the merged 40-season sequence of the four consecutive testing periods. Confidence intervals of the skill scores were calculated by a nonparametric bootstrap approach following Jolliffe (2007): the cross-validated time series of predicted quantiles and predictands were resampled 1000 times with replacement, and 1000 skill score values were derived to approximate 95% confidence intervals.

TABLE 5. Brier skill scores (%) of the logistic regression for wet day probabilities against the climatological wet day probability. The 95% confidence intervals are given in parentheses.

Station	DJF		JJA	
Kinlochewe	38	(35, 40)	23	(21, 26)
Balmoral	12	(10, 14)	9	(7, 11)
Blyth Bridge	15	(13, 17)	10	(8, 12)
Belfast	23	(20, 25)	17	(14, 19)
Anglesey	19	(17, 21)	15	(12, 17)
Sheffield	15	(13, 17)	11	(9, 13)
Bude	22	(19, 24)	18	(15, 20)
Cambridge	13	(10, 15)	13	(11, 16)
Hastings	18	(15, 20)	15	(13, 18)

To examine the capability of our logistic model to predict dry and wet days (defined as more than 1 mm of precipitation), we employ the Brier score (BS) (e.g., Wilks 2006). The Brier score measures the averaged squared error between N pairs of probabilistic forecasts (f_i) and binary observations (o_i), where f_i is the predicted wet day probability p_i from Eq. (7), a wet day observed is $o_i = 1$, and a dry day is $o_i = 0$,

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2. \tag{12}$$

The Brier skill score (BSS) measures the improvement of the Brier score of the considered model relative to that of a reference model $\mathrm{BS}_{\mathrm{ref}}$,

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{logistic}}}{\mathrm{BS}_{\mathrm{ref}}}. \tag{13}$$

Here we consider the climatological wet day probability as reference model.

The resulting BSS are shown in Table 5. Consistently positive values reflect that our model predicts the wet and dry day sequence considerably better than the climatology. This result indicates that precipitation simulated by a spectrally nudged RCM driven with perfect

TABLE 6. Quantile skill scores (%) of the VGLM mixture model for different quantiles against the best climatological model (either stationary mixture or gamma) for (top) winter and (bottom) summer. The 95% confidence intervals are given in parentheses.

(DJF) Station	0.5		0.75		0.9		0.95		0.98	
Kinlochewe	30	(28, 31)	36	(34, 38)	35	(31, 37)	32	(26, 33)	28	(18, 28)
Balmoral	8	(7, 10)	13	(12, 15)	15	(13, 19)	13	(10, 18)	12	(7, 18)
Blyth Bridge	11	(9, 12)	15	(13, 17)	13	(10, 16)	11	(9, 15)	9	(6, 16)
Belfast	18	(16, 20)	25	(23, 28)	23	(20, 26)	22	(17, 25)	17	(11, 22)
Anglesey	15	(13, 17)	19	(18, 22)	17	(14, 20)	16	(12, 20)	15	(9, 21)
Sheffield	10	(8, 11)	15	(13, 17)	14	(11, 17)	14	(9, 16)	11	(6, 16)
Bude	15	(14, 17)	19	(17, 21)	16	(13, 18)	12	(9, 15)	5	(4, 11)
Cambridge	6	(4, 8)	12	(9, 14)	11	(8, 14)	12	(7, 15)	13	(6, 19)
Hastings	11	(9, 13)	21	(18, 23)	19	(16, 22)	17	(14, 21)	11	(7, 16)
Mean	14		20		18		16		13	

(JJA) Station	0.5		0.75		0.9		0.95		0.98	
Kinlochewe	19	(17, 20)	23	(21, 25)	24	(21, 27)	21	(19, 26)	17	(13, 24)
Balmoral	3	(3, 4)	10	(8, 11)	11	(9, 13)	11	(8, 14)	8	(6, 13)
Blyth Bridge	4	(3, 6)	11	(9, 13)	11	(9, 14)	10	(7, 13)	9	(4, 12)
Belfast	7	(6, 9)	17	(14, 19)	15	(12, 18)	13	(8, 17)	12	(7, 17)
Anglesey	5	(4, 7)	17	(14, 20)	19	(16, 22)	18	(14, 22)	15	(9, 20)
Sheffield	3	(2, 5)	11	(9, 13)	15	(11, 17)	12	(8, 16)	11	(5, 16)
Bude	8	(6, 9)	17	(15, 20)	16	(14, 20)	14	(11, 18)	9	(3, 14)
Cambridge	3	(2, 5)	14	(11, 16)	16	(12, 19)	15	(11, 19)	10	(6, 15)
Hastings	4	(2, 6)	17	(14, 19)	15	(12, 19)	16	(11, 21)	16	(10, 23)
Mean	6		15		16		15		12	

boundary conditions is an informative predictor for observed local-scale wet day probabilities.

To quantify the capability of our model to predict specific quantiles, we employ the quantile skill score (QSS; Friederichs and Hense 2007; Friederichs and Thorarinsdottir 2012). For a given set of observations y_i and predictors x_i, where $i = 1, \ldots, N$, the quantile score (QS) for the predicted α-quantile q_α as a function of the predictors x_i is defined as the weighted average of the distance of each observation from the α-quantile estimate,

$$\mathrm{QS}_\alpha = \sum_{i=1}^{N} \rho_\alpha[y_i - q_\alpha(x_i)], \tag{14}$$

where

$$\rho_\alpha(u) = \begin{cases} \alpha u & \text{for} \quad u \geq 0; \\ (\alpha - 1)u & \text{for} \quad u < 0. \end{cases} \tag{15}$$

The QS measures the resolution and reliability of a conditional quantile forecast and thus penalizes noninformative and biased quantile forecasts (Friederichs 2010). The quantile skill score compared to a reference model is defined as

$$\mathrm{QSS}_\alpha = 1 - \frac{\mathrm{QS}_\alpha}{\mathrm{QS}_{\alpha,\mathrm{ref}}}. \tag{16}$$

First we quantify the performance of our model relative to the climatology for selected quantiles at all example gauges (see Table 6). Here, the climatology is defined as the combination of climatological wet day probability and either stationary mixture or gamma model, depending on which of the two yields the better quantile score (i.e., the best stationary model is chosen as reference). For both winter and summer, the QSS is consistently positive and ranges for most quantiles on average between 12% and 20%. In general, it is highest for medium-high intensities (75th to 90th percentile), with slightly lower values for extremes. Interestingly, the relative performance in predicting the median is apparently weak, in particular for summer. The reason for this behavior might be that the (time dependent) median is often zero and is thus well predicted also by the climatological distribution, in particular for a low climatological wet day probability, as in summer.

Second, we choose the VGLM gamma model as a reference to quantify the improvement of explicitly accounting for extremes (see Table 7). For both winter and summer, the improvement compared to the VGLM gamma model is negligible. This finding contrasts considerably with the QQ plots, in which, for both winter and summer, the 99th percentile (e.g., around 12 mm day^{-1} for winter and 20 mm day^{-1} for summer in the QQ plot for Cambridge) is considerably underestimated by the gamma model yet relatively well captured by the mixture model. In other words, biases in high quantiles are reduced by the mixture model. As stated above, the QS rewards low biases and penalizes

TABLE 7. As in Table 6, but against VGLM gamma model.

(DJF) Station	0.5		0.75		0.9		0.95		0.98	
Kinlochewe	−1	(−3, 0)	−1	(−3, 0)	−2	(−6, −2)	−4	(−11, −5)	−9	(−23, −9)
Balmoral	0	(0, 1)	0	(0, 1)	0	(0, 1)	−1	(−1, 2)	0	(−3, 3)
Blyth Bridge	0	(0, 0)	0	(0, 1)	0	(0, 1)	−1	(−1, 1)	−2	(−2, 2)
Belfast	1	(0, 1)	0	(−1, 1)	0	(−1, 1)	−1	(−3, 1)	−3	(−9, 1)
Anglesey	0	(0, 1)	0	(0, 1)	−1	(−1, 0)	−1	(−2, 0)	−3	(−5, −1)
Sheffield	0	(0, 1)	0	(0, 1)	0	(−1, 0)	0	(−2, 1)	−1	(−4, 2)
Bude	0	(0, 0)	0	(−1, 0)	0	(−1, 0)	0	(−1, 1)	−2	(−2, 1)
Cambridge	0	(0, 0)	0	(0, 1)	0	(−1, 2)	2	(−1, 4)	3	(−2, 10)
Hastings	0	(0, 0)	0	(0, 1)	0	(0, 0)	0	(−1, 1)	−3	(−5, 0)
Mean	0		0		0		−1		−2	
(JJA) Station	0.5		0.75		0.9		0.95		0.98	
Kinlochewe	1	(0, 3)	0	(0, 2)	0	(0, 2)	−1	(−1, 2)	−3	(−5, 2)
Balmoral	1	(−1, 3)	−1	(−3, 1)	−3	(−5, −1)	−3	(−4, −1)	−4	(−5, 0)
Blyth Bridge	0	(0, 1)	0	(0, 1)	1	(0, 2)	2	(0, 3)	2	(−2, 4)
Belfast	1	(0, 1)	0	(0, 0)	0	(0, 1)	1	(0, 1)	1	(−2, 2)
Anglesey	1	(0, 2)	0	(0, 1)	1	(0, 1)	2	(0, 3)	1	(−3, 3)
Sheffield	0	(0, 1)	1	(0, 1)	1	(0, 2)	1	(0, 3)	2	(0, 5)
Bude	1	(0, 1)	0	(0, 0)	−1	(0, 0)	0	(0, 2)	−1	(−5, 1)
Cambridge	0	(0, 1)	0	(−1, 0)	0	(−1, 0)	2	(0, 3)	−1	(−3, 1)
Hastings	1	(0, 2)	0	(0, 1)	0	(−1, 1)	1	(0, 3)	1	(0, 4)
Mean	1		0		0		1		0	

noninformative predictions. The fact that the QSS of the VGLM mixture relative to the VGLM gamma model is negligible thus indicates that the gain in QSS by the decrease in bias is outweighed by a decrease in predictive power of the model. The latter effect is likely caused by the higher number of parameters in the VGLM mixture model. Note that the QSS only assesses the model performance for the range of observed values and does not state anything about the performance when extrapolating to unobserved extremes.

c. Example application

In this section, we present an initial application of our stochastic MOS downscaling model to the rain gauge data at Cambridge Botanic Garden for both winter and summer. Based on the calibrated VGLM mixture model, we use RCM-simulated precipitation to predict the local precipitation distribution at each day. Figure 8 shows a range of quantiles from the 50th to 95th percentile (colored lines) compared to the actual observations (black spikes) for the winters (top) and summers (bottom) from 1961 to 1969. Overall, our model exhibits a considerable sharpness; it deviates strongly from the climatology (constant quantiles) and shows long dry spells (where the 75th percentile is small and the median essentially zero) as well as extreme events. In general, the predictions of dry and wet spells are accurate compared to the actual observed spells, consistent with the results for the Brier skill scores (Table 5). Also, higher intensities and extreme events are well predicted, consistent with the quantile skill score results (Table 6).

6. Conclusions

We developed a stochastic MOS approach for bias correcting and downscaling climate model output. The key idea is to use RCM-simulated precipitation as a predictor for the full local-scale intensity distribution ranging from dry days to extreme events. To enable the calibration of such a regression model, a pairwise calibration is necessary where predictor and predictand correspond on a day-by-day basis. In our implementation, temporal correspondence was ensured by perfect boundary conditions for the RCM from NCEP1 data and additional spectral nudging.

Traditional bias correction methods in climate research are deterministic and therefore only correct systematic biases but do not account for unexplained local-scale variability (Maraun 2013). That is, these methods, by construction, cannot overcome representativeness problems. Our approach is probabilistic and, by generating random small-scale variability, can additionally downscale processes highly variable in space and time such as precipitation from the grid scale to the point scale.

Our specific implementation employs a logistic regression to model wet day probabilities, with RCM-simulated precipitation as a predictor. A mixture model (Frigessi et al. 2002; Vrac and Naveau 2007) is used to

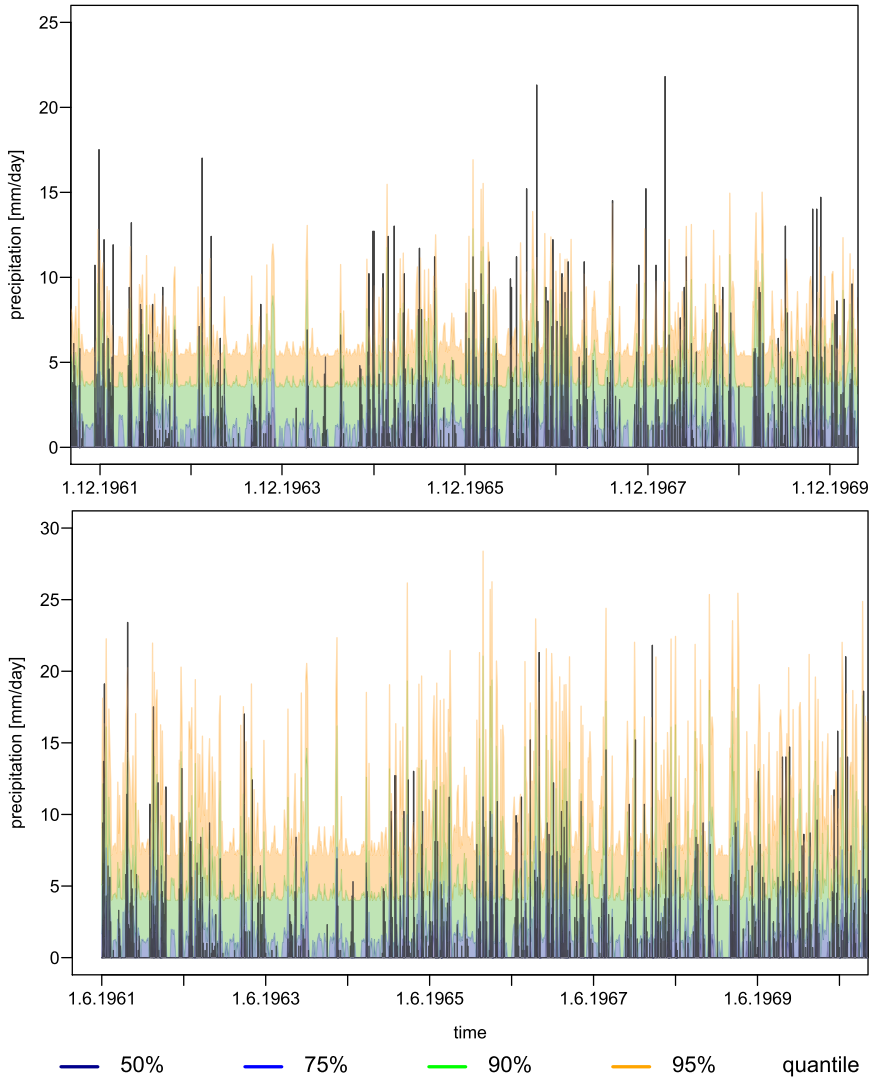

FIG. 8. Predicted time series of quantiles at Cambridge for (top) DJF and (bottom) JJA: observations (black spikes) and predicted quantiles of the VGLM mixture model (color).

describe precipitation intensities; that is, moderate precipitation is represented by a gamma distribution, and extremes are represented by a generalized Pareto (GP) distribution. RCM-simulated precipitation is included as predictor for the intensities via a vector generalized linear model (VGLM) (Yee and Stephenson 2007; Maraun et al. 2010a, 2011). The specific model structure—which model parameters are affected by the predictor—had to be selected individually for each season and rain gauge based on the Akaike information criterion. The proposed model can, in principle, also be used for weather forecasting: for example, as an extension of models describing only the bulk of a distribution (e.g., Thorarinsdottir and Johnson 2012) or solely extreme events (e.g., Friederichs 2010). QQ plots revealed that the VGLM mixture model effectively corrects systematic biases and provides a well-calibrated estimate of the local precipitation distribution for a wide range of quantiles.

In our context, downscaling refers to predicting the distribution of small-scale precipitation from gridbox-simulated precipitation. We found that the predicted quantiles of our mixture model depend strongly on the predictor, implying a considerable forecast resolution (predictive power) and thus downscaling capability. This finding was further quantified by skill scores: our model substantially improves the Brier skill score for predicting local-scale wet and dry days compared to the climatology. Furthermore, our model considerably improves the quantile score for predicting a wide range of quantiles relative to the climatology. In terms of forecast

verification attributes, our model thus provides a well-calibrated (with a low bias) estimate with considerable resolution (the predicted distribution depends strongly on simulated precipitation) of local-scale wet day probabilities and precipitation intensities.

The calibration and model selection procedure for the VGLM mixture model is computationally rather expensive in the order of hours per gauge on one node of a state-of-the-art processor. The strength of this model therefore lies in an accurate representation of the full precipitation distribution for a relatively small number of gauges. For the description of a large number of rain gauges, one might consider using a VGLM gamma model (i.e., a simplified version of the mixture model where the tail distribution is omitted). Calibration of this model takes less than a second per gauge. For low to moderate simulated precipitation (predictor) values, the VGLM gamma model predicts basically the same quantiles as the mixture model: only very high quantiles are underestimated because of the light tail of the gamma distribution. This effect is stronger for summer, where it is already visible in the observed range of values. For winter, it is only relevant when extrapolating to unobserved intensities. For high simulated precipitation (predictor) values, however, the VGLM gamma model is not flexible enough and predicts too broad a distribution. Here, local precipitation will be overestimated for most quantiles and even for extremes (for very high return levels the light gamma tail will lead to an underestimation again). Comparisons based on the quantile score, however, indicate that, for the range of simulated and observed precipitation, predictions based on the VGLM gamma model are compatible with those from the mixture model. To summarize, the VGLM gamma is a fast and feasible alternative to the complex VGLM mixture model, as long as one is not concerned with very high extremes and a low number of potential outliers caused by the light gamma tail and the inflexibility of the VGLM gamma model, respectively.

Our stochastic MOS approach sets a new framework for bias correction to combine the advantages of precipitation generators and RCMs. Similarly, to change factor–based weather generators, it utilizes RCM output to produce a random sequence of local-scale weather. As a key advantage, however, our approach is not only consistent with the gridbox climate change signal of the RCM but also with the daily precipitation sequence produced by the RCM. Whereas change factor–based weather generators only produce internal climate variability up to several weeks based on Markov chains, our approach, by construction, captures all the climate variability simulated by the RCM, ranging from interannual to multidecadal fluctuations.

A further advantage results from the pairwise calibration. Traditional distribution-wise correction approaches implicitly assume that RCM-simulated precipitation is a realistic representation of observed precipitation. In the (pairwise) perfect boundary setting, we are able to assess this skill explicitly by evaluating the predictive power of our regression model.

The requirement of perfect boundary conditions restricts our approach to solely correct RCM biases; biases of the driving GCM will be preserved. Our MOS approach shares this property with PP approaches that are calibrated with observational predictors and then transferred to AOGCM predictors. Yet, to our knowledge, it is currently not clear to what extent a correction of biases in a free-running AOGCM or RCM–AOGCM modeling chain is justified. For instance, Eden et al. (2012) and Eden and Widmann (2014) argue that large-scale circulation errors in the AOGCM may not reasonably be corrected by postprocessing model output.

The framework of stochastic MOS opens a completely new research avenue in climate science, and our study should be regarded as a first step, rather than a fully developed tool. Time series can in principle be simulated from the current version of our model, but the day-to-day memory appeared to be slightly weaker than in reality. Including the previous day's predicted precipitation as predictor for wet day probabilities is expected to improve the simulation. In fact, the pairwise calibration provides a framework to include other predictors than simulated precipitation [i.e., to extend the simple bias correction/downscaling approach to a full multipredictor MOS; see Themeßl et al. (2011) for an initial study]. In a MOS context, these predictors can represent regional-scale processes and might, therefore, have much higher predictive power than typical large-scale predictors of PP approaches. Furthermore one might also transfer probabilistic multistation models (e.g., Yang et al. 2005) to our stochastic MOS that explicitly model spatial dependence. Similarly, the approach is, in principle, extendable to a multivariate approach. Such models would then establish a new framework for weather generators that are consistent with the weather sequence of a bias-corrected driving climate model, from daily to multidecadal scales.

Acknowledgments. We thank B. Geyer and colleagues from the Helmholtz Centre Geesthacht for kindly providing the CCLM simulations. We thank T. Thorarinsdottir for helpful discussions. This study has been undertaken within the PLEIADES project, funded by the Volkswagen Foundation (Grants 85423 and 85425). M. Vrac received additional funding by the McSIM and StaRMIP French ANR projects. T. Kent has received a Short-Term

Scientific Mission Grant from the EU COST Action ES1102 VALUE.

REFERENCES

Akaike, H., 1973: Information theory and an extension of the maximum likelihood principle. *Proc. Second Int. Symp. on Information Theory,* Budapest, Hungary, Institute of Electrical and Electronics Engineers, 267–281.

Bates, B., Z. W. Kundzewicz, S. Wu, and J. Palutikov, Eds., 2008: Climate change and water. IPCC Tech. Paper 6, 200 pp. [Available online at http://www.ipcc.ch/pdf/technical-papers/climate-change-water-en.pdf.]

Berrocal, V., A. Gelfand, and D. Holland, 2010: A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.,* **15,** 176–197, doi:10.1007/s13253-009-0004-z.

Chandler, R. E., and H. S. Wheater, 2002: Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland. *Water Resour. Res.,* **38,** 1192, doi:10.1029/2001WR000906.

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

——, F. Boberg, O. B. Christensen, and P. Lucas-Picher, 2008: On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophys. Res. Lett.,* **35,** L20709, doi:10.1029/2008GL035694.

Coles, S., 2001: *An Introduction to Statistical Modeling of Extreme Values.* Springer Series in Statistics, Vol. 14, Springer, 209 pp.

Davison, A. C., 2003: *Statistical Models.* Cambridge Series in Statistical and Probabilistic Mathematics, Vol. 11, Cambridge University Press, 738 pp.

Dobson, A. J., 2001: *An Introduction to Generalized Linear Models*. 2nd ed. Chapman and Hall, 240 pp.

Eden, J., and M. Widmann, 2014: Downscaling of GCM-simulated precipitation using model output statistics. *J. Climate,* **27,** 312–324, doi:10.1175/JCLI-D-13-00063.1.

——, ——, D. Grawe, and S. Rast, 2012: Skill, correction, and downscaling of GCM-simulated precipitation. *J. Climate,* **25,** 3970–3984, doi:10.1175/JCLI-D-11-00254.1.

Friederichs, P., 2010: Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes,* **13,** 109–132, doi:10.1007/s10687-010-0107-5.

——, and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.,* **135,** 2365–2378, doi:10.1175/MWR3403.1.

——, and T. L. Thorarinsdottir, 2012: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics,* **23,** 579–594, doi:10.1002/env.2176.

Frigessi, A., O. Haug, and H. Rue, 2002: A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes,* **5,** 219–235, doi:10.1023/A:1024072610684.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118, doi:10.1175/MWR2904.1.

Gregory, J. M., P. D. Jones, and T. M. L. Wigley, 1991: Precipitation in Britain: An analysis of area-average data updated to 1989. *Int. J. Climatol.,* **11,** 331–345, doi:10.1002/joc.3370110308.

Grotch, S. L., and M. C. MacCracken, 1991: The use of general circulation models to predict regional climate change. *J. Climate,* **4,** 286–303, doi:10.1175/1520-0442(1991)004<0286:TUOGCM>2.0.CO;2.

Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. *Wea. Forecasting,* **22,** 637–650, doi:10.1175/WAF989.1.

——, and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Wiley, 240 pp.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Katz, R., 1977: Precipitation as a chain-dependent process. *J. Appl. Meteor.,* **16,** 671–676, doi:10.1175/1520-0450(1977)016<0671:PAACDP>2.0.CO;2.

Kilsby, C. G., and Coauthors, 2007: A daily weather generator for use in climate change studies. *Environ. Modell. Softw.,* **22,** 1705–1719, doi:10.1016/j.envsoft.2007.02.005.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate,* **26,** 2137–2143, doi:10.1175/JCLI-D-12-00821.1.

——, T. J. Osborn, and N. P. Gillett, 2008: United Kingdom daily precipitation intensity: Improved early data, error estimates and an update from 2000 to 2006. *Int. J. Climatol.,* **28,** 833–842, doi:10.1002/joc.1672.

——, H. W. Rust, and T. J. Osborn, 2009: The annual cycle of heavy precipitation across the United Kingdom: A model based on extreme value statistics. *Int. J. Climatol.,* **29,** 1731–1744, doi:10.1002/joc.1811.

——, ——, and ——, 2010a: Synoptic airflow and UK daily precipitation extremes. *Extremes,* **13,** 133–153, doi:10.1007/s10687-010-0102-x.

——, and Coauthors, 2010b: 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.

——, T. J. Osborn, and H. W. Rust, 2011: The influence of synoptic airflow on UK daily precipitation extremes. Part I: Observed spatio-temporal relationships. *Climate Dyn.,* **36,** 261–275, doi:10.1007/s00382-009-0710-9.

Meehl, G. A., and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis.* S. Solomon et al., Eds., Cambridge University Press, 747–846.

Piani, C., J. O. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.,* **99,** 187–192, doi:10.1007/s00704-009-0134-9.

Rockel, B., A. Will, and A. Hense, 2008: The regional climate model COSMO-CLM (CCLM). *Meteor. Z.,* **17,** 347–348, doi:10.1127/0941-2948/2008/0309.

Rummukainen, M., 2010: State-of-the-art with regional climate models. *Wiley Int. Rev. Climate Change,* **1,** 82–96, doi:10.1002/wcc.8.

Seneviratne, S., and Coauthors, 2012: Changes in climate extremes and their impacts on the natural physical environment. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation,* C. B. Field et al., Eds., Cambridge University Press, 109–184.

Shao, J., 1997: An asymptotic theory for linear model selection. *Stat. Sin.,* **7,** 221–264.

Themeßl, M. J., A. Gobiet, and A. Leuprecht, 2011: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.,* **31,** 1530–1544, doi:10.1002/joc.2168.

Thorarinsdottir, T. L., and M. S. Johnson, 2012: Probabilistic wind gust forecasting using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.,* **140,** 889–897, doi:10.1175/MWR-D-11-00075.1.

van der Linden, P., and J. F. B. Mitchell, Eds., 2009: ENSEMBLES: Climate change and its impacts: Summary of research and results from the ENSEMBLES project. Met Office Hadley Centre Tech. Rep., 160 pp. [Available online at http://ensembles-eu.metoffice.com/docs/Ensembles_final_report_Nov09.pdf.]

von Storch, H., 1999: On the use of ''inflation'' in statistical downscaling. *J. Climate,* **12,** 3505–3506, doi:10.1175/1520-0442(1999)012<3505:OTUOII>2.0.CO;2.

——, H. Langenberg, and F. Feser, 2000: A spectral nudging technique for dynamical downscaling purposes. *Mon. Wea. Rev.,* **128,** 3664–3673, doi:10.1175/1520-0493(2000)128<3664:ASNTFD>2.0.CO;2.

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.

Widmann, M., C. S. Bretherton, and E. P. Salathé, 2003: Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as a predictor. *J. Climate,* **16,** 799–816, doi:10.1175/1520-0442(2003)016<0799:SPDOTN>2.0.CO;2.

Wigley, T. M. L., J. M. Lough, and P. D. Jones, 1984: Spatial patterns of precipitation in England and Wales and a revised, homogeneous England and Wales precipitation series. *J. Climatol.,* **4,** 1–25, doi:10.1002/joc.3370040102.

Wilby, R. L., and T. M. L. Wigley, 2000: Precipitation predictors for downscaling: Observed and general circulation model relationships. *Int. J. Climatol.,* **20,** 641–661, doi:10.1002/(SICI)1097-0088(200005)20:6<641::AID-JOC501>3.0.CO;2-1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 627 pp.

Xu, C., 1999: From GCMs to river flow: A review of downscaling methods and hydrologic modelling approaches. *Prog. Phys. Geogr.,* **23,** 229–249, doi:10.1177/030913339902300204.

Yang, C., R. E. Chandler, V. S. Isham, and H. S. Wheater, 2005: Spatial-temporal rainfall simulation using generalized linear models. *Water Resour. Res.,* **41,** W11415, doi:10.1029/2004WR003739.

Yee, T. W., and C. J. Wild, 1996: Vector generalized additive models. *J. Roy. Stat. Soc.* **58B,** 481–493.

——, and A. G. Stephenson, 2007: Vector generalized linear and additive extreme value models. *Extremes,* **10,** 1–19, doi:10.1007/s10687-007-0032-4.

Zwiers, F., and Coauthors, 2013: Climate extremes: Challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. *Climate Science for Serving Society: Research, Modeling and Prediction Priorities,* G. R. Asrar and J. W. Hurrell, Eds., Springer, 339–389, doi:10.1007/978-94-007-6692-1_13.