Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes

William Kleiber,[1] Richard W. Katz,[1] and Balaji Rajagopalan[2,3]

Received 1 July 2011; revised 14 October 2011; accepted 3 December 2011; published 19 January 2012.

[1] A daily stochastic spatiotemporal precipitation generator that yields spatially consistent gridded quantitative precipitation realizations is described. The methodology relies on a latent Gaussian process to drive precipitation occurrence and a probability integral transformed Gaussian process for intensity. At individual locations, the model reduces to a Markov chain for precipitation occurrence and a gamma distribution for precipitation intensity, allowing statistical parameters to be included in a generalized linear model framework. Statistical parameters are modeled as spatial Gaussian processes, which allows for interpolation to locations where there are no direct observations via kriging. One advantage of such a model for the statistical parameters is that stochastic generator parameters are immediately available at any location, with the ability to adapt to spatially varying precipitation characteristics. A second advantage is that parameter uncertainty, generally unavailable with deterministic interpolators, can be immediately quantified at all locations. The methodology is illustrated on two data sets, the first in Iowa and the second over the Pampas region of Argentina. In both examples, the method is able to capture the local and domain aggregated precipitation behavior fairly well at a wide range of time scales, including daily, monthly, and annually.

Citation: Kleiber, W., R. W. Katz, and B. Rajagopalan (2012), Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes, *Water Resour. Res.*, *48*, W01523, doi:10.1029/2011WR011105.

1. Introduction

[2] Stochastic precipitation generators have been key features of downscaling, climate impact studies, hydrologic models and agricultural models for a number of decades. At individual locations, these random generators often rely on a Markov chain to describe the temporal dependence of precipitation occurrence [*Katz*, 1977] and either an exponential distribution, gamma distribution or some mixture thereof to model the intensity of rainfall given its occurrence [*Richardson*, 1981; *Stern and Coe*, 1984; *Woolhiser and Pegram*, 1979]. Recent reviews of statistical downscaling, including use of stochastic precipitation generators, have been covered by *Maraun et al.* [2010] and *Wilks* [2010]. *Wilks and Wilby* [1999] give a thorough introduction and history of stochastic weather generators (i.e., for which precipitation is one component).

[3] Recent interest in this field has moved away from individual location models to spatial models that are able to generate spatially and temporally correlated fields of precipitation. This is especially difficult, given the highly variable nature of precipitation over small spatial and temporal scales, including its intermittency. It is critical to capture the domain aggregate behavior of precipitation intensity and dry or wet spells, which play important roles in hydrologic planning and water resource management. There are a number of approaches to spatial-temporal modeling of precipitation, including hidden Markov models for occurrence [*Hughes and Guttorp*, 1999] and intensity [*Ailliot et al.*, 2009; *Charles et al.*, 1999], resampling based on nearest neighbors [*Apipattanavis et al.*, 2007; *Buishand and Brandsma*, 2001; *Rajagopalan and Lall*, 1999], generalized chain-dependent processes [*Zheng and Katz*, 2008; *Zheng et al.*, 2010], power transformation of precipitation to normality [*Sansó and Guenni*, 2000; *Yang et al.*, 2005], artificial neural networks [*Cannon*, 2008], or copula-based approaches [*Bárdossy and Pegram*, 2009]. One of the main advantages of stochastic precipitation generators is uncertainty quantification over a spatial domain, which is used, for example, in assessment of impacts of climate change [*Kilsby et al.*, 2007; *Mehrotra and Sharma*, 2010]; see also *Burton et al.* [2008] for recent spatiotemporal precipitation simulation software. While herein we focus on daily simulation, there is a large body of literature on simulation at finer temporal scales [see, e.g., *Onof et al.*, 2000; *Rodríguez-Iturbe et al.*, 1988; *Valdes and Rodríguez-Iturbe*, 1985].

[4] The early work of *Wilks* [1998] paved the way for many current approaches to this problem, relying on correlated latent multivariate normals and transformed multivariate normals to generate correlated occurrence and intensity fields. Recent advances have been made by *Brissette et al.* [2007] and *Thompson et al.* [2007] who discuss efficient simulation and more sophisticated estimation approaches,

[1]Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, Colorado, USA.
[2]Department of Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, Colorado, USA.
[3]Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, Colorado, USA.

Copyright 2012 by the American Geophysical Union
0043-1397/12/2011WR011105

respectively. *Mehrotra et al.* [2006] compared the resampling, hidden Markov model, and *Wilks* [1998] approaches and found that all are flexible, but the approach of *Wilks* [1998] has the advantage of replicating the observed spatial dependence while still allowing for significant flexibility in temporal dependence at individual locations.

[5] Most spatial precipitation generators only simulate precipitation at locations with observations. In order to generate simulations between observation sites, model parameters are required over the entire simulation domain. Assuming spatially constant model parameters is typically unjustified because of variable orography, so various authors have proposed estimating these parameters at observation locations, and interpolating this information to any location usually with a type of weighted regression [*Johnson et al.*, 2000; *Wilks*, 2008]. While these methods yield parameters throughout the domain, they do not produce estimates of parameter uncertainty at ungauged locations. Developing a gridded precipitation generator with locally varying parameters and uncertainty characterization is crucial for applications such as crop modeling and water resource modeling, especially since data are usually not available at all locations of interest.

[6] We propose a parametric model for spatially correlated precipitation occurrence and intensity that is similar to the approach of *Wilks* [1998]. Seasonal variation and other covariates can be included in a generalized linear model (GLM) framework, which imparts significant flexibility in the statistical model [*Yang et al.*, 2005; *Furrer and Katz*, 2007]. Along with seasonally varying occurrence rate and intensities, we are also able to capture seasonally varying spatial correlation within our parametric framework. The basic probabilistic structure we use involves Gaussian processes, which are simply stochastic processes whose finite dimensional distributions are multivariate normal. The attraction of Gaussian processes is that they can be completely described by very few parameters, and have displayed a wide array of flexibility in numerous scientific contexts. In particular, we use a two-stage model, where precipitation occurrence is driven by a latent spatial process and precipitation intensity is modeled as a transformed Gaussian process. At individual stations, our model reduces to a Markov chain for occurrence and a gamma distribution for intensity.

[7] Parameters, those coefficients in the GLM framework, are modeled as Gaussian processes. Applying our statistical model with the same GLM coefficients at all locations is inappropriate, especially in regions of complex terrain [*Johnson et al.*, 2000]. Using a stochastic model for the statistical parameters allows for a locally varying precipitation model that is available at any point of interest, and in particular gridded simulations are readily produced. The Gaussian process representation lends itself to the flexible method of spatial interpolation known as kriging. An equally important benefit of using a stochastic model for the parameters is that parameter uncertainty is immediately available at all locations, which has been emphasized by *Lima and Lall* [2009]. Incorporating parameter uncertainty can also reduce the common problem of overdispersion in precipitation generators [*Katz and Zheng*, 1999].

[8] The use of Gaussian processes for precipitation modeling is becoming more widespread, as they can capture a wide variety of spatial correlation, while requiring only a few parameters. *Berrocal et al.* [2008] used latent Gaussian processes in the context of short-term mesoscale precipitation forecasting. *Wilks* [2009] used these stochastic functions to generate gridded precipitation data sets. Our approach departs from that of *Wilks* [2009] in that we put the occurrence and intensity models in a GLM framework, our approach to seasonally varying spatial correlation is more parsimonious and requires only a few parameters, and we produce uncertainty estimates of model parameters at any location, even without direct observations.

[9] We illustrate our model on two data sets, the first a network of 22 stations in Iowa that are taken from the U.S. Historical Climatology Network [*Menne et al.*, 2010], and the second a more sparsely populated network of 19 stations in the Pampas region of Argentina. Neither of these regions have complex terrain, but both do exhibit a substantial degree of seasonality in precipitation patterns and have spatially varying average occurrence and intensity components. In both cases we illustrate the flexibility of our model to capture local and domain aggregated precipitation behavior at various time scales.

2. Stochastic Model

[10] Our approach to generating spatially and temporally correlated precipitation is to first generate precipitation occurrence and then, at locations with positive precipitation, simulate precipitation intensities. We begin describing our approach with the occurrence model.

2.1. Precipitation Occurrence

[11] At site $\boldsymbol{s}$ on day t, denote precipitation occurrence $O(\boldsymbol{s}, t) = 0$ if there is no rainfall, and $O(\boldsymbol{s}, t) = 1$ otherwise. We introduce a latent (i.e., unobserved) Gaussian process $W_1(\boldsymbol{s}, t)$ with mean $\mu(\boldsymbol{s}, t)$ and covariance function $C_1(\boldsymbol{h}, t)$ where $\boldsymbol{h} = \boldsymbol{s}_1 - \boldsymbol{s}_2$ is the spatial lag vector between two locations. Then precipitation occurrence relies on the latent spatial field as

$$O(\boldsymbol{s}, t) = 0 \quad \text{if} \quad W_1(\boldsymbol{s}, t) < 0 \tag{1}$$

$$O(\boldsymbol{s}, t) = 1 \quad \text{if} \quad W_1(\boldsymbol{s}, t) \geq 0 \tag{2}$$

The physical motivation for conditioning on the latent process is that locations in small neighborhoods will tend to have correlated precipitation occurrence, which is preserved in the spatial correlation function $C_1(\boldsymbol{h}, t)$.

[12] The mean function $\mu(\boldsymbol{s}, t)$ will typically be a regression on small-order harmonics and the previous day's occurrence. If available, one can include covariates such as climate model output or broad scale atmospheric conditions. For our purposes, write

$$\mu(\boldsymbol{s}, t) = \beta_\mu(\boldsymbol{s})' X_\mu(\boldsymbol{s}, t) \tag{3}$$

where $X_\mu(\boldsymbol{s}, t)$ are the covariates, and $\beta_\mu(\boldsymbol{s}) = (\beta_{\mu,1}(\boldsymbol{s}), \ldots, \beta_{\mu,p}(\boldsymbol{s}))'$ are spatially varying regression parameters. Although we consider only linear functions, this basic model accommodates significant flexibility which may be desired on the basis of knowledge about a specific simulation domain.

[13] The covariance function $C_1(\boldsymbol{h}, t)$ is assumed to depend only on spatial lag $\boldsymbol{h}$ and time point t; below we use an exponential covariance with temporally varying scale,

$$C_1(\boldsymbol{h}, t) = \exp(-\|\boldsymbol{h}\|/A(t)) \tag{4}$$

where $A(t)$ is the time-dependent scale parameter. Notice this correlation function is isotropic in that it only depends on the length of $\boldsymbol{h}$; hence we implicitly are assuming no directional dependence of spatial correlation, but such anisotropy can be included by adjusting C_1. It is convenient to view $A(t)$ as a regression on two harmonics, allowing for spatial occurrence correlation that is either weaker or stronger depending on the season, as is often observed in practice.

[14] Estimation of the $\beta_\mu(\boldsymbol{s})$ parameters for individual sites proceeds by maximum likelihood. In particular, at individual locations our conditional model reduces to a probit regression [*McCullagh and Nelder*, 1989], with the maximum likelihood estimates (MLEs) being obtained at each location independently. This yields, for each observation location $\boldsymbol{s}$, estimates denoted by $\widehat{\beta}_\mu(\boldsymbol{s})$.

[15] Direct estimates of $\beta_\mu(\boldsymbol{s})$ are available only at locations with observed data, but in order to generate a complete precipitation field across the entire simulation domain, we require a method that yields estimates of these regression parameters between observation stations. A second consideration is the underlying uncertainty in these parameter estimates, which we seek to communicate along with the weather generator. Hence, as a second stage of modeling, we view the MLEs $\widehat{\beta}_{\mu,i}(\boldsymbol{s})$ as realizations from spatial Gaussian processes, where $\beta_{\mu,i}(\boldsymbol{s})$ is the ith component of $\beta_\mu(\boldsymbol{s})$. In particular, we decompose

$$\beta_{\mu,i}(\boldsymbol{s}) = Z_{\mu,i}(\boldsymbol{s}) + \epsilon_{\mu,i}(\boldsymbol{s}) \tag{5}$$

where $Z_{\mu,i}(\boldsymbol{s})$ is a spatial Gaussian process with mean $\theta_{\mu,i}$ and Matérn covariance function [*Stein*, 1999], which has parameters variance $\sigma^2_{\mu,i}$, scale $a_{\mu,i}$ and smoothness $\nu_{\mu,i}$. Here $\epsilon_{\mu,i}(\boldsymbol{s})$ is a normally distributed error term, centered at zero, with variance $\tau^2_{\mu,i}$. In this context, $\epsilon_{\mu,i}(\boldsymbol{s})$ can be thought of as small scale variability that is indistinguishable from measurement error, and this quantifies our uncertainty in parameter estimates at observation locations. We use the Matérn covariance function, as it allows the statistical model to be smooth across space, with the parameter ν controlling smoothness. This is as opposed to precipitation realizations that are more intermittent, using the exponential covariance function (4) which coincides with the Matérn class when $\nu = 0.5$, yielding rougher stochastic realizations. The spatial process $Z_{\mu,i}(\boldsymbol{s})$ will allow us to smooth the individual estimates $\widehat{\beta}_{\mu,i}(\boldsymbol{s})$ over the simulation domain, allowing for stochastic realizations at any location. We estimate parameters $\theta_{\mu,i}, \sigma^2_{\mu,i}, a_{\mu,i}, \nu_{\mu,i}$ and $\tau^2_{\mu,i}$ by maximum likelihood, conditional on MLEs $\widehat{\beta}_{\mu,i}(\boldsymbol{s})$. We use this two-step procedure to estimate parameters, as there are so many parameters that direct maximization in one step would be extremely difficult and time consuming. Using a two-step procedure such as ours has proven effective for similar types of models [*Berrocal et al.*, 2008; *Kleiber et al.*, 2011, 2012].

[16] At this point, the individual location model has been described, and the crucial step for the spatially correlated occurrence model is in estimating the temporally varying scale parameter $A(t)$ of (4). It is possible to estimate $A(t)$ by maximum likelihood using a stochastic EM algorithm [*Nielsen*, 2000], but it is well known that MLEs will underestimate spatial dependence if an overly simple mean function is used [*Kitanidis and Lane*, 1985]. On the basis of our experiments, the MLEs result in spatial correlation that is substantially too weak. As an alternative, we propose a method of moments approach that does not make any likelihood assumptions, instead minimizing the squared distance between model correlations and observed spatial correlations. Specifically, our estimates are obtained through the following minimization,

$$\begin{aligned}\min_{A(\cdot)} \sum_t \sum_{i \neq j} \sum_{k,\ell=0,1} &(\widehat{P}(O(\boldsymbol{s}_i, t) = k, O(\boldsymbol{s}_j, t) = \ell) \\ &- P_M(O(\boldsymbol{s}_i, t) = k, O(\boldsymbol{s}_j, t) = \ell))^2\end{aligned} \tag{6}$$

where $\widehat{P}(O(\boldsymbol{s}_i, t) = k, O(\boldsymbol{s}_j, t) = \ell)$ is the observed relative frequency of site $\boldsymbol{s}_i$ taking on value k and site $\boldsymbol{s}_j$ taking on the value ℓ simultaneously on day t, and $P_M(O(\boldsymbol{s}_i, t) = k, O(\boldsymbol{s}_j, t) = \ell)$ is the model probability of the same event. Note that the model probability is not a simple function of pairwise correlation, for example when $k = \ell = 0$, the probability is $P(W_1(\boldsymbol{s}_i, t) < 0, W_1(\boldsymbol{s}_j, t) < 0) \neq \mathrm{Cor}(W_1(\boldsymbol{s}_i, t), W_1(\boldsymbol{s}_j, t))$, and actually involves a two-dimensional integral of a bivariate normal probability density function, which is readily approximated in most scientific computing languages, for instance using the mvtnorm package in R. Other authors have considered similar estimation approaches to spatial correlations [*Allcroft and Glasbey*, 2003; *Baigorria and Jones*, 2010; *Durban and Glasbey*, 2001]. Such an optimization in terms of joint probabilities of occurrence is effectively equivalent to optimization in terms of pairwise correlations for the occurrence process. Recall that $A(t)$ typically has only a few free parameters, and the minimization is taken over these parameters.

2.2. Precipitation Intensity

[17] At site $\boldsymbol{s}$ on day t, consider precipitation amount denoted by $Y(\boldsymbol{s}, t)$ (i.e., the intensity conditional on $O(\boldsymbol{s}, t) = 1$). We propose a general intensity model, and discuss a special case that involves fewer parameters. In general, at a single site we model $Y(\boldsymbol{s}, t)$ as a gamma distributed random variable.

[18] The general approach uses spatially and seasonally varying shape and scale parameters of the gamma distribution. With a seasonally varying shape parameter, we can indirectly take account of the frequency of different types of precipitation, such as convective and frontal. We have scale $\alpha(\boldsymbol{s}, t)$ and shape $\gamma(\boldsymbol{s}, t)$, both of which typically are regressions on covariates, with possibly different covariates and parameter values than those used in the occurrence model. In particular, we have

$$\log \alpha(\boldsymbol{s}, t) = \beta_\alpha(\boldsymbol{s})' X_\alpha(\boldsymbol{s}, t) \tag{7}$$

$$\log \gamma(\boldsymbol{s}, t) = \beta_\gamma(\boldsymbol{s})' X_\gamma(\boldsymbol{s}, t) \tag{8}$$

where the log transformation guarantees the parameters take on positive values. The scale and shape parameters are estimated by maximum likelihood at each individual observation location, which in turn yields local simulations tuned to observational records.

[19] In order to produce locally realistic simulations between observation stations, we require a spatial model for the coefficients $\alpha(\boldsymbol{s},t)$ and $\gamma(\boldsymbol{s},t)$. To this end, we view the MLEs $\widehat{\beta}_{\alpha,i}(\boldsymbol{s})$ and $\widehat{\beta}_{\gamma,i}(\boldsymbol{s})$ as realizations from spatial Gaussian processes. In particular, all coefficient processes are decomposed as in (5), where $Z_{\alpha,i}(\boldsymbol{s})$ and $Z_{\gamma,i}(\boldsymbol{s})$ are spatial Gaussian processes with constant means and Matérn covariance functions. As in the occurrence case, $\epsilon_{\alpha,i}(\boldsymbol{s})$ and $\epsilon_{\gamma,i}(\boldsymbol{s})$ are normally distributed error terms, centered at zero, with variances $\tau^2_{\alpha,i}$ and $\tau^2_{\gamma,i}$, respectively. The spatial parameters that govern behavior of the $Z(\boldsymbol{s})$ and $\epsilon(\boldsymbol{s})$ processes are estimated by maximum likelihood, conditional on MLEs $\widehat{\beta}_{\alpha,i}(\boldsymbol{s})$ and $\widehat{\beta}_{\gamma,i}(\boldsymbol{s})$.

[20] A special case of the general model was discussed by *Furrer and Katz* [2007], who fixed the gamma shape parameter across time at a given site, which has been found to be adequate for some data sets such as precipitation intensity at one location in the Pampas region, see also *Yang et al.* [2005]. In this special case, the log of the mean of the gamma distribution is regressed on a set of covariates, perhaps the same as those specified in (7). Specifically, with scale $\alpha(\boldsymbol{s},t)$ and shape $\gamma(\boldsymbol{s})$, we have

$$\log\big(\alpha(\boldsymbol{s},t)\gamma(\boldsymbol{s})\big) = \beta_{\alpha\gamma}(\boldsymbol{s})' X_{\alpha\gamma}(\boldsymbol{s},t) \tag{9}$$

where the mean of the gamma distribution is $\alpha(\boldsymbol{s},t)\gamma(\boldsymbol{s})$. As the shape $\gamma(\boldsymbol{s})$ does not depend on time t, we are effectively modeling only the scale $\alpha(\boldsymbol{s},t)$ as a function of the temporal covariates. The mean and shape parameters are estimated by maximum likelihood at each observation location, which implies a seasonally varying scale parameter $\alpha(\boldsymbol{s},t)$ at a given site. The parameters $\alpha(\boldsymbol{s},t)$ and $\gamma(\boldsymbol{s})$ are modeled as transformed spatial Gaussian processes, similar to the general model.

[21] To simulate spatially correlated fields of precipitation, we introduce a Gaussian process $W_2(\boldsymbol{s},t)$ with mean zero and covariance function $C_2(\boldsymbol{h},t)$, such that

$$Y(\boldsymbol{s},t) = G^{-1}_{\boldsymbol{s},t}(\Phi(W_2(\boldsymbol{s},t))) \tag{10}$$

where $G_{\boldsymbol{s},t}$ is the cumulative distribution function (CDF) of the gamma distribution at site $\boldsymbol{s}$ and time t, and Φ is the CDF of a standard normal. This transformation approach is called a spatially varying anamorphosis function [*Chilès and Delfiner*, 1999], retaining the gamma distribution at individual locations while allowing for any degree of correlation between locations; *Berrocal et al.* [2008] used a similar model for mesoscale precipitation forecasting, but without spatially varying coefficient processes.

[22] The covariance function $C_2(\boldsymbol{h},t)$ of $W_2(\boldsymbol{s},t)$ is a time varying stationary spatial covariance function with a form identical to that of (4), but with distinct parameters. Producing stochastic realizations of spatial precipitation intensities relies on an appropriate estimate of $A(t)$. Along the same lines as with the occurrence model, we avoid using maximum likelihood to estimate $A(t)$, and instead obtain an estimate through the following method of moments minimization:

$$\min_{A(\cdot)} \sum_t \sum_{i\neq j} \big(\widehat{\mathrm{Cor}}(Y(\boldsymbol{s}_i,t), Y(\boldsymbol{s}_j,t)) - \mathrm{Cor}_M(Y(\boldsymbol{s}_i,t), Y(\boldsymbol{s}_j,t))\big)^2 \tag{11}$$

where $\widehat{\mathrm{Cor}}(Y(\boldsymbol{s}_i,t), Y(\boldsymbol{s}_j,t))$ is the empirical correlation between site $\boldsymbol{s}_i$ and site $\boldsymbol{s}_j$ on day t, while Cor_M is the corresponding model correlation. Note that only terms for which precipitation occurs at both sites simultaneously enter into the above equation. Because of the highly nonlinear transformation involved, the model correlation is not available in closed form. But it is straightforward to simulate from the bivariate distributions and, in practice, we approximate the model correlation by Monte Carlo approximations. In particular, for any fixed trial value of $A(t)$, we approximate $\mathrm{Cor}_M(Y(\boldsymbol{s}_i,t), Y(\boldsymbol{s}_j,t))$ by Monte Carlo sampling for all pairs $i \neq j$ and calculate the distance (11). This procedure is performed over a grid of potential values of $A(t)$, with the value minimizing the distance (11) being chosen.

[23] We note that *Wilks* [1998] identifies an edge effect, where generating a field of intensities independent of the occurrences can lead to unrealistically large values of precipitation near the boundary of a dry area. If this is a concern, one option is to generate the spatial fields $W_1(\boldsymbol{s},t)$ and $W_2(\boldsymbol{s},t)$ from the same random number seed. Mathematically, this is equivalent to generating a single set of uniform random numbers and using this same set in the inverse probability integral transformations for both the occurrence and intensity process. This has the effect of imposing positive correlation between Gaussian processes for both field realizations, while allowing for the two fields to have differing length scales. Using the same random number seed will decrease the variability of field realizations, as occurrence and intensity will no longer be independent.

[24] We end section 2.2 with an outline of operational use of our model. We assume model parameters have already been estimated as described above. For simulation at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ on day t, perform the following steps.

[25] 1. Simulate the errors $\varepsilon_{\cdot,i}(\boldsymbol{s}_j)$ and put $\beta_{\cdot,i}(\boldsymbol{s}_j) = \widehat{\beta}_{\cdot,i}(\boldsymbol{s}_j) + \varepsilon_{\cdot,i}(\boldsymbol{s}_j)$, where the dot denotes μ, α and γ, and $j = 1, \ldots, n$.

[26] 2. Randomly generate a realization from each multivariate normal vector $(W_1(\boldsymbol{s}_1,t), \ldots, W_1(\boldsymbol{s}_n,t))'$, and $(W_2(\boldsymbol{s}_1,t), \ldots, W_2(\boldsymbol{s}_n,t))'$, where the mean of $W_1(\boldsymbol{s},t)$ uses the simulated values obtained in Step 1. The covariance matrices used for generation of the multivariate normal vectors are obtained from (4), using the estimates of $A(t)$ for W_1 and W_2, respectively.

[27] 3. For each location $\boldsymbol{s} = \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ there is zero precipitation if $W_1(\boldsymbol{s},t) < 0$.

[28] 4. For those locations with positive precipitation $(W_1(\boldsymbol{s},t) \geq 0)$, the simulated intensity is set to $Y(\boldsymbol{s},t) = G^{-1}_{\boldsymbol{s},t}(\Phi(W_2(\boldsymbol{s},t)))$, where the shape and scale of the gamma CDF $G_{\boldsymbol{s},t}$ rely on the simulated values of $\beta_{\alpha,i}(\boldsymbol{s})$ and $\beta_{\gamma,i}(\boldsymbol{s})$ obtained in Step 1.

3. Applications

[29] We test our model on two data sets, the first in Iowa consists of 22 observation stations that make up a subsection

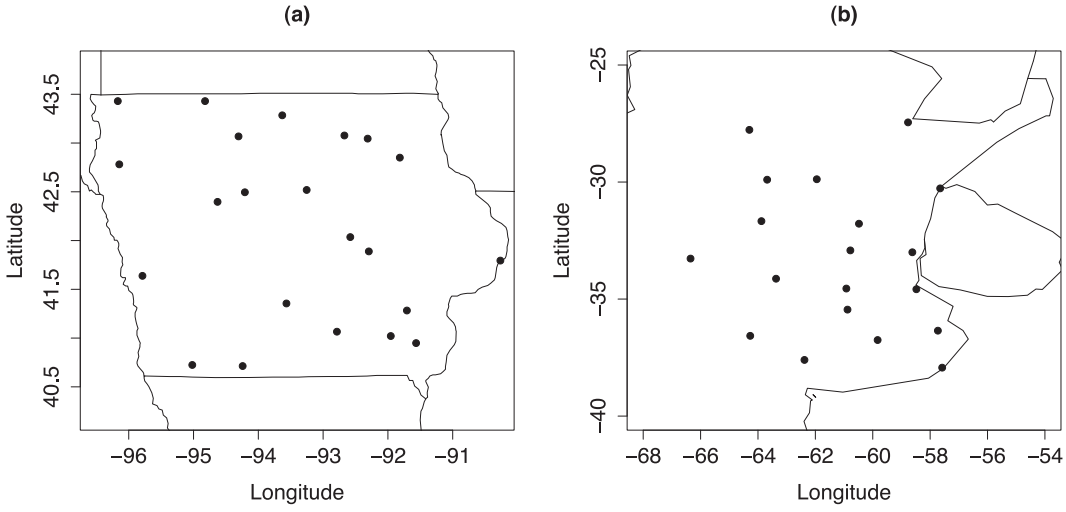

Figure 1. Observation networks in (a) Iowa and (b) Pampas region of Argentina.

of the U.S. Historical Climatology Network (USHCN) [*Menne et al.*, 2010]. Station locations are displayed in Figure 1. Station data are available from 1893 to 2009, with the shortest record still being 111 years in length. One station has precipitation data available on only 61% of days, while all other locations have observations on at least 90% of days. The Iowa observation network is fairly dense, with the minimal intersite distance at 29 km and maximal distance at 516 km, covering a range of approximately 145,000 km^2. The second data set is from the Pampas region of Argentina, consisting of 19 stations with data being available from 1908 to 2010 but with most stations beginning recording during the 1930s. The observation network in the Pampas is significantly sparser than that in Iowa, with the minimal intersite distance at 100 km and maximal distance at 1293 km, covering a range of approximately 750,000 km^2.

[30] Precipitation simulation in these regions is challenging because of the marked seasonality of precipitation and spatially varying average intensity and occurrence. For convenience, we have removed leap days from both sets of data, so that all years have 365 days. For sake of space, we include only some illustrative plots for the Pampas data, but the proposed stochastic model fits equally well (some corresponding plots for the Pampas are included in the auxiliary materials).[1] For both domains, the observation networks were sparse enough that the edge effect discussed by *Wilks* [1998] was not apparent, and independently simulating occurrence and intensity processes is a reasonable approach. Ensuring the continuity between dry and wet areas would also be more important on smaller time scales than for daily aggregated precipitation.

3.1. Iowa Precipitation

[31] The latent process in the occurrence model of section 2.1 requires a mean and covariance function. The covariance function takes on the form specified in (4), with a scale function specified by

$$A(t) = \exp\big(a_0 + a_1\cos(2\pi t/365) + a_2\sin(2\pi t/365)\big) \quad (12)$$

The motivation for this functional form is that, in Iowa, spatial correlation of precipitation occurrence is dependent on season. During winter, most precipitation events occur because of the passage of weather fronts, whereas summer precipitation is often driven by highly localized convective storms. Note that the parameters a_0, a_1 and a_2 are only estimated once, which yield temporally varying spatial correlation for the entire year. The estimated parameters imply, at the average intersite distance (206 km), a maximal correlation of 0.73 and a minimal correlation of 0.54 on the latent Gaussian process scale, corresponding to winter and summer respectively. We specify the mean function as

$$\begin{aligned}\mu(\boldsymbol{s},t) = {} & \beta_{\mu,0}(\boldsymbol{s}) + \beta_{\mu,1}(\boldsymbol{s})\,O(\boldsymbol{s},t-1) \\ & + \beta_{\mu,2}(\boldsymbol{s})\cos(2\pi t/365) + \beta_{\mu,3}(\boldsymbol{s})\sin(2\pi t/365) \\ & + \beta_{\mu,4}(\boldsymbol{s})\cos(4\pi t/365) + \beta_{\mu,5}(\boldsymbol{s})\sin(4\pi t/365)\end{aligned} \quad (13)$$

which defines a first-order, two-state Markov chain at individual locations. Using the Bayesian information criterion (BIC) [*Schwarz*, 1978], the covariates included in $\mu(\boldsymbol{s},t)$ were selected over other options including fewer harmonics or interactions. In particular, at each location, we computed the BIC values for a set of possible covariates, and exactly the same set of covariates in $\mu(\boldsymbol{s},t)$ was favored at all locations. The same approach was used below to choose all remaining mean function covariates for precipitation intensity in Iowa, and in the Pampas.

[32] We use the general precipitation intensity model proposed in section 2.2 (i.e., as specified by (7) and (8)) because of the distinct types of precipitation that occur in Iowa throughout the year. The scale and shape parameters of the gamma intensity model are

$$\begin{aligned}\log\alpha(\boldsymbol{s},t) = {} & \beta_{\alpha,0}(\boldsymbol{s}) + \beta_{\alpha,1}(\boldsymbol{s})\,O(\boldsymbol{s},t-1) \\ & + \beta_{\alpha,2}(\boldsymbol{s})\cos(2\pi t/365) \\ & + \beta_{\alpha,3}(\boldsymbol{s})\sin(2\pi t/365)\end{aligned} \quad (14)$$

[1] Auxiliary materials are available in the HTML. doi: 10.1029/2011WR011105.

$$\log\gamma(\boldsymbol{s},t) = \beta_{\gamma,0}(\boldsymbol{s}) + \beta_{\gamma,1}(\boldsymbol{s})\,O(\boldsymbol{s},t-1) + \beta_{\gamma,2}(\boldsymbol{s})\cos(2\pi t/365) + \beta_{\gamma,3}(\boldsymbol{s})\sin(2\pi t/365). \quad (15)$$

Through these regressions, we allow for seasonally varying precipitation intensity, with heavier precipitation during summer. We also allow intensity to depend on the previous day's occurrence, allowing for daily regime shifts. Not only is the spatial correlation of precipitation occurrence seasonally varying, but the spatial correlation of intensities also varies by season. The localized convective storms that characterize summer precipitation exhibit lower spatial correlation of precipitation intensities than the more widespread frontal events during winter. Hence, we decompose the temporally varying scale $A(t)$ in the same way as (12), but with distinct parameters. At the average intersite distance, the maximal correlation during winter is approximately 0.44, which drops to 0.15 during summer, on the Gaussian process scale.

3.2. Pampas Precipitation

[33] Data from a single station in the Pampas, Argentina network of stations was analyzed by *Furrer and Katz* [2007], and we follow essentially the same model adopted by them. However, we do not include ENSO as a covariate to simplify comparisons with the results from Iowa. The latent process in the occurrence model of section 2.1 requires a mean and covariance function. The covariance function takes on the form specified in (4), with a scale function the same as (12). The average intersite distance in the Pampas data set is 525 km where the maximal and minimal correlation at this distance are 0.48 and 0.39 on the latent Gaussian process scale, respectively. Following *Furrer and Katz* [2007], we specify the mean function as

$$\begin{aligned}\mu(\boldsymbol{s},t) = {} & \beta_{\mu,0}(\boldsymbol{s}) + \beta_{\mu,1}(\boldsymbol{s})O(\boldsymbol{s},t-1) + \beta_{\mu,2}(\boldsymbol{s})\cos(2\pi t/365) \\ & + \beta_{\mu,3}(\boldsymbol{s})\sin(2\pi t/365) \\ & + \beta_{\mu,4}(\boldsymbol{s})\cos(2\pi t/365)O(\boldsymbol{s},t-1) \\ & + \beta_{\mu,5}(\boldsymbol{s})\sin(2\pi t/365)O(\boldsymbol{s},t-1)\end{aligned} \quad (16)$$

which defines a Markov chain at single locations; notice we use probit regression instead of logistic regression as done by *Furrer and Katz* [2007].

[34] The log of the mean of the gamma intensity model in section 2.2 (i.e., as specified by (9)) is the regression

$$\beta_0(\boldsymbol{s}) + \beta_1(\boldsymbol{s})\cos(2\pi t/365) + \beta_2(\boldsymbol{s})\sin(2\pi t/365) \quad (17)$$

with a fixed shape parameter. Through these regressions, we allow for seasonally varying precipitation intensity, with heavier precipitation during summer. Not only is the spatial correlation of precipitation occurrence seasonally varying, but the spatial correlation of intensities also varies by season. The seasonality of precipitation in the Pampas is markedly stronger than in Iowa. We also decompose the temporally varying scale $A(t)$ of the transformed intensity process in the same way as (12), but with distinct parameters. Here the implied correlation on the Gaussian process scale is 0.06 during winter and 0.20 during summer, at the average intersite distance. Notice the harmonics in our model for $A(t)$ allows the correlation to be high during summer in Iowa (JJA) which is winter in the Pampas, where we have the lowest correlation.

3.3. Model Validation in Iowa and the Pampas

[35] We simulated trajectories of all available years of data using our stochastic spatiotemporal precipitation model from both data sets, and now examine the results. In particular, we generated 100 trajectories of the 117 years of data for Iowa, and 111 years of data for the Pampas. All figures and discussion refer to the Iowa data, unless otherwise noted. We begin by validating performance of the spatial occurrence model described in section 2.1. For one example location in Iowa (i.e., Washington), Figure 2 shows empirical 1 day transition probabilities for rainfall occurrence with the model transition probabilities superimposed. As is well known, the previous day's occurrence is an important covariate to include; in this case, the probability of precipitation is approximately 0.15 higher on days preceded by precipitation. The higher-order harmonics chosen by BIC allow for a substantial increase in probability during springtime, which levels out during the other seasons.

[36] An essential feature that stochastic precipitation generators attempt to replicate is the length of dry and wet spells. Figure 3 shows the log frequency of empirical dry spells from 1 day to up to 70 days at the example locations, Rock Rapids, Iowa, and Buenos Aires, Argentina, with 90% pointwise confidence intervals based on the 100 simulations overlaid. Our single station Markov chain model shows extremely good characterization of dry spell behavior at individual locations. Even for extreme spell lengths, the precipitation generator reasonably replicates the average length and occurrence of these dry spells.

[37] Our precipitation generator satisfactorily simulates occurrence behavior at individual locations, but the power in the method is in correlating occurrences across space. Figure 4 shows simulated and observed probabilities for pairs of stations in Iowa being simultaneously dry or simultaneously wet in winter (DJF) and in summer (JJA). Our latent Gaussian process model accurately replicates the observed pairwise probabilities well. In particular, the model is able to reproduce pairwise probabilities of anywhere from approximately 0.05 to up to 0.80, displaying very flexible behavior with a relatively simple parametric model. We also examined pairwise probabilities of locations being oppositely wet and dry, which showed similar extent of replication as in the plots of Figure 4.

[38] To assess the model climatology of occurrences, we examine the seasonally varying mean and variability of occurrence rate. The unconditional occurrence rate is effectively equivalent to the average number of wet days in any given month. Table 1 shows the mean and standard deviation of occurrences across the domain of Iowa by each month. Our simulations replicate the observed average to within $0.01 - 0.02$, and capture the variability extremely well, typically differing by less than 0.015 from the observed variability.

[39] One of the most difficult features of multisite precipitation data to replicate is the length of spatially aggregated wet and dry spells, another feature essential for

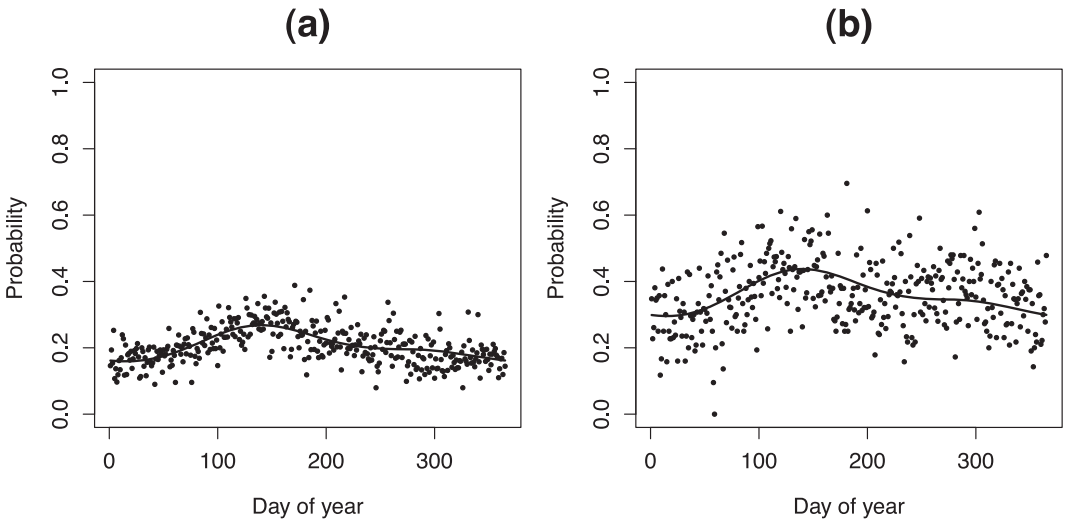

Figure 2. Empirical transition probabilities at Washington, Iowa, with model transition probabilities as the line for (a) $O(\mathbf{s}, t-1) = 0$ and $O(\mathbf{s}, t) = 1$ and (b) $O(\mathbf{s}, t-1) = 1$ and $O(\mathbf{s}, t) = 1$.

hydrological planning and water resource management. We define an aggregate dry spell as the maximum number of consecutive days without rain at any location within the domain, and an aggregate wet spell as the maximum number of consecutive days on which precipitation occurred at at least one location within the domain. Figure 5 shows aggregate dry and wet spells with 90% pointwise model confidence intervals based on our 100 simulations in Iowa and the Pampas. In Iowa, our approach was unable to generate sufficiently long domain dry spells; however, it captures aggregate wet spell behavior very accurately, except at short lengths. In the Pampas region, our latent Gaussian process model replicates both the dry and wet spells extremely accurately at long lengths. We reiterate that our model is not tuned to reproduce this statistic; our positive performance is due to capturing the individual station temporal structure and the spatial correlation at any fixed point in time.

[40] Figure 6 displays the observed frequencies of total number of stations in Iowa with positive precipitation, as well as 90% pointwise model confidence intervals based on our 100 simulations. We see that the stochastic generator faithfully replicates this statistic well, except for slightly undersimulating the frequency of all sites being dry. This suggests that precipitation occurs at least one location more often than specified by our model, although we replicate the average number of daily spatial occurrences for all other values well.

[41] We conclude our discussion of the spatial occurrence model by examining a plot similar to that of *Wilks* [1998], pairwise lagged simultaneous occurrence probabilities. Each point in Figure 7 represents the probability

$$P(O(\boldsymbol{s}_i, t-1) = 0, \quad O(\boldsymbol{s}_j, t) = 1) \tag{18}$$

$$P(O(\boldsymbol{s}_i, t-1) = 1, \quad O(\boldsymbol{s}_j, t) = 0) \tag{19}$$

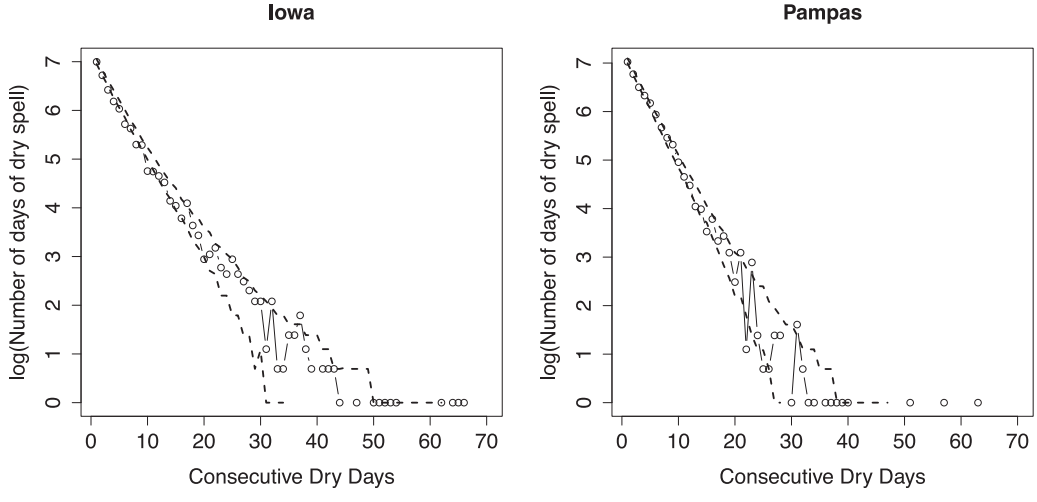

Figure 3. Log frequency of empirical dry spells at Rock Rapids, Iowa, and Buenos Aires, Argentina, with 90% pointwise confidence intervals based on 100 simulations as dashed lines.

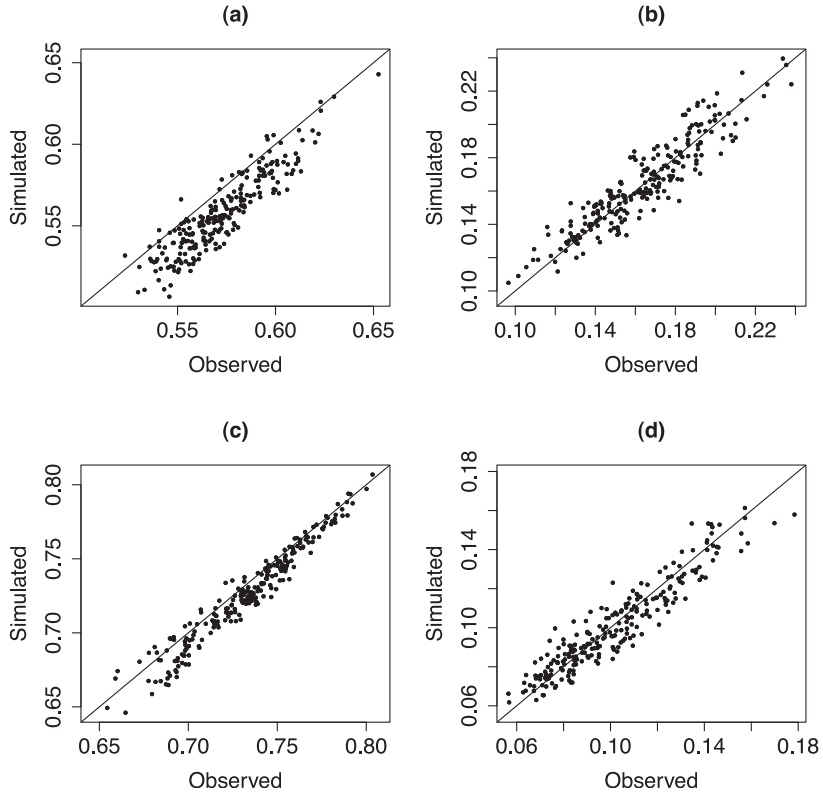

Figure 4. Pairwise empirical simultaneous occurrence probabilities in Iowa for both stations being (a) dry in summer (June–August), (b) wet in summer, (c) dry in winter (December–February), and (d) wet in winter.

for the simulated and observed data at Iowa. These statistics are difficult to replicate, especially because our model is not directly tuned to do so. Notice our approach seems to imply a slight additive bias in the simulated probabilities, but removes the quite noticeable multiplicative bias exhibited by *Wilks* [1998] in an application to another precipitation data set. Other approaches have been able to further reduce this challenging criterion [*Lee et al.*, 2010], but often come at the cost of significant model complications. Finally, note that this additive bias is nearly negligible, only on the order of 0.01 to 0.02, which is unlikely to have a significant impact in practice.

[42] We now turn to validating the spatial intensity model, with special attention to spatially aggregated performance. Table 2 displays monthly statistics of daily precipitation intensity across the domain of Iowa. In particular, the average simulated intensity in any given month differs from observed averages by typically approximately 1 mm, while the simulated standard deviation of intensity differs from the observed variability by less than 1 mm on average. Hence, the model exhibits no apparent bias in the first or second distributional moments.

[43] To examine the interannual variability, Figure 8 shows the standard deviation of total monthly precipitation at a single example location, Belle Plaine, Iowa. The two plots include our model where the parameters are simulated at each location (thereby incorporating parameter uncertainty), and the same model but using fixed parameters. Our approach shows an ability to scale to time frames longer than daily, which is a crucial desired feature of stochastic precipitation generators [*Gregory et al.*, 1993]. The greater variability in model simulated monthly total precipitation seen in Figure 8a is directly due to incorporation of parameter uncertainty, with the model not being trained to replicate this statistic. Our model performs similarly well for the other locations in Iowa.

[44] Figure 9 shows observed pairwise station correlations of precipitation intensity (i.e., only on days with simultaneous positive precipitation) against the simulated correlations for summer and winter in Iowa. The ability to

Table 1. Average and Standard Deviation (SD) of Daily Occurrence Rate by Month Over the Domain of Iowa

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed mean	0.180	0.189	0.236	0.305	0.349	0.336	0.273	0.272	0.267	0.220	0.196	0.185
Simulated mean	0.178	0.199	0.249	0.309	0.343	0.336	0.302	0.273	0.253	0.236	0.212	0.186
Observed SD	0.384	0.391	0.424	0.460	0.477	0.473	0.446	0.445	0.442	0.415	0.397	0.388
Simulated SD	0.382	0.399	0.432	0.462	0.475	0.472	0.459	0.445	0.435	0.424	0.409	0.389

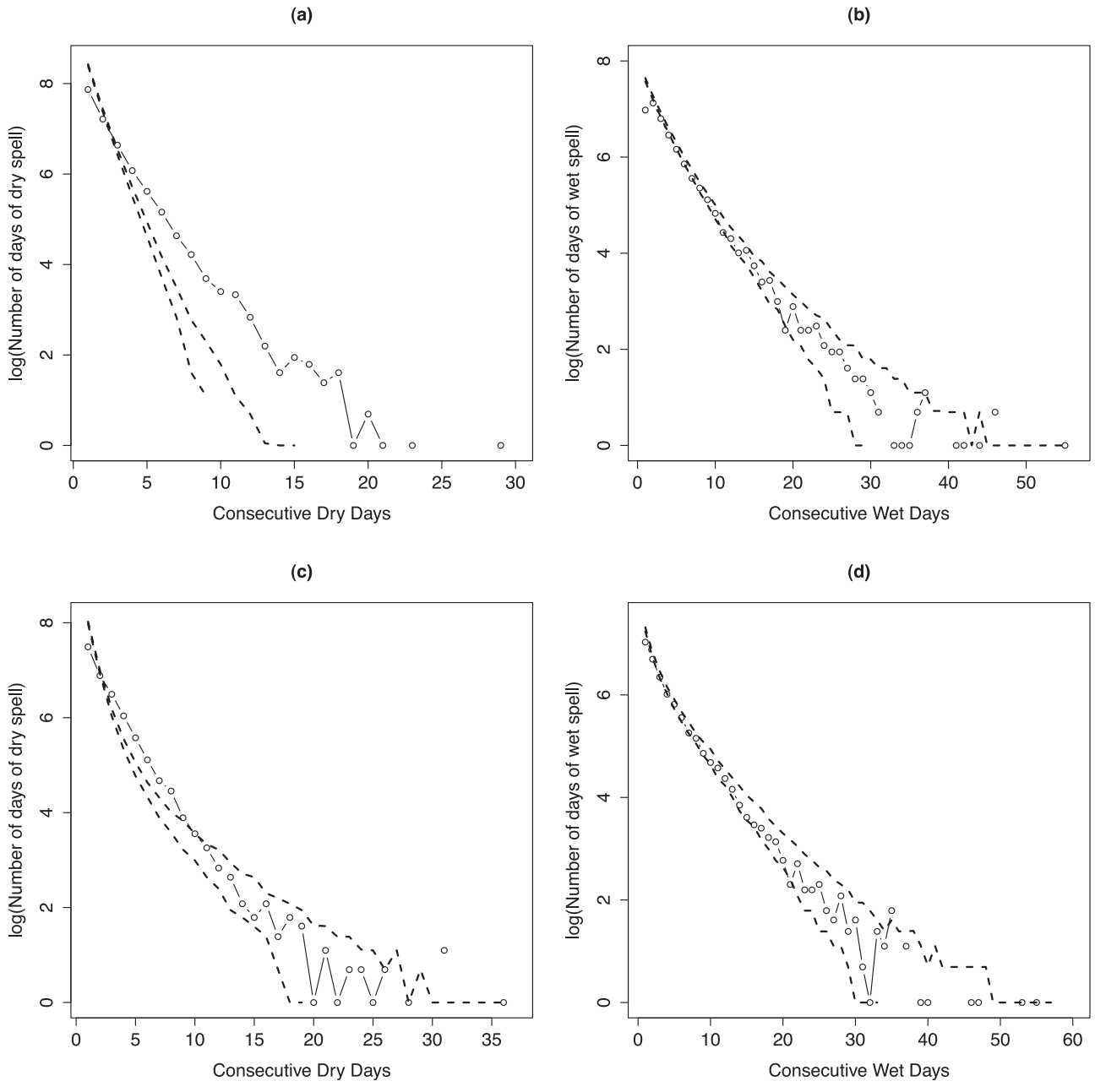

Figure 5. Log frequency of domain aggregate spells: (a) dry in Iowa, (b) wet in Iowa, (c) dry in the Pampas, and (d) wet in the Pampas, with 90% pointwise confidence intervals based on 100 simulations as dashed lines.

adapt to seasons is due to the temporally evolving spatial correlation function (4), which allows for a substantial increase in spatial correlation during winter as compared to summer. It should be unsurprising that the pairwise correlations of Figure 9 are more variable than for occurrence (recall Figure 4), since the spatial correlation of intensity has a shorter length scale and is much more variable. To illustrate the seasonally varying spatial correlation, Figure 10 shows the spatial correlation function for the Iowa domain for the two example dates of 1 January and 1 July, with empirical correlation functions being on the Gaussian process scale. Our method of moments approach to estimating spatial correlation shows good performance.

[45] One of the main motivations for spatially consistent precipitation is to capture the domain aggregated behavior, that is, the total precipitation averaged over all locations on a given day. Figure 11 shows Q-Q plots for daily regional average precipitation under both our spatial model and a marginal model in Iowa and the Pampas region. The marginal model retains the same marginal probability distribution for precipitation intensity at each individual location, but has no transformed Gaussian processes to correlate spatial precipitation. Our spatial approach significantly improves over the marginal model, and retains the same marginal distribution. The main room for improvement in Iowa lies above about the 90% quantile in Figure 11; at

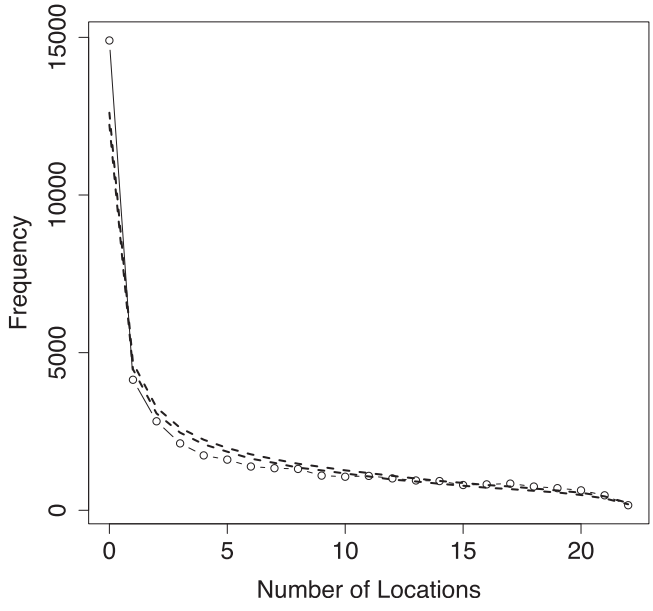

Figure 6. Observed number of stations in Iowa at which precipitation occurred with 90% pointwise model confidence intervals based on 100 simulations as dashed lines.

this extreme level, domain average precipitation intensity in Iowa is not well modeled by a sum of gamma variables, although extreme spatial precipitation is reasonably accurately replicated in the Pampas.

[46] A final key requirement for a spatially consistent stochastic generator is to replicate the observed variability of annual total precipitation averaged over the domain, as important for some water management purposes. To this end, Figures 12 and 13 display Q-Q plots for annual regional mean precipitation, with 90% pointwise confidence bounds based on our 100 simulations for both data sets. Figure 12 also includes a Q-Q plot for annual mean precipitation at Belle Plaine, Iowa, which illustrates that our model scales well from daily to annual time frames both locally and regionally. It is important to be aware that our model has not been trained to reproduce this statistic.

3.4. Gridded Simulations

[47] The advantage of the spatial model (5) is that model parameters can be interpolated to any location of interest, including parameter uncertainty for both the occurrence and intensity processes. The Gaussian process spatial model lends itself to using a spatial interpolation technique called kriging [*Cressie*, 1993]. The kriging predictor is the best linear unbiased predictor (in the sense of quadratic loss), and coincides with conditional expectation for normally distributed variables. In particular, the conditional distribution of $\beta(\boldsymbol{s}_0)$ based on the partial realization $\widehat{\beta}(\boldsymbol{s})$ at $\boldsymbol{s} = \boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ is normal with kriging mean and variance

$$\mathrm{E}(\beta(\boldsymbol{s}_0) \mid \widehat{\beta}(\boldsymbol{s}_1), \ldots, \widehat{\beta}(\boldsymbol{s}_n)) = \mu + \boldsymbol{c}'\Sigma^{-1}(\widehat{\beta} - \mu) \tag{20}$$

$$\mathrm{Var}(\beta(\boldsymbol{s}_0) \mid \widehat{\beta}(\boldsymbol{s}_1), \ldots, \widehat{\beta}(\boldsymbol{s}_n)) = \sigma^2 + \tau^2 - \boldsymbol{c}'\Sigma^{-1}\boldsymbol{c} \tag{21}$$

where μ and σ^2 are the mean and marginal variance of the $Z(\boldsymbol{s})$ process, respectively, τ^2 is the nugget effect arising from the $\epsilon(\boldsymbol{s})$ process, Σ is the covariance matrix of $(\beta(\boldsymbol{s}_1), \ldots, \beta(\boldsymbol{s}_n))'$, and the ith entry of the vector $\boldsymbol{c}$ is $\mathrm{Cov}(\beta(\boldsymbol{s}_0), \beta(\boldsymbol{s}_i))$. One has two options for interpolating parameters with the kriging model: (1) use the deterministic point prediction of (20); or (2) use a simulation which is normally distributed with mean and variance given by (20) and (21), respectively. This second approach recognizes that the kriging mean is a good average value, but is also able to communicate the uncertainty in this predictor to the weather generator.

[48] Accounting for parameter uncertainty is important to reduce overdispersion, which is otherwise magnified because of the inflexibility of the fixed statistical model (as we have already seen; recall Figure 8). To this end, Figure 14 shows monthly and annual variability of an observed intensity series and 100 simulated trajectories at a held out location (Mount Ayr, Iowa). That is, the simulations are based on interpolated statistical parameters, using either the

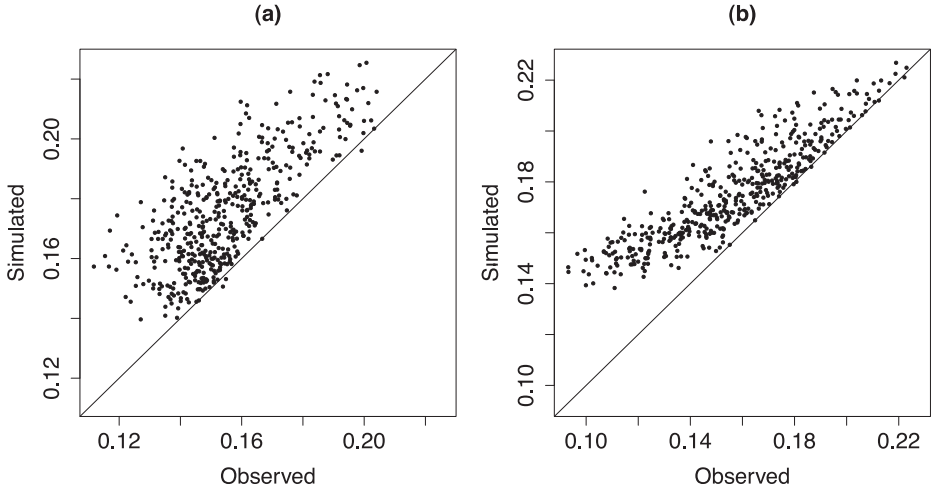

Figure 7. Joint probability of precipitation occurrence state of two sites in Iowa, where each site was oppositely (a) dry and wet on successive days or (b) wet and dry on successive days.

Table 2. Mean and Standard Deviation (SD) of Daily Precipitation Intensity by Month Over the Domain of Iowa[a]

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed mean	4.66	5.12	6.94	8.56	9.83	11.69	11.67	11.74	11.48	8.69	7.50	5.28
Simulated mean	5.44	5.70	6.66	8.36	10.55	12.51	13.51	12.96	11.17	8.94	7.08	5.88
Observed SD	5.99	6.29	8.35	10.23	12.13	14.85	15.14	15.9	15.07	10.86	9.85	6.75
Simulated SD	6.12	6.46	7.69	9.89	12.68	15.22	16.59	16.09	13.9	10.94	8.39	6.72

[a]Units are millimeters.

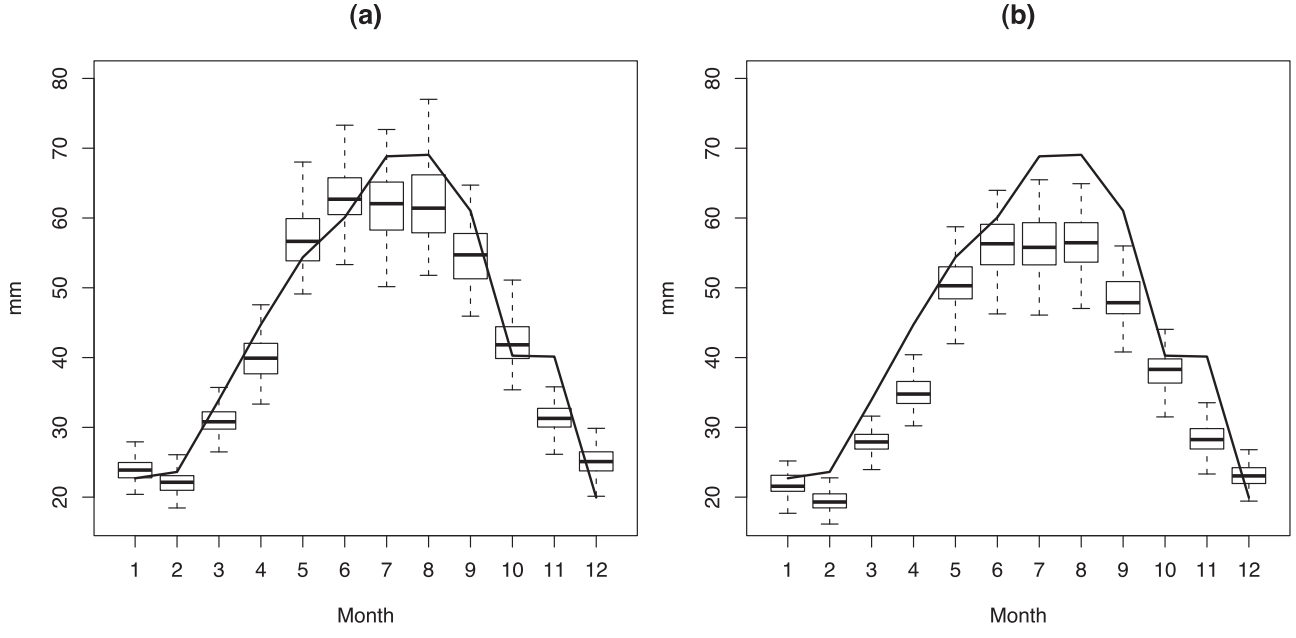

Figure 8. Standard deviation of monthly total precipitation at an example location (Belle Plaine, Iowa) indicated by the solid curve, with box plots showing the simulated interannual variability over 100 simulations for the case where (a) model parameters are simulated on each day and (b) model parameters are fixed on each day. Box plot whiskers extend to 1.5 times the interquartile range.

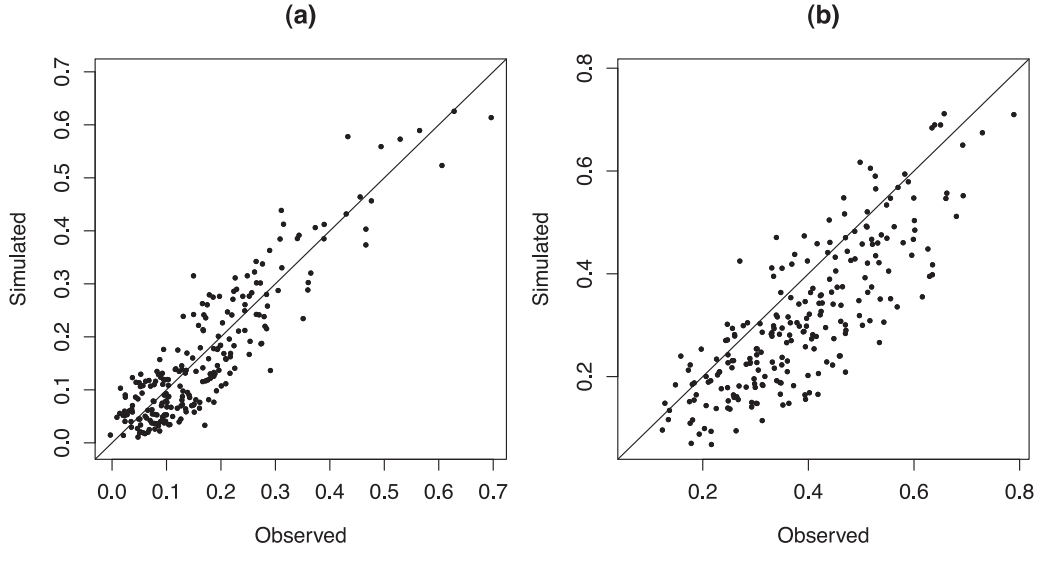

Figure 9. Pairwise empirical correlations for precipitation intensity in Iowa in (a) summer and (b) winter.

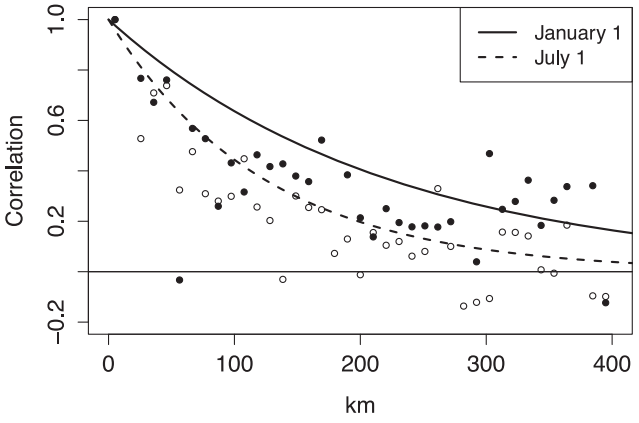

Figure 10. Spatial correlation function (4) for the intensity model on the Gaussian process scale with empirical covariances for the domain of Iowa for 1 January (solid line and solid circles) and 1 July (dashed line and open circles).

fixed kriging mean (20), or using parameter values centered around (20) with normal deviations with variance defined by (21). Immediately, we see that using a deterministic interpolation does not provide variable enough simulations on either the monthly or annual scale, whereas including parameter uncertainty yields an appropriate amount of spread at both of these scales. The effect of parameter uncertainty on precipitation intensity is more readily seen, whereas the increased variability is less clear for occurrence simulation.

[49] One consequence of simulating the interpolated parameter, rather than fixing it at the kriging mean (20) is that the simulated series will typically be slightly more variable than the observations at that particular location. Mathematically, this is due to the fact that the kriging variance (21) is always greater than the local parameter uncertainty contained in τ^2. Heuristically, this is due to the fact that we cannot directly estimate the parameters at this location, hence our prediction uncertainty along with the variability endowed by the statistical model imply a greater spread of stochastic realizations.

[50] To complete this point, consider Figure 15, which shows interannual variability for both types of model interpolation at Mount Ayr, Iowa. In particular, the parameters governing the simulations at this location are unknown, and are interpolated from surrounding locations. Here, we calculate interannual variability as the standard deviation of monthly totaled precipitation at this particular location. Figure 15a is from the model where the precipitation generator parameters (β) are simulated at this location on the basis of the distribution of (20) and (21), whereas Figure 15b uses the deterministic kriging interpolator from (20). Clearly accounting for parameter uncertainty is a critical consideration for stochastic precipitation generators, as we see the overdispersion if left unaccounted for.

[51] We close section 3.4 with an example of gridded precipitation that illustrates our model's practical use. Figure 16 shows two randomly generated precipitation fields for Iowa on adjacent days. Parameters were simulated at each grid point using the kriging model, and then a field of occurrence and intensity were randomly generated. The first field's covariates were set to 1 January conditions, and the second field occurred on 2 January, conditional on the observed precipitation as simulated on 1 January. We see the temporal momentum of the precipitation field, with high intensities and positive precipitation tending to occur near the grid cells where there was rain on the previous day. The model is available at every location, and once the first day of precipitation has been generated the chains of precipitation fields can be simulated for any arbitrary number of days.

4. Summary and Discussion

[52] We have presented a general framework for simulating spatially correlated fields of daily precipitation. The method relies on a latent Gaussian process that drives precipitation occurrence, and the consequent field of

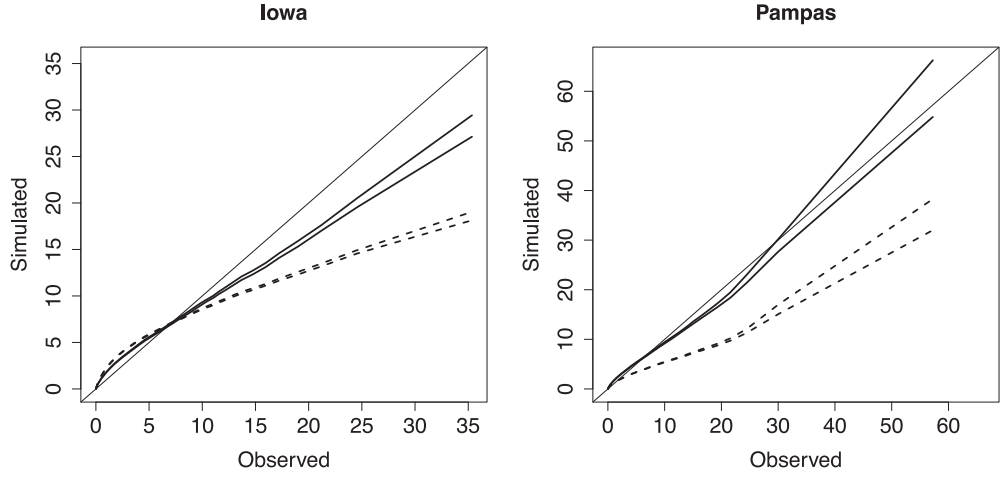

Figure 11. Q-Q plot for domain aggregated average daily precipitation in Iowa and the Pampas with 90% pointwise confidence intervals for the spatial model (solid line) and a marginal model (dashed line) that retains the gamma distribution at individual locations but has no spatial correlation included in either the occurrence or intensity processes; units are mm.

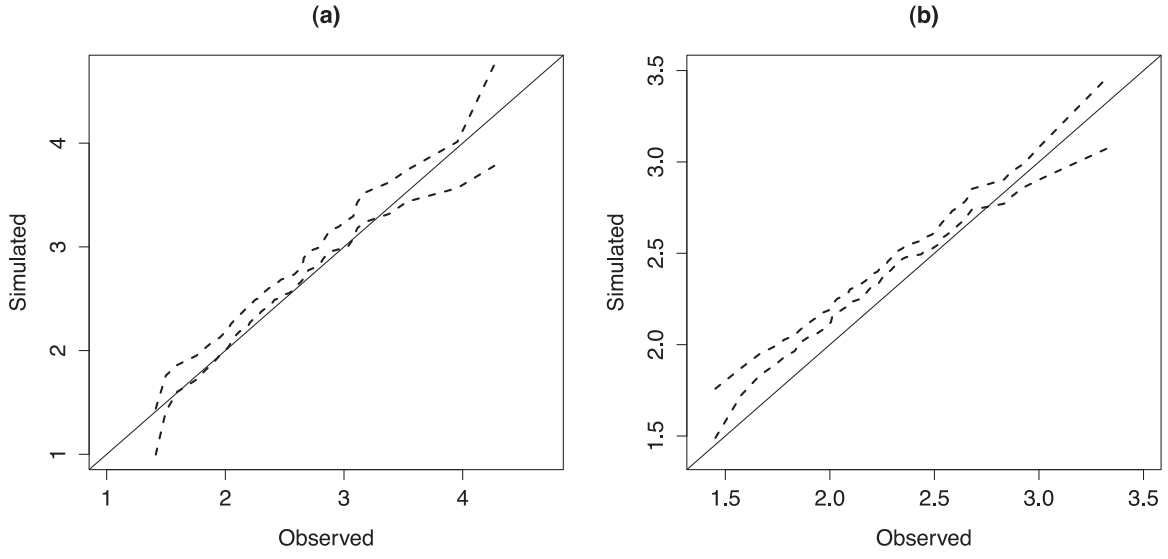

Figure 12. Q-Q plots for (a) annual mean precipitation at Belle Plaine, Iowa, and (b) annual domain mean precipitation in Iowa with 90% pointwise confidence intervals based on 100 simulations as dashed lines; units are mm.

precipitation intensities is modeled as a transformed Gaussian process which reduces to a gamma distribution at individual locations. Seasonally varying occurrence and intensity is achieved through a GLM, which can be easily extended to include other covariates for the purposes of downscaling, climate impacts [*Qian et al.*, 2002; *Mehrotra and Sharma*, 2010], climate change scenarios [*Katz*, 1996], or observational record infilling [*Kyriakidis et al.*, 2004].

[53] Modeling parameters as spatial Gaussian processes allows for quantification of parameter uncertainties and spatial interpolation using kriging. In effect, this allows the stochastic model to vary by location, which is crucial when dealing with large domains or regions of complex topography [*Guan et al.*, 2005; *Hay et al.*, 1998; *Thornton et al.*, 1997]. For example, *Goovaerts* [2000] discusses methods of incorporating elevation information into the geostatistical framework. Kriging is a preferred technique for spatial interpolations, as not only are point estimates available at any arbitrary location, but parameter uncertainty is immediately described. Deterministic interpolations can result in underdispersion of the stochastic realizations, whereas including interpolation uncertainty yields greater variability in the simulations. Indeed, simulating model parameters is a straightforward and easy way to reduce overdispersion, whereas other authors have needed to postprocess their daily simulations to align with monthly and annual scales [*Srikanthan and Pegram*, 2009].

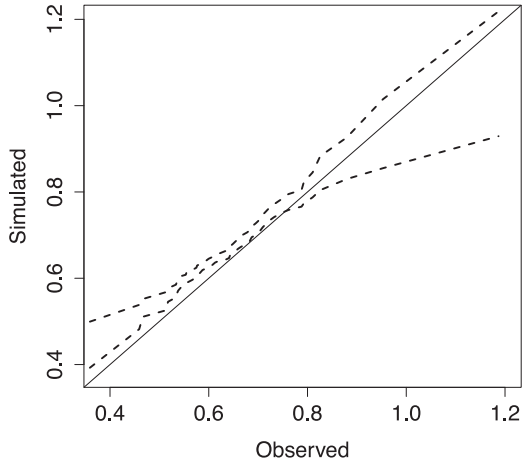

Figure 13. Q-Q plot for annual domain mean precipitation in the Pampas with 90% pointwise confidence intervals based on 100 simulations; units are mm.

[54] When interpolating model parameters to locations with no observations, our kriging conditional distribution should be contrasted with the approach of *Wilks* [2008, 2009]. He considered a deterministic interpolation scheme, an extension of that used by *Lall et al.* [2006]. Their locally weighted regression scheme can take account of correlation decay across distance as well as elevation. The drawback to this approach is that uncertainty in parameter predictions is not directly used, whereas uncertainty is immediately available at all locations in closed form using kriging. Correlation dependence on elevation can be built into the covariance function as well; see *Kleiber et al.* [2011] for one approach.

[55] We illustrated our model on two data sets, the first was a network of 22 stations in Iowa, and the second a network of 19 stations over a significantly larger domain in the Pampas region of Argentina. The method shows a good ability to replicate dry and wet spells both locally and on a domain aggregated level. In Iowa, our approach generated slightly too few aggregate dry spells of extreme length, while in the Pampas the model captured domain aggregated wet and dry spells very well. Our tentative hypothesis for this apparent disagreement in performance is that the model in Iowa is failing to capture the two types of precipitation common during the summer, the local convective storms and the widespread frontal precipitation events. In the Pampas region, the convective occurrences tend to be much more widespread and long lasting than in the midwestern

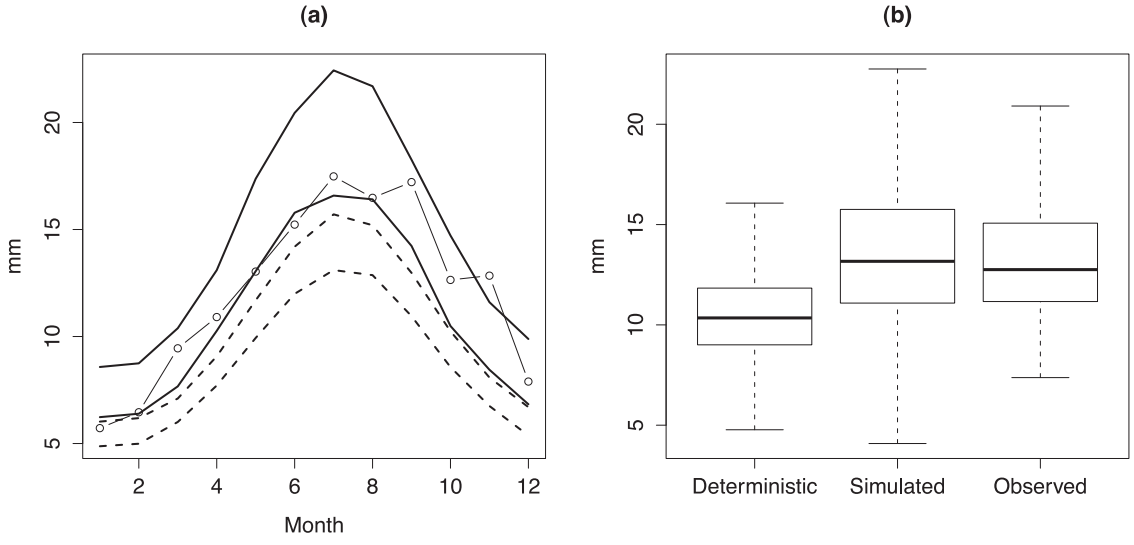

Figure 14. Standard deviation of precipitation intensity at a held out location (Mount Ayr, Iowa), using an interpolated model: (a) Standard deviation of intensity by month with pointwise 90% confidence intervals based on 100 simulations with parameter uncertainty included (solid lines) and deterministic interpolations (dashed lines). (b) Box plots of annual standard deviation of precipitation intensity with deterministic interpolation (deterministic), simulated parameters (simulated) and observed standard deviation (observed). Box plot whiskers extend to 1.5 times the interquartile range.

United States [*Durkee and Mote*, 2009]. Hence, the temporally varying spatial correlation in the Pampas is sufficient to effectively distinguish between the two types of precipitation, while a more complicated model that explicitly accounts for both types of precipitation may be required in Iowa. In particular, during the summer our spatial length scale of occurrence is shrunk via the temporally varying range parameter in both Iowa and the Pampas, and hence our model exists somewhere between highly localized convective storms and widespread frontal events. One possible solution is to consider a model that separately accounts for both of these types of precipitation, which might in turn replicate the regional dry spells more accurately; this would be a difficult step to take but would certainly be

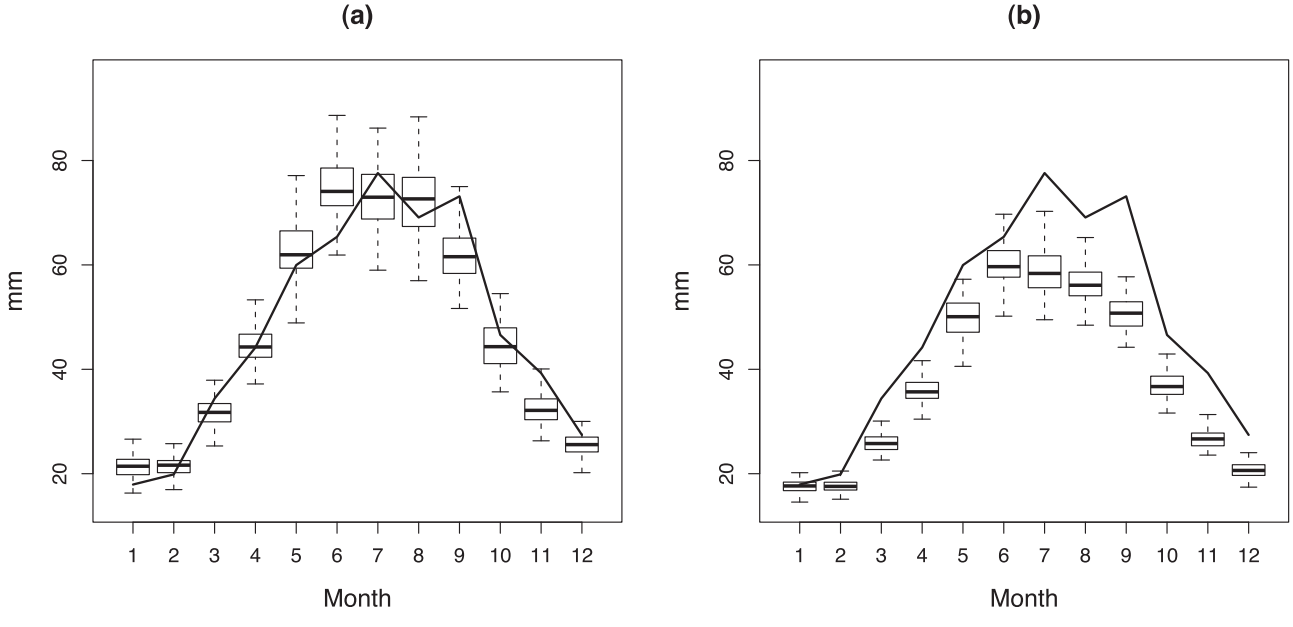

Figure 15. Interannual variability for each month in terms of standard deviation of monthly summed rainfall at a held out location (Mount Ayr, Iowa). Box plots are based on 100 simulations from an interpolated model with (a) parameter uncertainty included and (b) deterministic parameter interpolation. The solid line indicates the observed standard deviation of observed monthly precipitation. Box plot whiskers extend to 1.5 times the interquartile range.

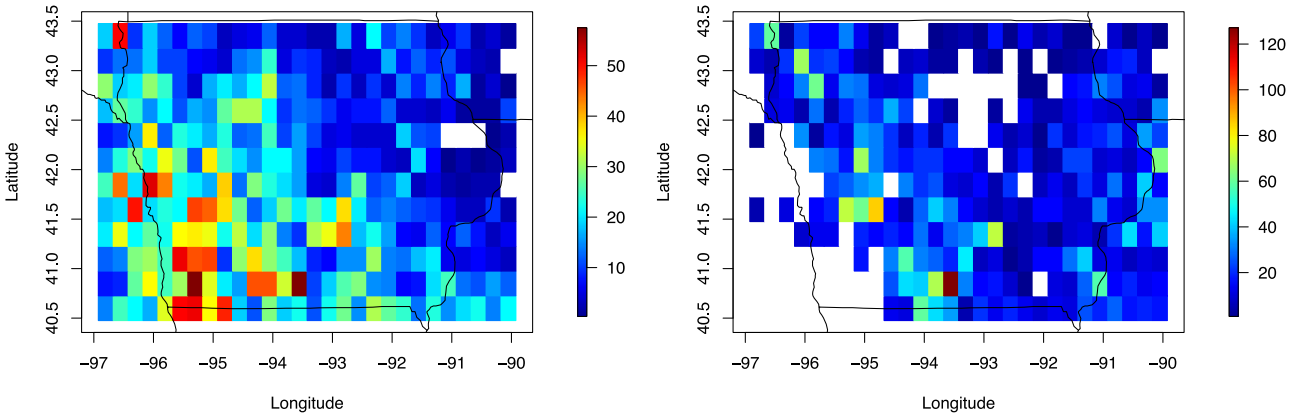

Figure 16. Two gridded precipitation simulations in Iowa with grid spacing of approximately 20 km, the first on 1 January and the second on 2 January; units are mm.

worthy of further research. In the Pampas, our model replicated the extreme wet and dry spells extremely accurately, displaying the flexibility of our approach in different domains.

[56] For both precipitation occurrence and intensity, we used an isotropic correlation function whose scale parameter varied with time. Alternative anisotropic and nonstationary models may be of interest in other settings with complex terrain or over a larger domain [*Baigorria et al.*, 2007]. Extending our model to include spatial anisotropy is straightforward. The nonstationary and anisotropic covariance models of *Paciorek and Schervish* [2006] may be of particular interest. In our domain, we had at most 22 stations, and with such sparse data it is extremely difficult to identify, let alone model spatial anisotropy. Potentially, with a larger network of stations, one could identify and successfully fit an anisotropic model, but this is beyond the scope of our present examples.

[57] At individual locations, our model reduces to a Markov chain for precipitation occurrence and, conditional on the occurrence of precipitation, to a gamma distribution for intensity. The drawback to gamma distributions is that they do not possess a heavy enough tail and hence cannot adequately represent extreme rainfall events [*Furrer and Katz*, 2008]. However, methods of combining spatial models for extremes with those for lower values are few and far between. It would be desirable to combine a model such as ours with that of *Buishand et al.* [2008], for example, who describe a model for spatially correlated extreme precipitation.

[58] Our daily precipitation model was able to reproduce the observed variability in daily rainfall occurrences and intensities on local and domain aggregated scales, as well as on longer temporal scales such as monthly and annually. Hence, our approach requires no extra effort to incorporate the important low-frequency behavior and interannual variability required of precipitation generators [*Gregory et al.*, 1993; *Wilks and Wilby*, 1999].

[59] Another future route of research is to develop a full stochastic weather generator that considers not only precipitation, but other variables such as temperature, wind speed, relative humidity and solar radiation across space simultaneously (e.g., as modeled at a single site by *Parlange and Katz* [2000] or *Kilsby et al.* [2007] and also in a multisite context by *Wilks* [2009]). A similar approach to ours for temperature field simulation can be implemented using the methods of *Berrocal et al.* [2007] and *Kleiber et al.* [2011]. Some of the multivariate spatial models currently of interest in the statistics community may also be useful to this end [*Apanasovich and Genton*, 2010; *Gelfand et al.*, 2004; *Gneiting et al.*, 2010].

[60] **Acknowledgments.** We gratefully acknowledge the comments of the three reviewers, Francesco Serinaldi and two anonymous reviewers. We thank Guillermo Podestá for facilitating the use of Argentina daily weather data, the National Meteorological Service of Argentina for providing the data, and Maria Skansi for quality control of the data. This research was supported by the NCAR Weather and Climate Assessment Science Program and by NSF Coupled Natural Human Systems Program grant CNH-0709681. NCAR is sponsored by the National Science Foundation.

References

Ailliot, P., C. Thompson, and P. Thomson (2009), Spacetime modelling of precipitation by using a hidden Markov model and censored Gaussian distributions, *Appl. Stat.*, *58*, 405–426.

Allcroft, D. J., and C. A. Glasbey (2003), A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation, *Appl. Stat.*, *52*, 487–498.

Apanasovich, T. V., and M. G. Genton (2010), Cross-covariance functions for multivariate random fields based on latent dimensions, *Biometrika*, *97*, 15–30.

Apipattanavis, S., G. Podestá, B. Rajagopalan, and R. W. Katz (2007), A semiparametric multivariate and multisite weather generator, *Water Resour. Res.*, *43*, W11401, doi:10.1029/2006WR005714.

Baigorria, G. A., and J. W. Jones (2010), GiST: A stochastic model for generating spatially and temporally correlated daily rainfall data, *J. Clim.*, *23*, 5990–6008.

Baigorria, G. A., J. W. Jones, and J. J. O'Brien (2007), Understanding rainfall spatial variability in southeast USA at different time scales, *Int. J. Climatol.*, *27*, 749–760.

Bárdossy, A., and G. G. S. Pegram (2009), Copula based multisite model for daily precipitation simulation, *Hydrol. Earth Syst. Sci.*, *13*, 2299–2314, doi:10.5194/hess-13-2299-2009.

Berrocal, V. J., A. E. Raftery, and T. Gneiting (2007), Combining spatial statistical and ensemble information for probabilistic weather forecasting, *Mon. Weather Rev.*, *135*, 1386–1402.

Berrocal, V. J., A. E. Raftery, and T. Gneiting (2008), Probabilistic quantitative precipitation field forecasting using a two-stage spatial model, *Ann. Appl. Stat.*, *2*, 1170–1193.

Brissette, F. P., M. Khalili, and R. Leconte (2007), Efficient stochastic generation of multi-site synthetic precipitation data, *J. Hydrol.*, *345*, 121–133.

Buishand, T. A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resour. Res.*, *37*(11), 2761–2776.

Buishand, T. A., L. deHaan, and C. Zhou (2008), On spatial extremes: With application to a rainfall problem, *Ann. Appl. Stat.*, *2*, 624–642.

Burton, A., C. G. Kilsby, C. G. Fowler, P. S. P. Cowpertwait, and P. E. O'Connell (2008), RainSim: A spatial-temporal stochastic rainfall modelling system, *Environ. Modell. Software*, *23*, 1356–1369.

Cannon, A. J. (2008), Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network, *J. Hydrometeorol.*, *9*, 1284–1300.

Charles, S. P., B. C. Bates, and J. P. Hughes (1999), A spatiotemporal model for downscaling precipitation occurrence and amounts, *J. Geophys. Res.*, *104*(D24), 31,657–31,669.

Chilès, J. P., and P. Delfiner (1999), *Geostatistics: Modeling Spatial Uncertainty*, John Wiley, New York.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, rev. ed., John Wiley, New York.

Durban, M., and C. A. Glasbey (2001), Weather modelling using a multivariate latent Gaussian model, *Agric. For. Meteorol.*, *109*, 187–201.

Durkee, J. D., and T. L. Mote (2009), A climatology of warm-season mesoscale convective complexes in subtropical South America, *Int. J. Climatol.*, doi:10.1002/joc.1893, in press.

Furrer, E. M., and R. W. Katz (2007), Generalized linear modeling approach to stochastic weather generators, *Clim. Res.*, *34*, 129–144.

Furrer, E. M., and R. W. Katz (2008), Improving the simulation of extreme precipitation events by stochastic weather generators, *Water Resour. Res.*, *44*, W12439, doi:10.1029/2008WR007316.

Gelfand, A. E., A. M. Schmidt, S. Banerjee, and C. F. Sirmans (2004), Nonstationary multivariate process modeling through spatially varying coregionalization (with discussion and rejoinder), *Test*, *13*, 263–312.

Gneiting, T., W. Kleiber, and M. Schlather (2010), Matérn cross-covariance functions for multivariate random fields, *J. Am. Stat. Assoc.*, *105*, 1167–1177.

Goovaerts, P. (2000), Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall, *J. Hydrol.*, *228*(1–2), 113–129.

Gregory, J. M., T. M. L. Wigley, and P. D. Jones (1993), Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals, *Clim. Dyn.*, *8*, 299–310.

Guan, H., J. L. Wilson, and O. Makhnin (2005), Geostatistical mapping of mountain precipitation incorporating autosearched effects of terrain and climate characteristics, *J. Hydrometeorol.*, *6*, 1018–1031.

Hay, L., R. Viger, and G. McCabe (1998), Precipitation interpolation in mountainous regions using multiple linear regression, *IAHS Publ.*, *248*, 33–38.

Hughes, J. P., and P. Guttorp (1999), A non-homogeneous hidden Markov model for precipitation occurrence, *Appl. Stat.*, *48*, 15–30.

Johnson, G. L., C. Daly, G. H. Taylor, and C. L. Hanson (2000), Spatial variability and interpolation of stochastic weather simulation model parameters, *J. Appl. Meteorol.*, *39*, 778–795.

Katz, R. W. (1977), Precipitation as a chain-dependent process, *J. Appl. Meteorol.*, *16*, 671–676.

Katz, R. W. (1996), Use of conditional stochastic models to generate climate change scenarios, *Clim. Change*, *32*, 237–255.

Katz, R. W., and X. Zheng (1999), Mixture model for overdispersion of precipitation, *J. Clim.*, *12*, 2528–2537.

Kilsby, C. G., P. D. Jones, A. Burton, A. C. Ford, H. J. Fowler, C. Harpham, P. James, A. Smith, and R. L. Wilby (2007), A daily weather generator for use in climate change studies, *Environ. Modell. Software*, *22*, 1705–1719.

Kitanidis, P. K., and R. W. Lane (1985), Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton method, *J. Hydrol.*, *79*, 53–71.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and E. Grimit (2011), Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging, *Mon. Weather Rev.*, *139*, 2630–2649, doi:10.1175/2010MWR3511.1.

Kleiber, W., A. E. Raftery, and T. Gneiting (2012), Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting, *J. Am. Stat. Assoc.*, in press.

Kyriakidis, P. C., N. L. Miller, and J. Kim (2004), A spatial time series framework for simulating daily precipitation at regional scales, *J. Hydrol.*, *297*, 236–255.

Lall, U., Y. I. Moon, H. H. Kwon, and K. Bosworth (2006), Locally weighted polynomial regression: Parameter choice and application to forecasts of the Great Salt Lake, *Water Resour. Res.*, *42*, W05422, doi:10.1029/2004WR003782.

Lee, D., H. An, Y. Lee, J. Lee, H. Lee, and H. Oh (2010), Improved multisite stochastic weather generation with applications to historical data in South Korea, *Asia Pac. J. Atmos. Sci.*, *46*, 497–504.

Lima, C. H. R., and U. Lall (2009), Hierarchical Bayesian modeling of multisite daily rainfall occurrence: Rainy season onset, peak and end, *Water Resour. Res.*, *45*, W07422, doi:10.1029/2008WR007485.

Maraun, D., et al. (2010), Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user, *Rev. Geophys.*, *48*, RG3003, doi:10.1029/2009RG000314.

McCullagh, P., and J. A. Nelder (1989), *Generalized Linear Models*, Chapman and Hall, London.

Mehrotra, R., and A. Sharma (2010), Development and application of a multisite rainfall stochastic downscaling framework for climate change impact assessment, *Water Resour. Res.*, *46*, W07526, doi:10.1029/2009WR008423.

Mehrotra, R., R. Srikanthan, and A. Sharma (2006), A comparison of three stochastic multi-site precipitation occurrence generators, *J. Hydrol.*, *331*, 280–292.

Menne, M. J., C. N. Williams Jr., and R. S. Vose (2010), United States Historical Climatology Network (USHCN) Version 2 Serial Monthly Dataset, Carbon Dioxide Inf. Anal. Cent., Oak Ridge Natl. Lab., Oak Ridge, Tenn. [available at http://cdiac.ornl.gov/epubs/ndp/ushcn/daily_doc.html].

Nielsen, S. F. (2000), The stochastic EM algorithm: Estimation and asymptotic results, *Bernoulli*, *6*, 457–489.

Onof, C., R. E. Chandler, A. Kakou, P. Northrop, H. S. Wheater, and V. Isham (2000), Rainfall modelling using Poisson-cluster processes: A review of developments, *Stochastic Environ. Res. Risk Assess.*, *14*, 384–411.

Paciorek, C. J., and M. J. Schervish (2006), Spatial modelling using a new class of nonstationary covariance functions, *Environmetrics*, *17*, 483–506.

Parlange, M. B., and R. W. Katz (2000), An extended version of the Richardson model for simulating daily weather variables, *J. Appl. Meteorol.*, *39*, 610–622.

Qian, B., J. Corte-Real, and H. Xu (2002), Multisite stochastic weather models for impact studies, *Int. J. Climatol.*, *22*, 1377–1397.

Rajagopalan, B., and U. Lall (1999), A *k*-nearest neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res.*, *35*, 3089–3101.

Richardson, C. W. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resour. Res.*, *17*, 182–190.

Rodríguez-Iturbe, I., D. R. Cox, and V. Isham (1988), A point process model for rainfall: Further developments, *Proc. R. Soc. London, Ser. A*, *417*, 283–298.

Sansó, B., and L. Guenni (2000), A nonstationary multisite model for rainfall, *J. Am. Stat. Assoc.*, *95*(452), 1089–1100.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*, 461–464.

Srikanthan, R., and G. G. S. Pegram (2009), A nested multisite daily rainfall stochastic generation model, *J. Hydrol.*, *371*, 142–153.

Stein, M. L., (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

Stern, R. D., and R. Coe (1984), A model fitting analysis of daily rainfall data, *J. R. Stat. Soc., Ser. A*, *147*, 1–34.

Thompson, C. S., P. J. Thomson, and X. Zheng (2007), Fitting a multisite daily rainfall model to New Zealand data, *J. Hydrol.*, *340*, 25–39.

Thornton, P. E., S. W. Running, and M. A. White (1997), Generating surfaces of daily meteorological variables over large regions of complex terrain, *J. Hydrol.*, *190*, 214–251.

Valdes, J. B., and I. Rodriguez-Iturbe (1985), Approximations of temporal rainfall from a multidimensional model, *Water Resour. Res.*, *21*, 1259–1270.

Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol.*, *210*, 178–191.

Wilks, D. S. (2008), High-resolution spatial interpolation of weather generator parameters using local weighted regressions, *Agric. For. Meteorol.*, *148*, 111–120.

Wilks, D. S. (2009), A gridded multisite weather generator and synchronization to observed weather data, *Water Resour. Res.*, *45*, W10419, doi:10.1029/2009WR007902.

Wilks, D. S. (2010), Use of stochastic weather generators for precipitation downscaling, *Wiley Interdiscip. Rev.*, *1*, 898–907, doi:10.1002/wcc.85.

Wilks, D. S., and R. L. Wilby (1999), The weather generation game: A review of stochastic weather models, *Prog. Phys. Geogr.*, *23*, 329–357.

Woolhiser, D. A., and G. G. S. Pegram (1979), Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models, *J. Appl. Meteorol.*, *18*, 34–42.

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.

Zheng, X., and R. W. Katz (2008), Simulation of spatial dependence in daily rainfall using multisite generators, *Water Resour. Res.*, *44*, W09403, doi:10.1029/2007WR006399.

Zheng, X., J. Renwick, and A. Clark (2010), Simulation of multisite precipitation using an extended chain-dependent process, *Water Resour. Res.*, *46*, W01504, doi:10.1029/2008WR007526.

R. W. Katz and W. Kleiber, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307-3000, USA. (wkleiber@uw.edu)

B. Rajagopalan, Department of Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, CO 80309, USA.