AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2016WR019741

Key Points:
- ABC algorithms approximate posterior distributions by bypassing the need for a likelihood function
- Spell length ABC metrics for single-site precipitation occurrence approximate true posteriors well
- Variogram-based metrics accurately capture spatial dependence for multisite occurrence

Correspondence to:
B. Olson,
brando6@uw.edu

Citation:
Olson, B., and W. Kleiber (2017), Approximate Bayesian computation methods for daily spatiotemporal precipitation occurrence simulation, *Water Resour. Res.*, *53*, doi:10.1002/2016WR019741.

Received 1 SEP 2016
Accepted 19 MAR 2017
Accepted article online 28 MAR 2017

Approximate Bayesian computation methods for daily spatiotemporal precipitation occurrence simulation

Branden Olson[1] and William Kleiber[2]

[1]Department of Statistics, University of Washington, Seattle, Washington, USA, [2]Department of Applied Mathematics, University of Colorado, Boulder, Colorado, USA

Abstract Stochastic precipitation generators (SPGs) produce synthetic precipitation data and are frequently used to generate inputs for physical models throughout many scientific disciplines. Especially for large data sets, statistical parameter estimation is difficult due to the high dimensionality of the likelihood function. We propose techniques to estimate SPG parameters for spatiotemporal precipitation occurrence based on an emerging set of methods called Approximate Bayesian computation (ABC), which bypass the evaluation of a likelihood function. Our statistical model employs a thresholded Gaussian process that reduces to a probit regression at single sites. We identify appropriate ABC penalization metrics for our model parameters to produce simulations whose statistical characteristics closely resemble those of the observations. Spell length metrics are appropriate for single sites, while a variogram-based metric is proposed for spatial simulations. We present numerical case studies at sites in Colorado and Iowa where the estimated statistical model adequately reproduces local and domain statistics.

Plain Language Summary Statistical simulations of precipitation and other weather quantities are commonly used in many sciences. Modern datasets are extremely high dimensional, which challenge traditional model estimation paradigms. We propose a novel technique specially adapted for estimation using approximate Bayesian computation (ABC), and show how important characteristics such as dry and wet spells can be used to quantify uncertain model parameters. The proposed method offers promising future directions for further research.

1. Introduction

Scientific fields such as hydrology, ecology, meteorology, agriculture, and climate impact assessment frequently utilize physical models that require daily spatiotemporal weather as input conditions under differing scientific scenarios [*Kustas et al.*, 1994; *Richardson*, 1981; *Wilks*, 1988; *Mearns et al.*, 1997; *Friend et al.*, 1997; *Strandman et al.*, 1993; *Semenov and Barrow*, 1997]. "Weather" typically refers to daily minimum and maximum temperature, and quantitative precipitation. However, in situ measurements are limited in their temporal and spatial coverage, and are often contaminated with missing values. Thus, it is desirable to produce synthetic weather scenarios that are statistically similar to direct measurements of weather quantities, either for historical network infilling, or to test future scientific scenarios. Stochastic weather generators (SWGs) are statistical mechanisms to generate weather data that are statistically similar to observational records. SWGs for daily weather sequences are the most common due to their widespread applicability and availability of observational records collected on a day-to-day basis [*Wilks and Wilby*, 1999]. SWGs can be loosely categorized into two approaches: model-based and empirical. Model-based SWGs use formal statistical models, whereas empirical approaches generally rely on resampling of historical patterns, requiring a large historical database [*Racsko et al.*, 1991; *Richardson*, 1981; *Lall and Sharma*, 1996; *Rajagopalan and Lall*, 1999].

Traditionally, SWGs were built to produce simulations colocated with observational data, but recent developments focus on generating consistent spatiotemporal simulations on space-time grids [*Wilks*, 1999; *Kleiber et al.*, 2012]. Precipitation is particularly at the forefront of the literature due to its challenging mixed discrete-continuous nature, skewness and non-Gaussianity [*Ailliot et al.*, 2009; *Allcroft and Glasbey*, 2003; *Brown et al.*, 2001; *Durban and Glasbey*, 2001; *Hughes et al.*, 1999; *Sansó and Guenni*, 2000]. These model-based approaches sometimes split modeling of the intensity and occurrence fields, and often use Markov chains to implement temporal dependence [*Katz*, 1977]. For intensities, modelers customarily turn to an exponential distribution, gamma distribution, or some mixture thereof [*Richardson*, 1981; *Stern and Coe*, 1984; *Woolhiser and Pegram*, 1979]. More recently, stochastic precipitation generators (SPGs) have played a key role in statistical downscaling [*Maraun et al*., 2010; *Wilks*, 2010]. Moreover, current advances on conditional rainfall simulation incorporate atmospheric circulation variables as predictors for weather generation [*Langousis and Kaleris*, 2014; *Langousis et al*., 2016]. The popularity, scope, and multifaceted nature of SWGs as has led to several review studies [*Hewitson and Crane*, 1996; *Zorita and von Storch*, 1997; *Wilby et al*., 2004; *Fowler et al*., 2007].

Interest has shifted from local precipitation, or precipitation at a single location, to spatiotemporally correlated precipitation across a spatial domain. The nature of local precipitation, including its variability and intermittency, compounds in the spatial case, making spatiotemporal precipitation simulation difficult. There are a wide variety of approaches for spatiotemporal precipitation modeling and simulation. Hidden Markov models have been used to model occurrence [*Hughes et al*., 1999] as well as intensity [*Ailliot et al*., 2009; *Charles et al*., 1999]. Other methods include nearest-neighbors resampling [*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 to normality [*Sansó and Guenni*, 2000; *Yang et al*., 2005], artificial neural network methods [*Cannon*, 2008], copula-based approaches [*Bárdossy and Pegram*, 2009], and multifractal rainfall models [*Gupta and Waymire*, 1993; *Lovejoy and Schertzer*, 1995; *Menabde et al*., 1997; *Deidda*, 2000; *Veneziano and Langousis*, 2005; *Langousis and Veneziano*, 2007; *Veneziano et al*., 2006; *Veneziano and Langousis*, 2010]. Current approaches to this problem typically seek the assistance of latent multivariate normals, sometimes including a transformation, to generate the occurrence/intensity values over a spatial domain. This approach was catalyzed by *Wilks* [1998], and advanced by *Brissette et al*. [2007] and *Thompson et al*. [2007]. The modern popularity of Wilks' approach can largely be attributed to a comparative study of his approach with the resampling and hidden Markov model approaches, which found the one of Wilks to best capture spatial dependence as well as local temporal dependence of precipitation [*Mehrotra and Sharma*, 2010].

For space-time data sets, observational data are often very high-dimensional: with even a handful of spatial locations and a few years of daily observations, the dimensionality can reach upward of tens of thousands of data points. Estimation schemes for model-based SWGs typically follow either a likelihood-based approach (whether frequentist or Bayesian), or a moment-matching approach. However, for modern data sets, likelihood-based methods are infeasible due to the dimensionality of the process; falling back on moment-based methods can be useful, but also implies a loss of statistical efficiency.

In this work, we explore a family of techniques known as Approximate Bayesian Computation (ABC) to approximate posterior densities of statistical model parameters. These algorithms yield accurately approximated posteriors while sidestepping the often problematic evaluation of analytic likelihood functions [*Sunnåker et al*., 2013]. In particular, we examine a daily spatiotemporal precipitation occurrence problem, and utilize ABC to estimate model parameters. Particular emphasis is put on the regularization function as it relates to scientifically meaningful statistics such as dry and wet spell lengths. We assess our approach on a large historical database of observed precipitation occurrence over the state of Iowa, and a case-study location in Colorado.

Successful ABC techniques for estimating SWGs of daily local maximum temperature were presented by *Olson* [2016], but presently, we are unaware of any other attempts to apply ABC for estimating SPGs. There has been some interest in using ABC for hydrological purposes, in comparing generalized likelihood uncertainty estimation (GLUE) with ABC in a case study of rainfall-runoff modeling [*Nott et al*., 2012]. Sadegh and Vrugt follow suit with some further exploration of ABC, again with rainfall-runoff applications in mind [*Sadegh and Vrugt*, 2013, 2014]. Otherwise, the utilization of ABC remains chiefly in the field of biology due to its origins in population genetics research. We contend that ABC constitutes a promising estimation toolset for creating reliable SWGs.

2. Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) refers to a family of techniques for approximating posterior densities of statistical parameters by bypassing direct likelihood evaluations. The first traces of ABC can be found in a 1984 essay by Donald Rubin, in which he puts forth a thought experiment to illustrate how Bayesian ideas should be interpreted and put into effect [*Rubin*, 1984; *Sunnåker et al*., 2013]. The first algorithm to incorporate ideas that comprise present-day ABC was proposed by *Tavaré et al*. [1997], concerning the chronology of common ancestors of humans based on homologous DNA sequence samples, and has since inspired much research in this discipline [*Fu and Li*, 1997; *Weiss and von Haeseler*, 1998; *Pritchard et al*., 1999; *Beaumont et al*., 2002].

ABC algorithms are designed to approximate posterior distributions of the form $f(\boldsymbol{\theta}|\mathbf{x})$ where $\boldsymbol{\theta}$ is a vector of model parameters, and $\mathbf{x}$ a vector of observations. Bayes' theorem yields the familiar relationship

$$f(\boldsymbol{\theta}|\mathbf{x}) \propto L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{1}$$

where $L(\mathbf{x}|\boldsymbol{\theta})$ is the likelihood function of $\mathbf{x}$ and $\pi(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$. Key difficulties arise when the likelihood L is either not available in closed form, or is computationally intractable to calculate numerically. In the following subsections, we describe two ABC algorithms that will be applicable to the ensuing precipitation problem.

2.1. ABC Acceptance-Rejection Schemes

In what follows, suppose that we are given observed data $\mathcal{D} \in \mathbb{R}^n$ which were generated by a statistical model $\mathcal{M}(\boldsymbol{\theta})$, parametrized by $\boldsymbol{\theta}=(\theta_1, \ldots, \theta_p)$. The key idea behind ABC is to generate a candidate parameter $\boldsymbol{\theta}' \sim \pi$, unconditionally generate a new data set $\mathcal{D}'$ according to $\mathcal{M}(\boldsymbol{\theta}')$, and if $\mathcal{D}'$ and $\mathcal{D}$ are "close" enough, view $\boldsymbol{\theta}'$ as a sample from the posterior distribution $\boldsymbol{\theta}' \sim f(\boldsymbol{\theta}|\mathcal{D})$.

To formalize the notion of "closeness," we introduce a similarity metric $\varrho : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^+$ that is negatively oriented, and such that $\varrho(\mathcal{D}, \mathcal{D})=0$. The ABC algorithm then follows by accepting $\boldsymbol{\theta}'$ if $\varrho(\mathcal{D}, \mathcal{D}') < \epsilon$ for some user-defined tolerance $\epsilon > 0$. Clearly the key scientific difficulty is to determine an appropriate $\varrho(\cdot, \cdot)$ and ϵ, both highly dependent on the scientific context and data. An important side note is that $\epsilon = 0$ yields samples from the true posterior, whereas allowing $\epsilon \rightarrow \infty$ yields samples from the prior. From a computational perspective, a higher ϵ will yield more accepted samples, but lose accuracy of posterior approximation, while a low ϵ will increase accuracy, but lose tractability. Empirical case studies demonstrate the effect of ϵ on the resultant posterior [*Sisson et al*., 2007], and theoretical bounds for parameter and posterior error estimates have been derived [*Dean et al*., 2011; *Fearnhead and Prangle*, 2011].

Typically, the dimensionality of $\mathcal{D}$ obfuscates the modeler's ability to choose a relevant ϱ, and so in practice we instead consider summary statistics $\mathbf{S}=\mathbf{s}(\mathcal{D})$ and $\mathbf{S}'=\mathbf{s}(\mathcal{D}')$. If $\mathbf{S}$ is a sufficient statistic for $\boldsymbol{\theta}$, then we have $f(\boldsymbol{\theta}|\mathcal{D})=f(\boldsymbol{\theta}|\mathbf{S})$ for all $\boldsymbol{\theta}$. Thus, in practice the metric $\varrho(\cdot, \cdot)$ is a function of these summary statistics. Even if the statistics are not sufficient, if they describe an important aspect of the data in the context of the problem, they should still act as acceptable summaries of the data. This gives rise to ABC Algorithm 1.

Algorithm 1: Generating samples distributed from $f(\boldsymbol{\theta}|\varrho(\mathbf{s}(\mathcal{D}), \mathbf{s}(\mathcal{D}')) < \epsilon)$

Input : Model $\mathcal{M}$, summary function $\mathbf{s}(\cdot)$, metric $\varrho(\cdot, \cdot)$, tolerance $\epsilon > 0$
desired number of samples N

Output: A vector V of samples distributed from $f(\boldsymbol{\theta}|\varrho(\mathbf{s}(\mathcal{D}), \mathbf{s}(\mathcal{D}')) < \epsilon)$

$\mathbf{S} \leftarrow \mathbf{s}(\mathcal{D})$

$V \leftarrow \{\}$

while $|V| < N$ **do**

  generate $\boldsymbol{\theta} \sim \pi(\cdot)$

  simulate $\mathcal{D}'$ from $\mathcal{M}$ parametrized by $\boldsymbol{\theta}$

  $\mathbf{S}' \leftarrow \mathbf{s}(\mathcal{D}')$

  if $\varrho(\mathbf{S}, \mathbf{S}') < \epsilon$ **then**
    append $\boldsymbol{\theta}$ to V

  end

end

return V

Algorithm 1 has its shortcomings, however. For example, if the prior distribution contains large intervals of values from which to sample parameters, the chances of accepting can be extremely low. This effect is compounded heavily as the number of statistical parameters increases. Additionally, the acceptance-rejection algorithm generates independent samples with each step, so each iteration relies on identifying a desirable $\boldsymbol{\theta}$ from $\pi(\cdot)$ without any other information.

2.2. ABC Using Markov Chain Monte Carlo

The operational issues of Algorithm 1 can be rectified by introduction of the Metropolis-Hastings (M-H) algorithm. M-H is a standard Markov chain Monte Carlo method used for distribution sampling [*Metropolis et al.*, 1953; *Hastings*, 1970]. In particular, M-H is a guided search through parameter space that uses a current value $\boldsymbol{\theta}$ to generate a candidate value $\boldsymbol{\theta}'$ according to a candidate-generating density $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ [*Chib and Greenberg*, 1995]. In this way, $\boldsymbol{\theta}'$ typically has a higher chance of acceptance than an unconditionally generated proposal. The ABC-MCMC algorithm is laid out in Algorithm 2. This will be our default variant of ABC, excepting the case of determining an initial guess where the rejection version is better suited.

As in standard M-H applications, Algorithm 2 shows that if $q(\cdot, \cdot)$ is symmetric, it drops out of the algorithm. In addition, if we sample from a uniform density for each parameter, $\theta_i \sim \mathcal{U}(a_i, b_i)$ for $i=1, \ldots, p$, then the condition $u < \min\{\pi(\boldsymbol{\theta}')/\pi(\boldsymbol{\theta}), 1\}$ reduces to simply checking that each parameter lies in the prior. That is, accept $\boldsymbol{\theta}' = (\theta_1', \ldots, \theta_n')$ only if $\theta_i' \in [a_i, b_i]$ for all i. In this case, the main filter of parameter acceptances is the ABC criterion, $\varrho(\mathcal{D}, \mathcal{D}') < \epsilon$.

Algorithm 2: Generating samples distributed from $f(\boldsymbol{\theta} | \varrho(\mathcal{D}, \mathcal{D}') < \epsilon)$ via ABC-MCMC

```
Input : Observed data D, arbitrary initial value θ_0, candidate-generating density q(·,·),
        model M, , metric ϱ(·,·), tolerance ϵ > 0, desired number of samples N
Output: A vector V of samples approximately distributed from f(θ|ϱ(D, D') < ϵ)
V ← {}
θ ← θ_0
while |V| < N do
    generate θ' ~ q(θ, ·)
    simulate D' from M parametrized by θ'
    if ϱ(D, D') < ϵ then
        u ~ U(0, 1)
        if u < min{ π(θ')q(θ',θ) / π(θ)q(θ,θ') , 1 } then
            θ ← θ'
        end
    end
    append θ to V
end
return V
```

It is important to note that Algorithm 2 is extremely sensitive to the initial guess $\boldsymbol{\theta}_0$ due to the relatively small jumping window in comparison to the prior. There are a couple of ways to sidestep this issue. For example, one can perform a grid search over $\boldsymbol{\theta}$ to approximate a minimizing value $\boldsymbol{\theta}_0 \approx \arg\min_{\boldsymbol{\theta}} \varrho(\mathcal{D}, \mathcal{D}'(\boldsymbol{\theta}))$ for our initial guess. Caution must be taken, however, since ϱ is a function of stochastic input, and therefore a suitable minimizing value may be difficult to ascertain. Another remedy is to start with Algorithm 1 until an accepted parameter has been found, and then proceed with Algorithm 2.

3. Statistical Spatiotemporal Precipitation Occurrence Model

We turn now to the spatiotemporal precipitation generator. We adapt the statistical model of *Kleiber et al.* [2012] to simulate spatially and temporally correlated binary values, relying on a latent Gaussian process. Let $Y(\mathbf{s},d)$ denote the precipitation occurrence at site $\mathbf{s}\in\mathbb{R}^2$ on day $d\in\mathbb{Z}^+$, so that $Y(\mathbf{s},d)=1$ for positive precipitation and 0 otherwise. Precipitation is driven by a latent Gaussian process $W(\mathbf{s},d)$ with mean function $\mu(\mathbf{s},d;\boldsymbol{\beta})$ and covariance function $C(\mathbf{h},d;\boldsymbol{\alpha})+\tau^2\mathbf{1}_{[\mathbf{h}=\mathbf{0}]}(\mathbf{h})$, where $\mathbf{1}$ is the indicator function, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are model parameter vectors and $\mathbf{h}\in\mathbb{R}^2$ is a spatial lag vector. In the covariance function, C represents continuous spatial variation while the nugget effect τ^2 accounts for microscale variation. The distribution of Y is specified through

$$Y(\mathbf{s},d)=\mathbf{1}_{[W(\mathbf{s},d)\geq 0]}, \tag{2}$$

which implies that $Y(\mathbf{s},d)$ is locally a probit regression.

Our choice for the mean function includes autoregressive and harmonic terms to capture temporal persistence and seasonal variations. In particular, we set

$$\begin{aligned}\mu(\mathbf{s},d;\boldsymbol{\beta})=\beta_0+\beta_1Y(\mathbf{s},d-1)+\beta_2\cos\left(\frac{2\pi d}{365}\right)+\beta_3\sin\left(\frac{2\pi d}{365}\right)\\ +\beta_4\cos\left(\frac{4\pi d}{365}\right)+\beta_5\sin\left(\frac{4\pi d}{365}\right)\end{aligned}, \tag{3}$$

where $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_5)^\top$. For the covariance structure, we assume an isotropic exponential covariance with a time-dependent range,

$$C(\mathbf{h},d;\boldsymbol{\alpha})=\exp\left(-\frac{||\mathbf{h}||}{A(d)}\right),$$

where $||\cdot||$ is the Euclidean norm. We elect

$$A(d)=\exp\left\{\alpha_0+\alpha_1\cos\left(\frac{2\pi d}{365}\right)+\alpha_2\sin\left(\frac{2\pi d}{365}\right)\right\}, \tag{4}$$

where $\boldsymbol{\alpha}=(\alpha_0,\alpha_1,\alpha_2)^\top$. The harmonics allow for seasonally dependent range (e.g., in the midwest US precipitation tends to have shorter length-scales during summer, being driven by locally convective storms, whereas length-scales are longer during winter due to frontal passages).

Thus our general spatiotemporal model is parameterized by $\boldsymbol{\theta}=\left(\boldsymbol{\beta}^\top,\boldsymbol{\alpha}^\top,\tau^2\right)^\top$. To illustrate the roles of the ABC metrics, we will explore two case studies of our model to which our metrics will be tailored. We begin with precipitation occurrence at a single site, and then generalize our methodology to a general spatial domain.

4. Single-Site Occurrence

An attractive feature of model (2) is that it reduces to a probit regression in the case of a single location $\mathbf{s}$. In particular, the only statistical parameters for the single-site generator are $\boldsymbol{\beta}$. While ABC bypasses the need to evaluate a likelihood function, for this single-site setup we can actually deduce a closed-form expression of the likelihood for our model, allowing us to check our posterior approximations from the ABC algorithm. We will assume uniform, independent marginal priors for β_i, i.e., $\beta_i\sim U(a_i,b_i), i=0,\ldots,5$. Then the log-posterior is

$$\begin{aligned}\log f(\boldsymbol{\beta}|\mathbf{o})=\sum_{d=2}^{T}\left\{\mathbf{1}_{[o_d=1]}\log\left(\Phi\left(\boldsymbol{\beta}^\top\mathbf{X}\right)\right)+\mathbf{1}_{[o_d=0]}\log\left(1-\Phi\left(\boldsymbol{\beta}^\top\mathbf{X}\right)\right)\right\}\\ +\sum_{i=0}^{5}\left\{\log\mathbf{1}_{[a_i\leq\beta_i\leq b_i]}-\log\left(b_i-a_i\right)\right\}+c\end{aligned}, \tag{5}$$

where c is a normalizing constant, $\mathbf{o}=(o_1,o_2,\ldots,o_T)^\top$ is the vector of observed occurrences $o_i=Y(\mathbf{s},i)$, and $\mathbf{X}$ is the design matrix for covariates (3). The derivation of $\log f$ is provided in Appendix A.

4.1. ABC Metrics for Single-Site Occurrence

For single-site precipitation occurrence, a natural similarity metric will be tailored to precipitation's temporal persistence. Heuristically, any simulated sequence of occurrences generated by the probit model should have similar short and long scales of temporal persistence as exhibited in the observational record. Thus, our metric is tailored to capture wet and dry spells of varying lengths.

Let ζ be a set of numbers that correspond to spell count lengths. Define $\varphi(\mathbf{x};\zeta)$ to be the function that computes the number of wet spells of length in ζ present in the time series $\mathbf{x}=(x_1,\ldots,x_T)$, and define $\psi(\mathbf{x};\zeta)$ to be the function that computes the number of dry spells of length in ζ present in $\mathbf{x}$. Our estimator $\widehat{\boldsymbol{\beta}}$ should adhere to the following criterion:

$$\widehat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\left\{\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left(\frac{c_w}{|\mathcal{Z}_1|}\sum_{\zeta_1\in\mathcal{Z}_1}\frac{|\varphi(\mathbf{s}_{\boldsymbol{\beta}}[t];\zeta_1)-\varphi(\mathbf{o}[t];\zeta_1)|}{\max\{\varphi(\mathbf{o}[t];\zeta_1),1\}}+\frac{c_h}{|\mathcal{Z}_2|}\sum_{\zeta_2\in\mathcal{Z}_2}\frac{|\psi(\mathbf{s}_{\boldsymbol{\beta}}[t];\zeta_2)-\psi(\mathbf{o}[t];\zeta_2)|}{\max\{\psi(\mathbf{o}[t];\zeta_2),1\}}\right)\right\}. \quad (6)$$

Here $\mathcal{T}$ is a partition of the months of the year, $\mathcal{Z}_1$ is a set of sets of wet spell lengths (e.g., $\{\{1,2,3\},\{4,5,6\},\{7,8,9\}\}$), $\mathcal{Z}_2$ is a set of sets of dry spell lengths, $\mathbf{o}$ is the observed occurrence time series, $\mathbf{s}_{\boldsymbol{\beta}}$ is a simulation of our model parameterized by $\boldsymbol{\beta}$, and the notation $\mathbf{x}[t]$ represents the time series $\mathbf{x}$ subsetted by the set of time points t. We would like to scale by the number of dry/wet spells in the observations, but it is possible that the particular isolated sequence $\mathbf{o}[t]$ does not have any of the given amount ζ_i, so we divide by $\max\{\varphi(\mathbf{o}[t];\zeta_i),1\}$ to ensure there is no division by zero. Furthermore, we allow $\mathcal{Z}_1$ and $\mathcal{Z}_2$ to be sets of sets to retain generality as well as allow for a faster algorithmic runtime in practice.

With this formulation of our estimator $\widehat{\boldsymbol{\beta}}$, we are in position to formalize our metric to estimate local precipitation occurrence for a site $\mathbf{s}$:

$$\varrho_{\text{Local}}(\mathcal{D},\mathcal{D}'):=\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left(\frac{1}{|\mathcal{Z}_1|}\sum_{\zeta_1\in\mathcal{Z}_1}\frac{|\varphi(\mathcal{D}[t];\zeta_1)-\varphi(\mathcal{D}'[t];\zeta_1)|}{\max\{\varphi(\mathcal{D}[t];\zeta_1),1\}}+\frac{1}{|\mathcal{Z}_2|}\sum_{\zeta_2\in\mathcal{Z}_2}\frac{|\psi(\mathcal{D}[t];\zeta_2)-\psi(\mathcal{D}'[t];\zeta_2)|}{\max\{\psi(\mathcal{D}[t];\zeta_2),1\}}\right). \quad (7)$$

While our methods are defined using specific functions, it should be noted that this procedure can be abstracted. For example, if the modeler were more interested in the total number of rainy days per month, one could replace our functions φ and ψ with a function that counts the number of days where precipitation occurred, and then use the relative error per month as the metric for the ABC criterion.

4.2. Numerical Results for Bonny Dam, Colorado

To validate the proposed approach, we examine Bonny Dam, Colorado, for which precipitation data are available in the Global Historical Climatology Network Database (GHCND) [*Peterson and Vose*, 1997]. Specifically, we use records for Bonny Dam from 1 June 1949 to 11 September 2011, yielding 21,911 days of actual precipitation recordings and 822 days of missing data values. Our uniform priors range between -1.5 and 1.5 for the intercept β_0, between 0 and 1 for the AR term β_1, and the remaining priors range from -0.5 to 0.5. We iterate through three choices of ϵ: $\epsilon_1=1.09$, $\epsilon_2=1.18$, and $\epsilon_3=1.27$, which yield acceptance rates of 23.1%, 37.7%, and 58.9%. Note that ϵ_1 was chosen to induce an acceptance rate close to the theoretical optimum of 23% [*Gelman et al.*, 1996]. In each case, we obtain $\boldsymbol{\beta}_0$ from a grid search based on our ϱ criterion. Moreover, we define our sets of spell lengths for our ϱ function as

$$\mathcal{Z}_1=\{\{1\},\{2,3\},\{4,5\},\{6,7\},\{8,9\},\{10,11\},\{12,13,\ldots\}\}, \quad (8)$$

for our wet spells and

$$\mathcal{Z}_2=\{\{1,2,3,4\},\{5,6,7,8\},\{9,10,11,12\},\{13,14,15,16\},\{17,18,19,20\},\{21,22,23,24\},\{25,26,27,28,29\},\{30,31,\ldots\}\}, \quad (9)$$

for our dry spells. These sets were constructed so that they span the range of realistic values well, highlight important lengths to be matched, and decrease algorithmic runtime. For example, when examining the observed spell counts, we see that the number of wet spells of length one is much larger in comparison to those of higher lengths, whereas the number of spells of length greater than 11 is relatively small. So, we keep {1} separate, bin spells of lengths of 12 and higher all in the same set, and group the rest in pairs for computational benefits. We encourage the modeler to adjust these sets to the problem at hand and ruminate on whether granularity or tractability is a higher priority. With these settings, we collect 10,000 ABC samples, after omitting a standard burn-in based on examining convergence of the trace plot of the sampling chain.

We compare priors, true posteriors, and ABC-approximated posteriors for each ϵ_i in Figure 1. Here we see, qualitatively, how closely the ABC posteriors align with the truth when ϵ yields a small enough acceptance rate, and how increasing our tuning parameter can lead to wider posteriors and straying MCMC chains. Nonetheless, all of the ABC posteriors are still significantly narrowed with respect to the priors. Figure 2 displays a table of the first three-empirical moments for the true and ABC posteriors, as well as the theoretical moments for the priors, for each choice of ϵ. We additionally compute and plot the mean functions $\mu(\mathbf{s}, d; \widehat{\boldsymbol{\beta}})$ using the empirical mean of each marginal posterior density $\widehat{\beta}_i$, and compare to the empirical probability of rain based on the observed data, seen in Figure 3. Indeed, both the true posterior and ABC mean functions present a credible description of the data's trend. Note that the probability of precipitation is not a feature of our metric ϱ, yet empirically we match the general observed pattern.

Moreover, we can examine statistics of sample trajectories from our ABC posteriors (with tuning parameter $\epsilon_1 = 1.09$) and true posteriors given by (5). First, we look at empirical standard deviation of our observations, 100 sample trajectories from our ABC posterior, and 100 sample trajectories from true posterior, for each of the 365 days. Box plots of these daily empirical standard deviations, binned by month, are shown in Figure 4. For example, for January, the box plot shown for the observations represents the distribution of $\widehat{\text{SD}}$(1 January), . . ., $\widehat{\text{SD}}$(31 January). For the ABC and true posterior box plots, each daily $\widehat{\text{SD}}(\cdot)$ is taken across all 100 trajectories, so that $\widehat{\text{SD}}$(1 January) is computed over all of the 100 empirical 1 January values. These statistics match well despite the absence of standard deviation comparisons in our penalization criterion ϱ. Of course, we are also interested in the correspondence of wet and dry spell counts, the crux of our ABC metric, for the observations and ABC simulations. Using the marginal ABC posterior means to estimate the parameters $\boldsymbol{\beta}$ and simulate a sample trajectory, we compute these spell counts for both time series and display them in Figures 5 and 6, respectively. Both wet and dry counts are well reproduced by our parameter estimates, at both short and long durations. Overall, the simulations produced by our ABC procedure are highly similar to the observations as seen by each summary statistic we tested.

5. Multisite Occurrence

We generalize our estimation approach for spatial occurrence data. Generally, estimation of mean function parameters $\boldsymbol{\beta}$ will proceed as in the previous section, but we will need a tool to estimate the spatial dependence between sites as controlled by the parameters $\boldsymbol{\alpha}$. For this, we propose a clustered variant of the classical empirical variogram, which allows for an effective ABC metric to estimate $\boldsymbol{\alpha}$.

5.1. Assessing Spatial Dependence: Variograms

A variogram γ is a function that quantifies the degree of spatial dependence of a stochastic process [*Cressie*, 1993; *Diggle et al.*, 1998]. A nonparametric empirical variogram is given by

$$\widehat{\gamma(\mathbf{u})} = \frac{1}{2|N(\mathbf{u})|} \sum_{i=1}^{n} \sum_{j \neq i} \left[y(\mathbf{x}_i) - y(\mathbf{x}_j) \right]^2, \tag{10}$$

for spatial data $y(\mathbf{x})$ at locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $N(\mathbf{u})$ is a user-defined neighborhood of location pairs $||\mathbf{x}_i - \mathbf{x}_j|| \approx ||\mathbf{u}||$ about $\mathbf{u}$, and $|\cdot|$ denotes the number of pairs in the neighborhood. Raw variograms (without averaging) tend to be extremely noisy; the binning is just a smoothing operation. We follow the standard rule of thumb by including at least 30 pairs of observations per bin. Moreover, for $|\mathbf{u}| \gg 0$, so few pairs are used in $\widehat{\gamma}(\mathbf{u})$ that we restrict distances so that $\mathbf{u} \leq \frac{1}{2} \max \|\mathbf{x}_i - \mathbf{x}_j\|$. Our interest in the variogram stems from a wealth of experience in the geostatistical community regarding using variogram-based metrics as

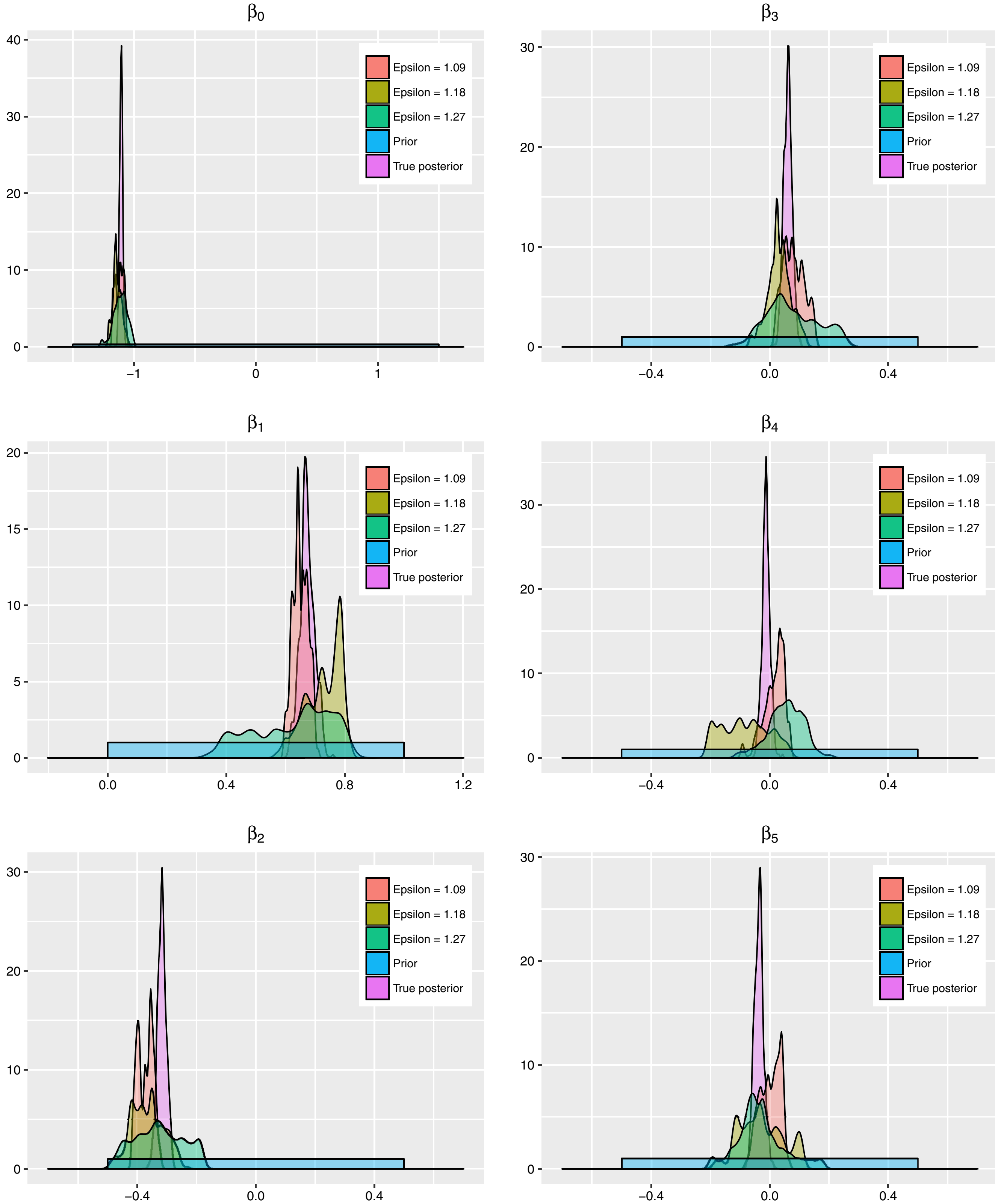

Figure 1. The marginal prior, true posterior (via MCMC sampling), and approximate posterior densities (for several values of ϵ) for $\boldsymbol{\beta}$.

posteriors with their corresponding ϵ values.

i	Moment	$\widehat{\beta}_{i,\text{Prior}}$	$\widehat{\beta}_{i,\text{Posterior}}$	$\widehat{\beta}_{i,\text{ABC}}(\epsilon=1.09)$	$\widehat{\beta}_{i,\text{ABC}}(\epsilon=1.18)$	$\widehat{\beta}_{i,\text{ABC}}(\epsilon=1.27)$
0	$\widehat{\mu_1}'$	0	-1.11	-1.12	-1.14	-1.11
	$\widehat{\mu_2}'$	0.750	1.22	1.27	1.30	1.25
	$\widehat{\mu_3}'$	0	-1.35	-1.43	-1.49	-1.40
1	$\widehat{\mu_1}'$	0.500	0.678	0.652	0.726	0.608
	$\widehat{\mu_2}'$	0.333	0.460	0.426	0.531	0.387
	$\widehat{\mu_3}'$	0.250	0.313	0.279	0.390	0.256
2	$\widehat{\mu_1}'$	0	-0.315	-0.377	-0.381	-0.339
	$\widehat{\mu_2}'$	0.0833	0.0993	0.143	0.147	0.122
	$\widehat{\mu_3}'$	0	-0.0314	-0.0543	-0.0577	-0.0457
3	$\widehat{\mu_1}'$	0	0.0637	0.0879	0.0322	0.0951
	$\widehat{\mu_2}'$	0.0833	0.00425	0.00880	0.00224	0.0157
	$\widehat{\mu_3}'$	0	0.000295	0.000968	0.000144	0.00296
4	$\widehat{\mu_1}'$	0	-0.0133	0.0180	-0.100	0.0620
	$\widehat{\mu_2}'$	0.0833	0.000430	0.00145	0.0153	0.00721
	$\widehat{\mu_3}'$	0	-0.0000122	0.0000217	-0.00244	0.000772
5	$\widehat{\mu_1}'$	0	-0.0307	-0.00288	-0.0237	-0.0278
	$\widehat{\mu_2}'$	0.0833	0.00122	0.00113	0.00663	0.00730
	$\widehat{\mu_3}'$	0	-0.0000555	-0.0000189	-0.000433	-0.000324

Figure 2. First three raw moments of the marginal priors, true posteriors, and three ABC posteriors with their corresponding ϵ values.

robust estimators of space-time dependence [*Cressie and Zimmerman*, 1992]. Indeed, variogram-based estimators of spatial dependence have been shown to perform nearly as well as likelihood-based procedures [*Zimmerman and Zimmerman*, 1991].

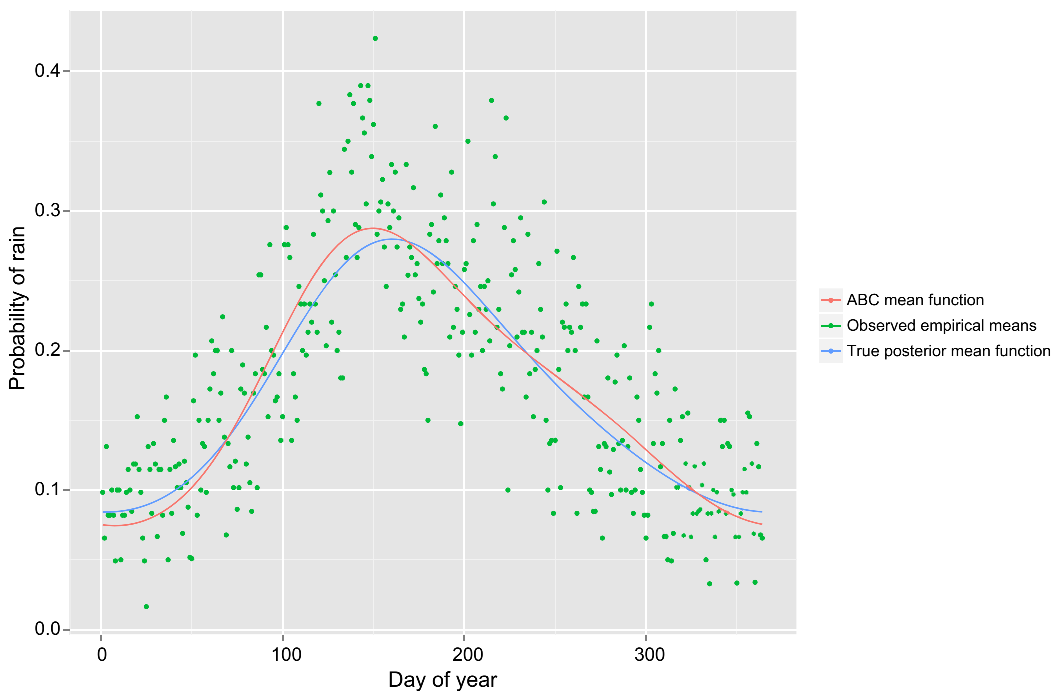

Figure 3. The mean function of precipitation occurrence for the ABC and true posterior estimates as well as the observed empirical probability of precipitation for Bonny Dam, Colorado.

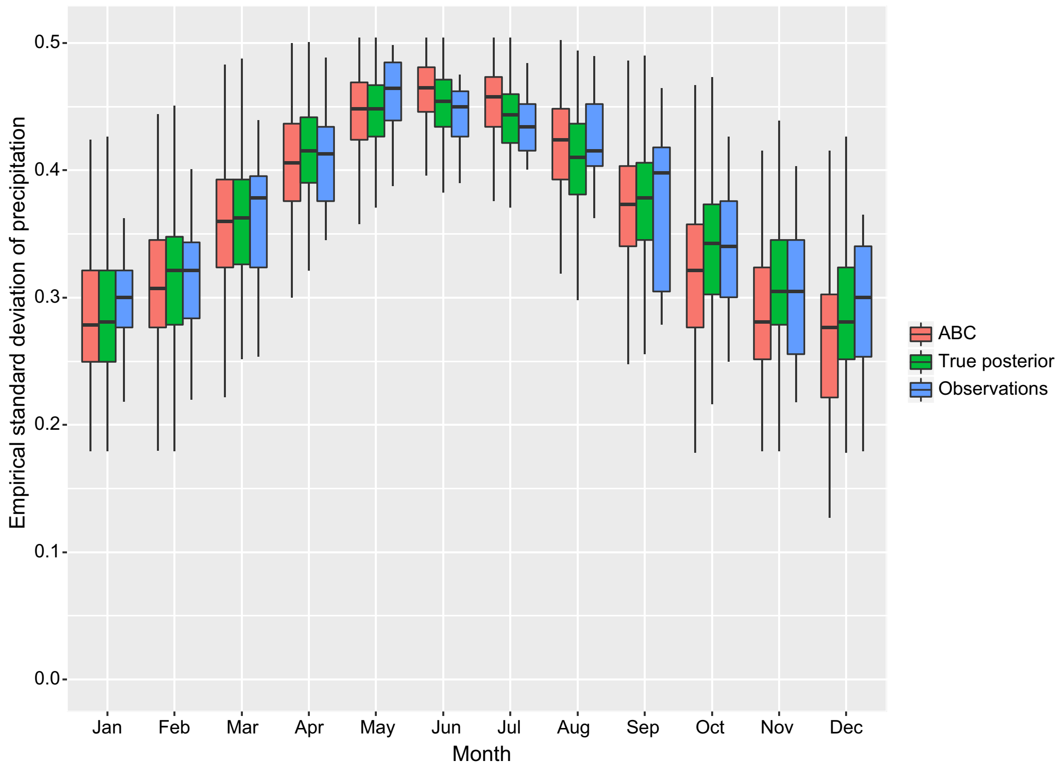

Figure 4. Daily empirical standard deviations of precipitation (in mm) for ABC simulations, true posterior simulations, and observations for Bonny Dam, Colorado, binned by month.

As our latent Gaussian spatial process is assumed to have seasonally dependent correlation lengths, we require empirical variograms that evolve throughout the year, e.g., $\widehat{\gamma(\mathbf{u})}=\widehat{\gamma(\mathbf{u},d)}$. We seek to choose enough days in the year to get a stable estimator, while still maintaining tractability for ABC. Thus, we construct a monthly aggregate variogram $\widehat{\Gamma}$ as a function of month m as well as $\mathbf{u}$, defined as

$$\widehat{\Gamma}(m,\mathbf{u})=\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}}\frac{1}{2\omega}\sum_{d=\delta_{m,y}-\omega}^{\delta_{m,y}+\omega}\widehat{\gamma}(\mathbf{u},d), \tag{11}$$

where $\mathcal{Y}$ is the set of sample years, $\delta_{m,y}$ is the sample center day of the month m in year y, and ω is the forward and backward offset for binning. For each month m, we loop through all sample years and select a sample center day for that month, $\delta_{m,\cdot}$, which is arbitrarily chosen, but consistent. Next, we compute a clustered variogram for δ_m in that we use only a few days before and after $\delta_{m,\cdot}$ to smooth the variogram data from an otherwise noisy set of points. This manifests as $\delta_{m,y}\pm\omega$, where we set $\omega=5$ which yields a group of 11 consecutive days for each month. Finally, we divide by the total number of aggregated variograms, $2\omega|\mathcal{Y}|$, to ensure that the values of $\widehat{\Gamma}$ are on the same scale as $\widehat{\gamma}$.

5.2. ABC Metrics for Multisite Occurrence

For our full spatiotemporal precipitation occurrence generator, we have $\boldsymbol{\theta}=\left(\boldsymbol{\beta}^\top,\boldsymbol{\alpha}^\top,\tau^2\right)^\top$, totalling 10 statistical parameters. Rather than simultaneously estimating all parameters, we separate out estimation of mean function parameters from spatial-dependence parameters, as is common in geostatistical modeling.

To estimate $\boldsymbol{\beta}$, we require a statistic that measures accuracy of the mean function over the entire domain. Assuming the terrain does not vary greatly and that precipitation behaves locally similarly at each

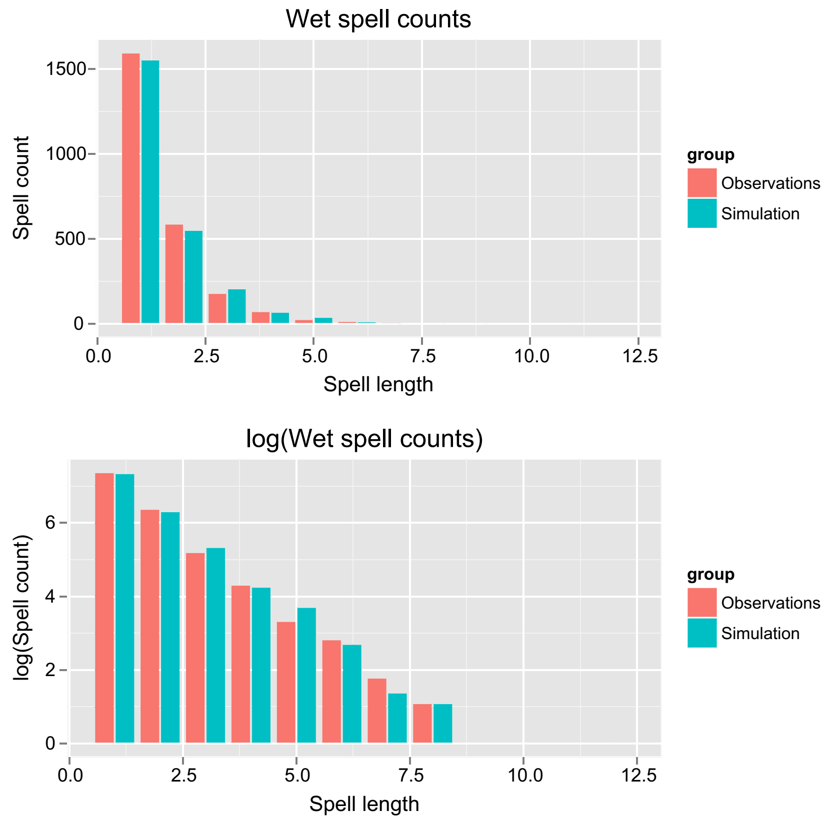

Figure 5. Wet spell counts and the logarithm of wet spell counts by spell length for the observations and ABC simulation for Bonny Dam, Colorado.

observation location $\mathbf{s}_i$, we consider an aggregation of our ϱ_{Local} metric as defined in (7). That is, we compute a standardized sum of the wet and dry spell count errors over all of our locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$,

$$\begin{aligned}\varrho_{\boldsymbol{\beta}}(\mathcal{D}, \mathcal{D}') &= \frac{1}{n}\sum_{i=0}^{n} \varrho_{\text{Local}}(\mathcal{D}[\mathbf{s}_i], \mathcal{D}'[\mathbf{s}_i]) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\left(\frac{1}{|\mathcal{Z}_1|}\sum_{\zeta_1\in\mathcal{Z}_1}\frac{|\varphi(\mathcal{D}[\mathbf{s}_i,t];\zeta_1)-\varphi(\mathcal{D}'[\mathbf{s}_i,t];\zeta_1)|}{\max\{\varphi(\mathcal{D}[\mathbf{s}_i[t];\zeta_1),1\}}\right., \\ &\quad \left. +\frac{1}{|\mathcal{Z}_2|}\sum_{\zeta_2\in\mathcal{Z}_2}\frac{|\psi(\mathcal{D}[\mathbf{s}_i,t];\zeta_2)-\psi(\mathcal{D}'[\mathbf{s}_i,t];\zeta_2)|}{\max\{\psi(\mathcal{D}[\mathbf{s}_i,t];\zeta_2),1\}}\right)\end{aligned} \tag{12}$$

where φ counts wet spells as in (6), ψ counts dry spells as in (6), $\mathcal{T}$ is a partition of the months of the year, $\mathcal{Z}_1$ is a set of sets of wet spell lengths, $\mathcal{Z}_2$ is a set of sets of dry spell lengths, $\mathcal{D}[\mathbf{s}_i]$ corresponds to the subset of the data at location $\mathbf{s}_i$, and $\mathcal{D}[\mathbf{s}_i, t]$ corresponds to the subset of the data at location $\mathbf{s}_i$ and time points t.

In a Gibbs sampling framework, we condition on an ABC estimate of $\boldsymbol{\beta}$, and estimate remaining spatial parameters $\gamma=(\boldsymbol{\alpha}^\top, \tau^2)^\top$ in an ABC-MCMC algorithm. To this end, we use our aggregate monthly variogram of (11) within our covariance similarity metric,

$$\varrho_\gamma(\mathcal{D}, \mathcal{D}') = \frac{1}{12}\sum_{m=1}^{12}\frac{|\widehat{\Gamma}(m, \mathcal{D}) - \widehat{\Gamma}(m, \mathcal{D}')|}{\widehat{\Gamma}(m, \mathcal{D})}, \tag{13}$$

where the sum is over the 12 months of the year. The combination of these metrics will constitute our ABC-MCMC algorithm for spatial precipitation occurrence, shown in Algorithm 3.

Algorithm 3: ABC-MCMC for daily spatiotemporal precipitation occurrence parameter estimation

Input : Observed data $\mathcal{D}$, initial values $\boldsymbol{\beta}_{\text{Initial}}$ and $\boldsymbol{\gamma}_{\text{Initial}}$, candidate-generating densities $q_1(\cdot,\cdot), q_2(\cdot,\cdot)$, tolerances $\epsilon_1, \epsilon_2 > 0$, desired number of samples N_1 for $\boldsymbol{\beta}$ and N_2 for $\boldsymbol{\gamma}$
Output: A vector V of samples approximately distributed from target distribution $f(\boldsymbol{\theta}|\mathcal{D})$

Define $\mathcal{M}$ using (2)
Define $\varrho_{\boldsymbol{\beta}}$ using (12)
Define $\varrho_{\boldsymbol{\gamma}}$ using (13)
$V_{\boldsymbol{\beta}} \leftarrow \{\}$
$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_{\text{Initial}}$
while $|V_{\boldsymbol{\beta}}| < N_1$ **do**
    generate $\boldsymbol{\beta}_{\text{candidate}} \sim q_1(\boldsymbol{\beta},\cdot)$
    simulate $\mathcal{D}'(\boldsymbol{\beta}_{\text{candidate}})$ from $\mathcal{M}$
    if $\varrho_{\boldsymbol{\beta}}\left(\mathcal{D}, \mathcal{D}'(\boldsymbol{\beta}_{candidate})\right) < \epsilon_1$ **then**
        generate $u \sim \mathcal{U}(0,1)$
        if $u < \min\left\{\frac{\pi(\boldsymbol{\beta}_{candidate})q_1(\boldsymbol{\beta}_{candidate},\boldsymbol{\beta})}{\pi(\boldsymbol{\beta})q_1(\boldsymbol{\beta},\boldsymbol{\beta}_{candidate})}, 1\right\}$ **then**
            $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta}_{\text{candidate}}$
        end
    end
    append $\boldsymbol{\beta}$ to $V_{\boldsymbol{\beta}}$
end
$V_{\boldsymbol{\gamma}} \leftarrow \{\}$
$\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}_{\text{Initial}}$
while $|V_{\boldsymbol{\gamma}}| < N_2$ **do**
    generate $\boldsymbol{\beta} \sim V_{\boldsymbol{\beta}}$
    generate $\boldsymbol{\gamma}_{\text{candidate}} \sim q_2(\boldsymbol{\gamma},\cdot)$
    simulate $\mathcal{D}'(\boldsymbol{\beta},\boldsymbol{\gamma}_{\text{candidate}})$ from $\mathcal{M}$
    if $\varrho_{\boldsymbol{\gamma}}(\mathcal{D}, \mathcal{D}'(\boldsymbol{\beta},\boldsymbol{\gamma}_{candidate})) < \epsilon_2$ **then**
        generate $u \sim \mathcal{U}(0,1)$
        if $u < \min\left\{\frac{\pi(\boldsymbol{\gamma}_{candidate})q_2(\boldsymbol{\gamma}_{candidate},\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma})q_2(\boldsymbol{\gamma},\boldsymbol{\gamma}_{candidate})}, 1\right\}$ **then**
            $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}_{\text{candidate}}$
        end
    end
    append $\boldsymbol{\gamma}$ to $V_{\boldsymbol{\gamma}}$
end

5.3. Simulation Study

To verify that our metric (13) can adequately capture spatial dependence of precipitation occurrence in a domain, we can conduct a simulation study to demonstrate that ABC recovers the true model parameters. We simulate a mean-zero GP with fixed covariance parameters with $\boldsymbol{\beta}=\mathbf{0}$ and $\boldsymbol{\alpha}, \tau$ fixed and known, and apply (2) to get the "observed" field. In particular, we choose $(\boldsymbol{\alpha}^{\mathsf{T}}, \tau)^{\mathsf{T}}=(6.2, 0.6, -1.5, 0.3)^{\mathsf{T}}$, with priors $\mathcal{U}(4, 7.3)$, $\mathcal{U}(-2, 2)$, $\mathcal{U}(-2, 2)$, and $\mathcal{U}(0.2, 0.6)$ for $\alpha_1, \alpha_2, \alpha_3$, and τ, respectively. For simplicity, we start our MCMC chain at the marginal midpoints of the prior. We select $\epsilon=1.0$, and collect 10,000 samples after discarding a burn-in of 2000. The priors and resultant ABC posteriors are shown in Figure 7, along with lines indicating the true parameter values, ABC posterior means, and ABC posterior medians.

The true values are captured well within central 95% credible intervals for each posterior density. Moreover, the means and medians align closely with the truth, especially considering the width of the priors. This suggests that the ABC metric (13) robustly captures spatial dependence in a thresholded mean-zero Gaussian process.

5.4. Numerical Results for the State of Iowa

We illustrate our approach on an observational data set from the GHCND over the state of Iowa, USA. The data set contains precipitation occurrence values for 22 different spatial locations from 1 January 1893 to 31 December 2009. In this range, there are 883,736 available (location, day) recordings and 55,774 (location, day) pairs with missing values. Nominally, estimation of our model parameters would then require evaluating a likelihood function over these 883,736 data points, which is essentially computationally impossible. For example, *Kleiber et al*. [2012] examined the same data set, but were forced to estimate parameters using moment-based estimators, which do not as readily quantify uncertainty as in a Bayesian framework.

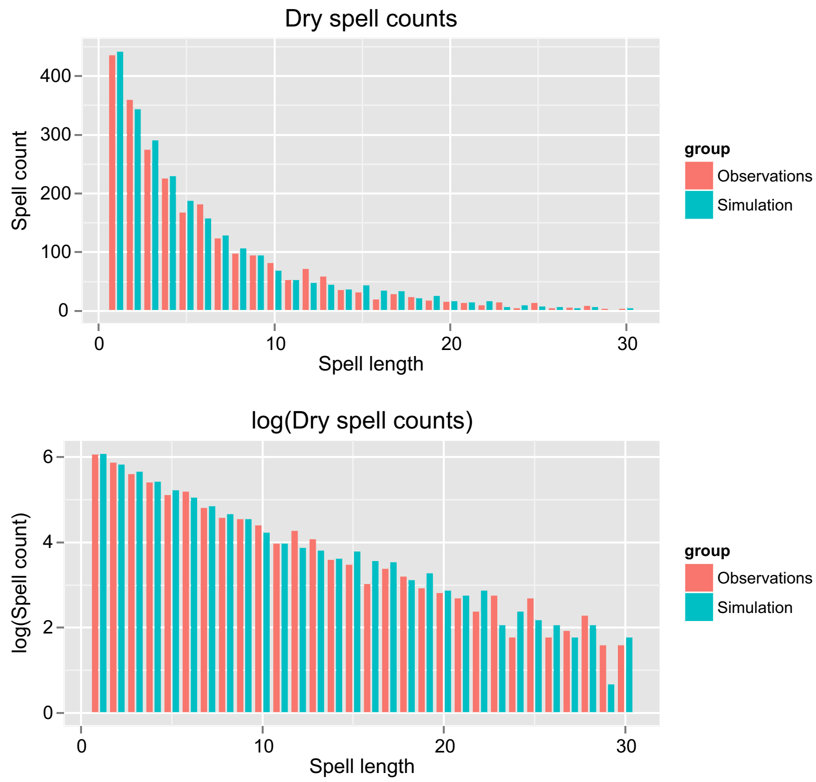

Figure 6. Dry spell counts and the logarithm of dry spell counts by spell length for the observations and ABC simulation for Bonny Dam, Colorado.

For the ABC-estimation of the mean function parameters $\boldsymbol{\beta}$, we choose a 20 year sequence of days for tractability, ranging from 1 January 1975 to 31 December 1995, which produces 156,135 actual recordings and only 4465 missing values. We reuse our wet and dry spell sets $\mathcal{Z}_1$ and $\mathcal{Z}_2$ from the previous section. For our tolerance ϵ_1, we first consider the level used for a single location, $\epsilon=1.09$, increase it slightly due to the variability throughout the domain, and then multiply it by the number of locations $n=22$, so that $\epsilon_1=1.15\times22=25.3$; this yields an acceptance rate of about 25.6%. Moreover, we conduct a grid search through the prior space to get an approximate minimizing value of ϱ, which we assign to be our initial guess $\boldsymbol{\beta}_0$. We elect for independent normal priors, centered around the initial guess $\boldsymbol{\beta}_0$ with variances of 0.1.

Figure 8 shows ABC posteriors after 4000 samples after discarding a burn-in of length 1000. We see considerable sharpness in the posteriors, confirmed by a plot of the expected probability of precipitation against the 22 empirical probabilities, as seen in Figure 9. As the means of the priors and ABC posteriors align incredibly well, we assessed the robustness of the posteriors against choice of prior by performing a mean shift of the priors. Indeed, similar ABC posteriors to the first case were recovered. Since Iowa was chosen due to its relatively homogeneous terrain, the behavior of the domain as a whole is replicated quite well. To estimate γ, we the same 20 year subset of the data, and perform a grid search for an initial guess γ_0. We select $\epsilon_2=1.0$ which induces an acceptance rate of 25.2%, and employ the following uniform priors: $\mathcal{U}(4,7.3)$ for α_0, $\mathcal{U}(-2,2)$ for α_1, $\mathcal{U}(-2,2)$ for α_2, and $\mathcal{U}(0.1,0.6)$ for τ. These were chosen according to a blend of the grid search minimizers $\boldsymbol{\alpha}_0$ and τ_0 and postulation of reasonable values for a range of an exponential covariance. After discarding a burn-in of 2000 values, we verify convergence via standard checks and obtain 10,000 samples, whose densities are displayed in Figure 10. The uncertainty is reduced compared to the priors, but is still substantial considering the large prior widths. To investigate this, we can look at the aggregate variograms $\widehat{\Gamma}(m)$ for each month, for simulations using both the means of the ABC posteriors for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and τ^2, as well as 100 random samples from each of these posteriors, shown in Figure 11. We see that while $\widehat{\Gamma}$ is estimated closely for many months, we see

10.1002/2016WR019741

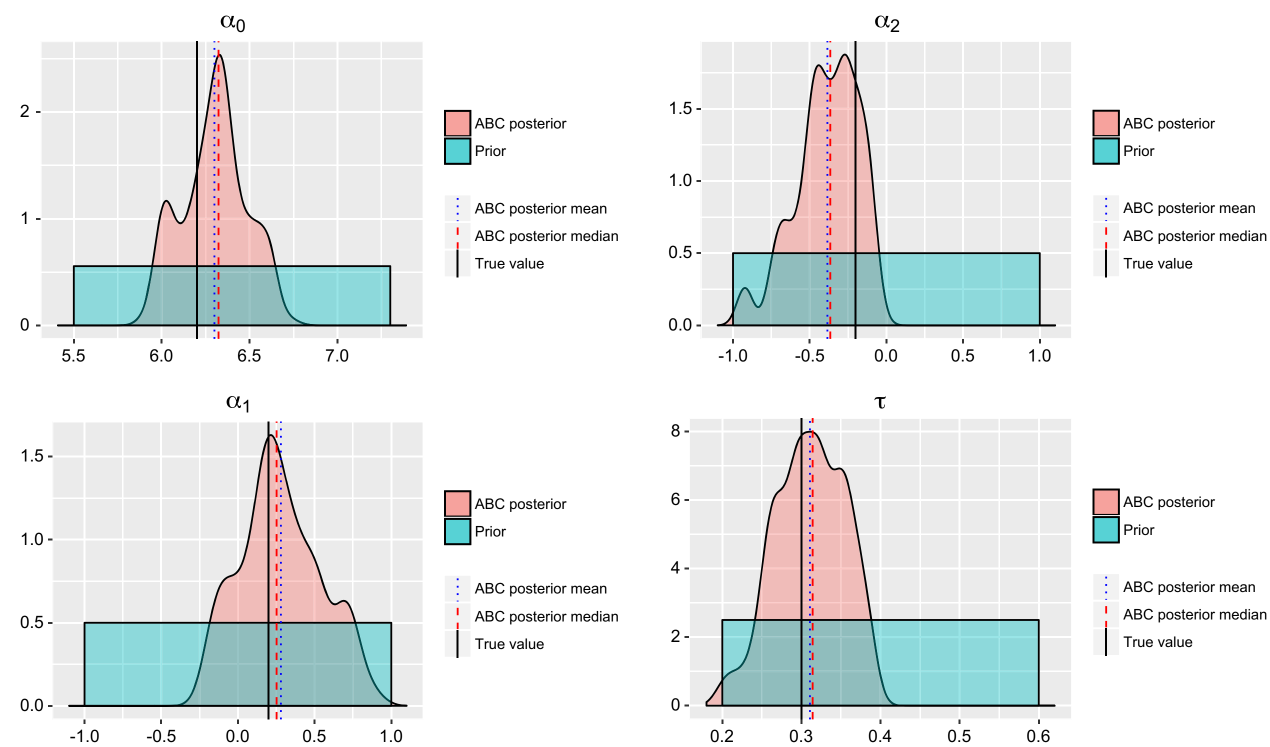

Figure 7. The prior and ABC posterior densities for parameters α_i from (4), as well as the (square root of the) nugget effect τ, for the simulation study.

consistent overestimation and underestimation for several months. One benefit of the variability allowed by the posteriors is that the sampled versions are doing a better job of containing the truth (in January, for instance), whereas the fixed parameter version misses the truth more often. As mentioned in the simulation study, the ability for ϱ_γ to recover the true parameters might suggest that much of the error seen here is the product of our model limitations. Still, the levels of bias are nontrivial and should be taken into high consideration by those who wish to implement such a procedure in a geophysical application. Nonetheless, the covariance structure is well identified and replicated.

One approach for spatial model validation is to examine the distribution of sequences of days where at least one location in the domain experienced precipitation. We generate 500 simulated trajectories of spatiotemporal precipitation occurrence over Iowa for the same time period as the data, each using different posterior samples based on our ABC-MCMC approach. We additionally introduce a zero-covariance process (i.e., forcing W to be spatial white noise), so that each locations' occurrence sequence is independent of the others. Figure shows the results of this experiment, comparing domain wet spell count lengths. Unsurprisingly, the white noise case produces too few long spell lengths, and too many short spell lengths. Both cases which utilize positive covariance have strong alignment with the observations, with the ABC-MCMC-sampled parameter case exhibiting considerably higher variability. This is encouraging, as uncertainty inherent in parameters is seen to propagate through the simulations, allowing for the uncertainty in domain statistics to capture the truth.

6. Discussion

Using a latent Gaussian process allows for correlated binary-valued simulations of precipitation occurrence, and variations of ABC algorithms allow us to adequately estimate the underlying statistical parameters while bypassing likelihood evaluations. In particular, ABC methods seem to be particularly suited to the stochastic weather generation problem, where there are often clear scientifically meaningful statistics that the generator attempts to reproduce. Our case studies in Colorado and Iowa suggest rigorously chosen ABC similarity metrics can yield meaningful and accurate samples from posterior distributions.

10.1002/2016WR019741

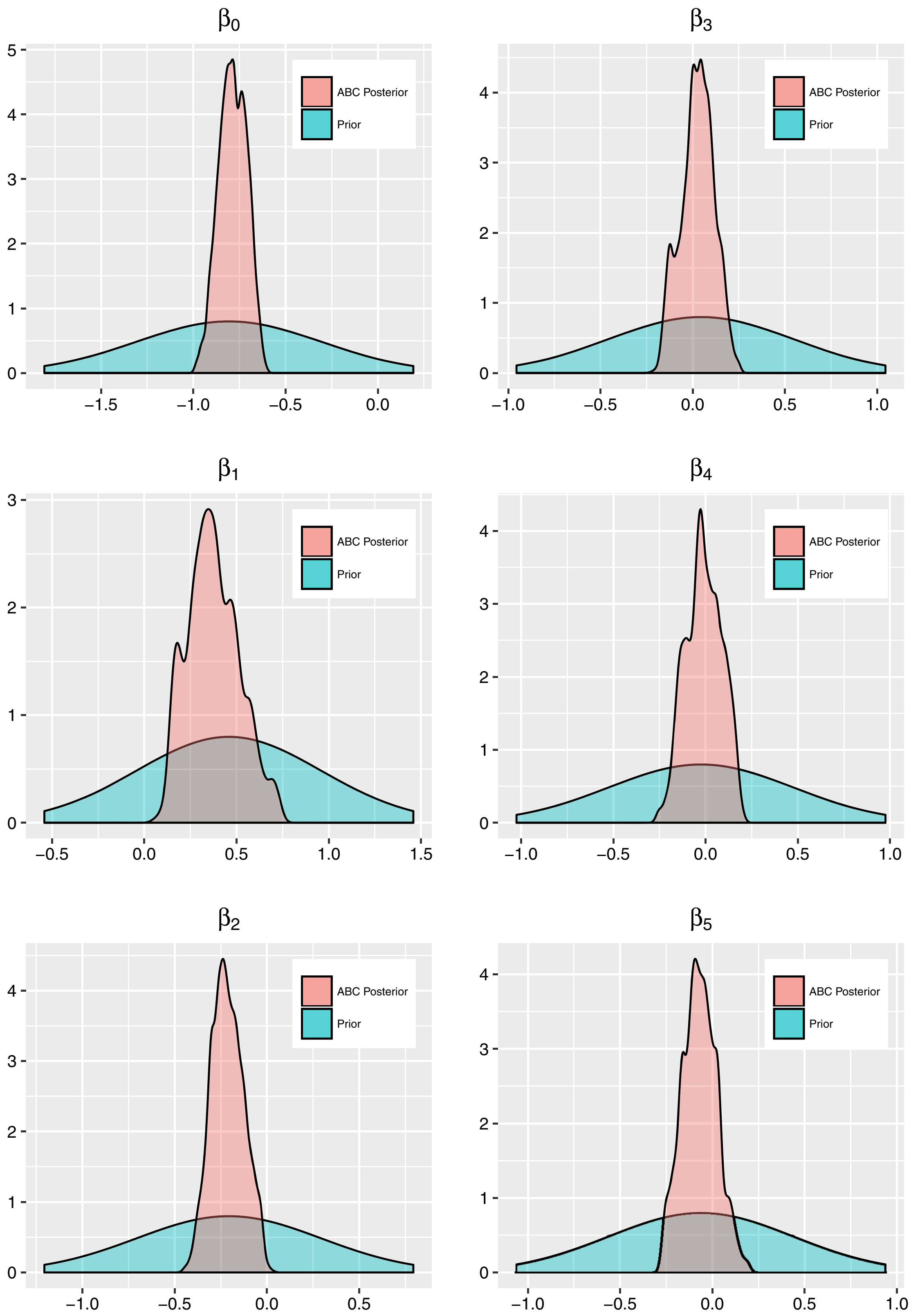

Figure 8. The prior and ABC posterior densities for each parameter β_i via (3), for the Iowa data study.

10.1002/2016WR019741

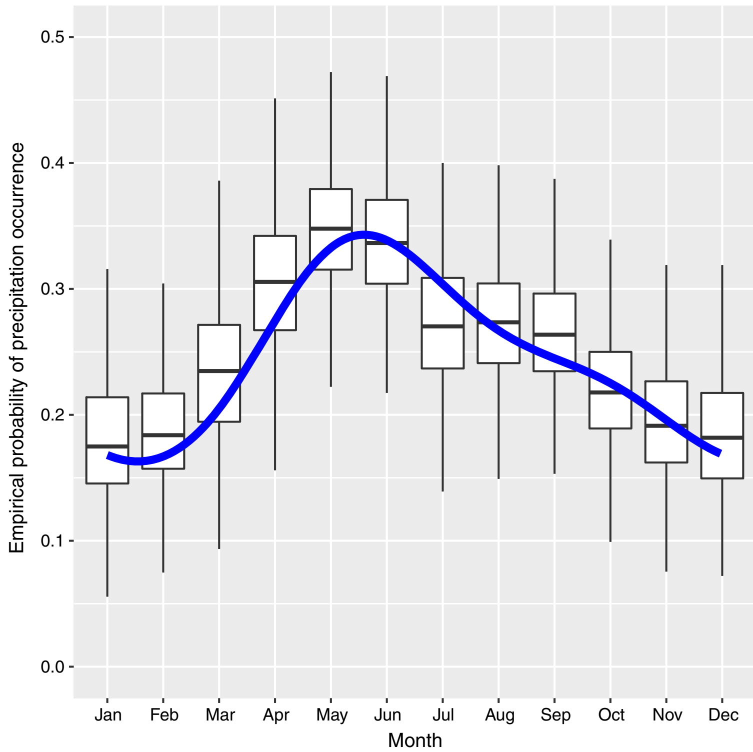

Figure 9. The empirical probability of precipitation by day averaged over each month and over all of the 22 locations in Iowa (represented by the box plots), as well as the mean daily probability of precipitation given by our $\boldsymbol{\beta}_{\mathrm{ABC-MCMC}}$ estimate (represented by the solid blue line).

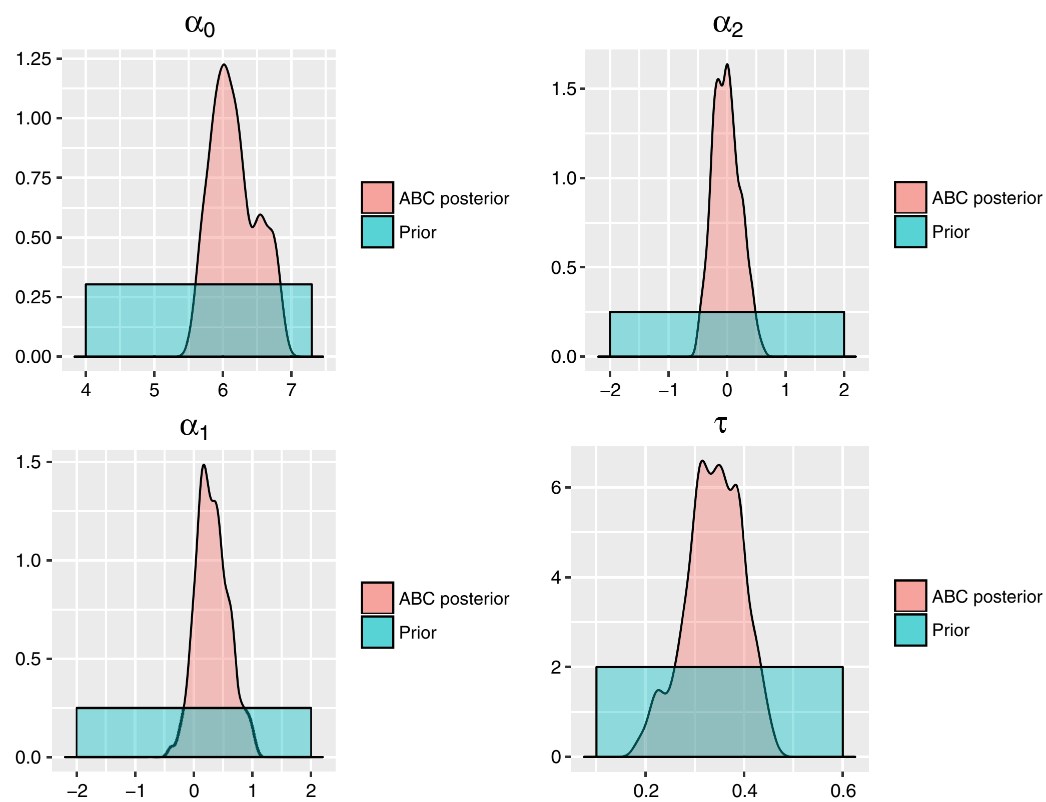

Figure 10. The prior and ABC posterior densities for parameters α_i from (4), as well as the (square root of the) nugget effect τ, for the Iowa data study.

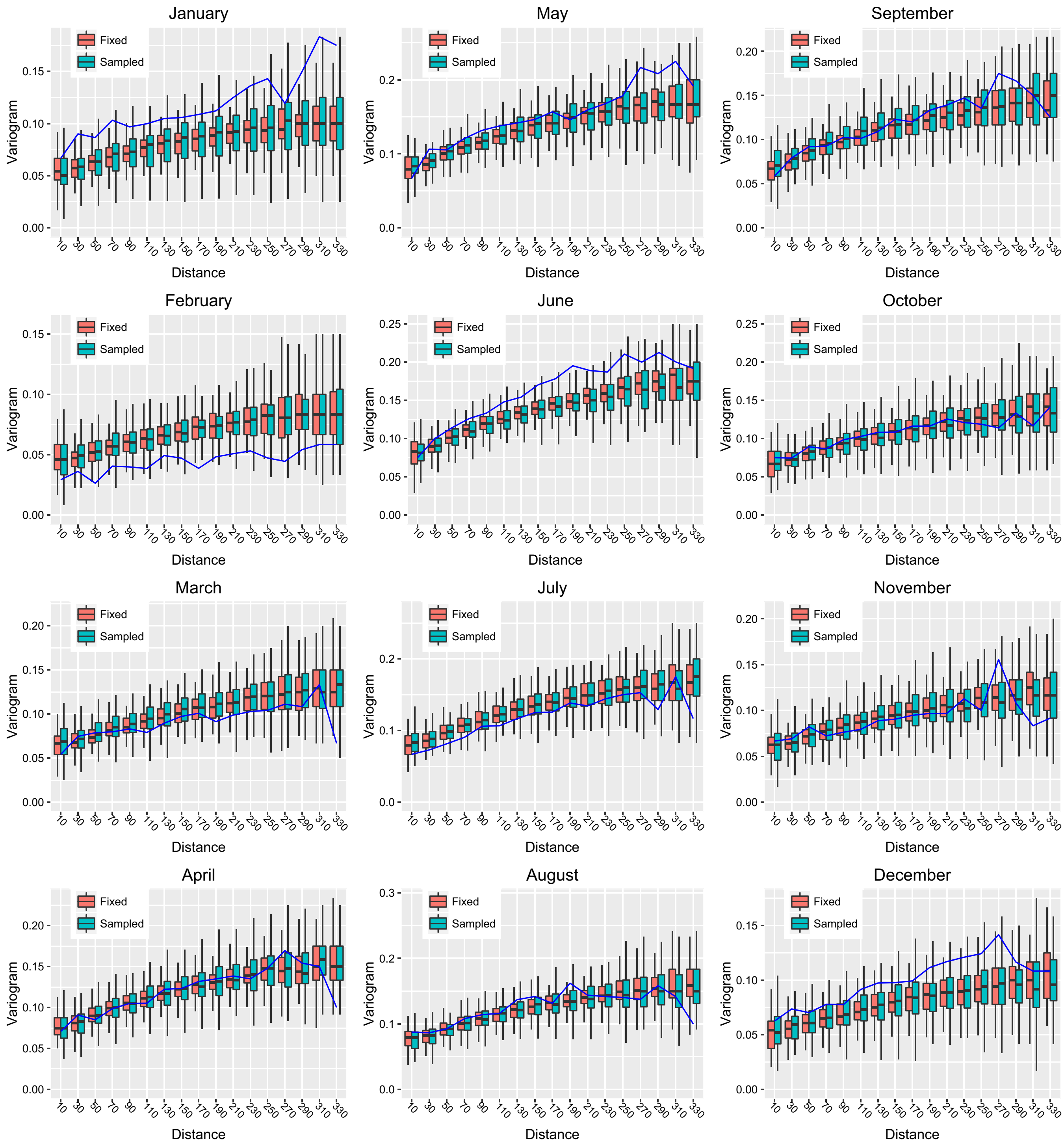

Figure 11. The aggregate variogram $\widehat{\Gamma}(m)$ of our simulated thresholded Gaussian process for each month for the state of Iowa (represented by the box plots), as well as the observed aggregate variogram for the state of Iowa (represented by the solid blue lines).

This opens up many avenues of future research, including exploration of other similarity metrics, applications to other weather variables, and applications to other statistical approaches to precipitation problems (especially in domains with complex terrain). A natural expansion upon our methodology would be the integration of precipitation intensity, although useful and reliable similarity metrics are unclear.

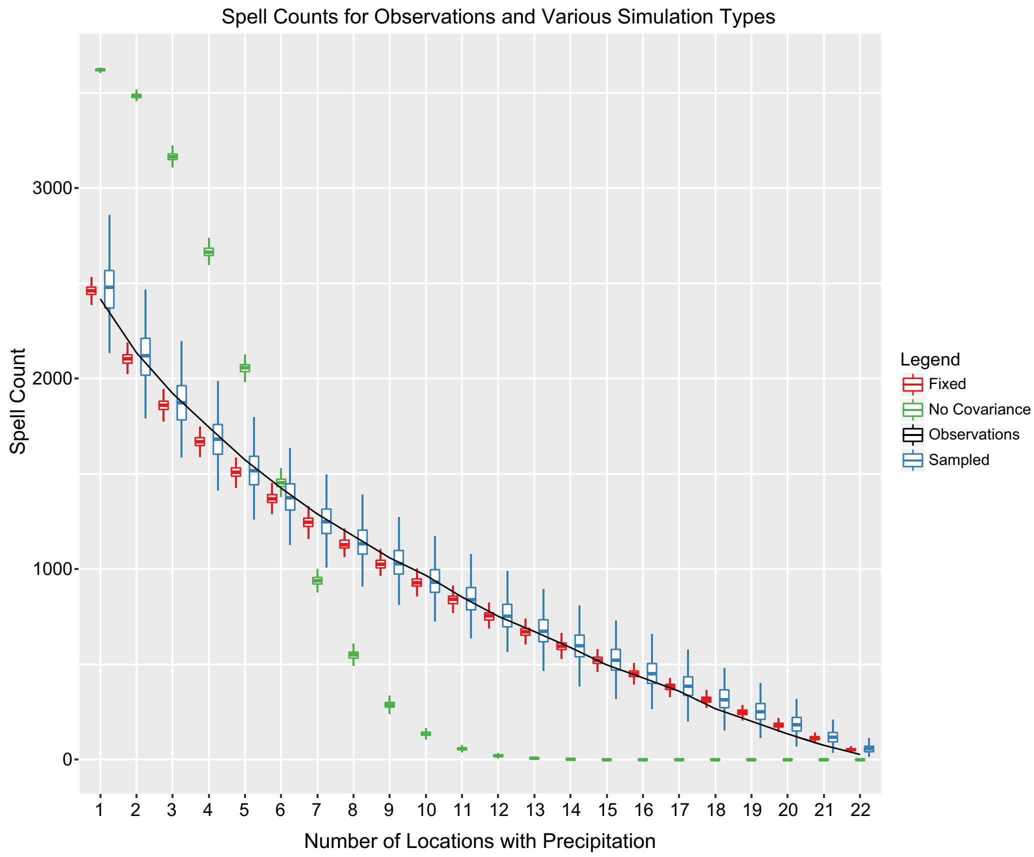

Figure 12. Domain spell count lengths for the observed occurrence values, as well as the three simulation examples which utilize the posterior distributions.

Appendix A: Analytic Solution of True Posterior for the Single-Site Instance of Model (2) With Uniform Priors

In what follows, let $\mathbf{o}=(o_1,\ldots,o_T)$ be a vector of observed precipitation events, so that $\Pr[o_i]=$ the probability that it rained on day i, and let $\boldsymbol{\beta}=(\beta_0,\ldots,\beta_k)^\top$ be our parameter vector. Also, let $p_0=\Pr[o_1|\boldsymbol{\beta}]$, the constant that corresponds to the probability of rain on the very first day.

First, we seek the aid of the multiplicative rule of probability, which states that, for some events $\{A_1,\ldots,A_n\}$,

$$\Pr\left[\bigcap_{i=1}^{n} A_i\right]=\Pr[A_1]\Pr[A_2|A_1]\Pr[A_3|A_1\cap A_2]\times\cdots\times\Pr\left[A_n\middle|\bigcap_{i=1}^{n-1} A_i\right] \quad \text{(A1)}$$

$$=\Pr[A_1]\prod_{i=2}^{n}\Pr\left[A_i\middle|\bigcap_{j=1}^{i-1} A_j\right]. \quad \text{(A2)}$$

In terms of our likelihood function $\Pr[\mathbf{o}|\boldsymbol{\beta}]=L(\boldsymbol{\beta}|\mathbf{o})$, we see that

$$\Pr[\mathbf{o}|\boldsymbol{\beta}]=\Pr[o_1|\boldsymbol{\beta}]\prod_{d=2}^{T}\Pr\left[o_i\middle|\bigcap_{j=1}^{d-1} o_j,\boldsymbol{\beta}\right]. \quad \text{(A3)}$$

However, by the Markov property of our precipitation chain, each occurrence of precipitation only depends on the precipitation of the previous day. Thus, this leaves

$$\Pr[\mathbf{o}|\boldsymbol{\beta}]=\Pr[o_1|\boldsymbol{\beta}]\prod_{d=2}^{T}\Pr[o_d|o_{d-1},\boldsymbol{\beta}]. \quad \text{(A4)}$$

Next we note that by (1),

$$\log f(\boldsymbol{\beta}|\mathbf{o}) \propto \log(L(\boldsymbol{\beta}|\mathbf{o})\pi(\boldsymbol{\beta})) \;\; \Rightarrow \;\; \log f(\boldsymbol{\beta}|\mathbf{o}) = \log L(\boldsymbol{\beta}|\mathbf{o}) + \log \pi(\boldsymbol{\beta}) + c_1, \tag{A5}$$

for some constant $c_1 \in \mathbb{R}$. Isolating the log-likelihood we find that

$$\log L(\boldsymbol{\beta}|\mathbf{o}) = \log(\Pr[\mathbf{o}|\boldsymbol{\beta}]), \tag{A6}$$

$$= \log\left(\Pr[o_1|\boldsymbol{\beta}]\prod_{d=2}^{T}\Pr[o_d|o_{d-1},\boldsymbol{\beta}]\right), \tag{A7}$$

$$= \log(\Pr[o_1|\boldsymbol{\beta}]) + \sum_{d=2}^{T}\log(\Pr[o_d|o_{d-1},\boldsymbol{\beta}]), \tag{A8}$$

$$= \log p_0 + \sum_{d=2}^{T}\log\left(\Phi(\boldsymbol{\beta}^\top\mathbf{X})^{o_d}\left[1-\Phi(\boldsymbol{\beta}^\top\mathbf{X})\right]^{1-o_d}\right), \tag{A9}$$

$$= \sum_{d=2}^{T}\left\{o_d \log(\Phi(\boldsymbol{\beta}^\top\mathbf{X})) + (1-o_d)\log(1-\Phi(\boldsymbol{\beta}^\top\mathbf{X}))\right\} + \log p_0, \tag{A10}$$

$$= \sum_{d=2}^{T}\left\{\mathbf{1}_{[o_d=1]}\log(\Phi(\boldsymbol{\beta}^\top\mathbf{X})) + \mathbf{1}_{[o_d=0]}\log(1-\Phi(\boldsymbol{\beta}^\top\mathbf{X}))\right\} + \log p_0. \tag{A11}$$

Plugging this into (A5) yields

$$\log f(\boldsymbol{\beta}|\mathbf{o}) \propto \begin{array}{l}\displaystyle\sum_{d=2}^{T}\left\{\mathbf{1}_{[o_d=1]}\log(\Phi(\boldsymbol{\beta}^\top\mathbf{X})) + \mathbf{1}_{[o_d=0]}\log(1-\Phi(\boldsymbol{\beta}^\top\mathbf{X}))\right\} \\ + \log \pi(\boldsymbol{\beta}) + c\end{array}, \tag{A12}$$

where $c = c_1 + \log p_0$. Moreover, since we are using uniform, independent priors, we can simplify the expression for $\log \pi(\boldsymbol{\beta})$:

$$\log \pi(\boldsymbol{\beta}) = \log\prod_{i=0}^{k}\pi_i(\beta_i), \tag{A13}$$

$$= \sum_{i=0}^{k}\log\left(\frac{1}{b_i-a_i}\times\mathbf{1}_{[a_i\le\beta_i\le b_i]}\right), \tag{A14}$$

$$= \sum_{i=0}^{k}\left\{\log \mathbf{1}_{[a_i\le\beta_i\le b_i]} - \log(b_i-a_i)\right\}. \tag{A15}$$

Therefore, we arrive at

$$\begin{aligned}\log f(\boldsymbol{\beta}|\mathbf{o}) \;\; &= \sum_{d=2}^{T}\left\{\mathbf{1}_{[o_d=1]}\log(\Phi(\boldsymbol{\beta}^\top\mathbf{X})) + \mathbf{1}_{[o_d=0]}\log(1-\Phi(\boldsymbol{\beta}^\top\mathbf{X}))\right\} \\ &\quad + \sum_{i=0}^{k}\left\{\log \mathbf{1}_{[a_i\le\beta_i\le b_i]} - \log(b_i-a_i)\right\} + c,\end{aligned} \tag{A16}$$

which was to be shown.

Acknowledgments
The authors thank the associate editor and reviewers for suggestions that helped improve the manuscript. Kleiber's portion of this research was supported by National Science Foundation (NSF) grants DMS-1417724, DMS-1406536, and BCS-1461576, and Olson's portion was supported by NSF EXTREEMS DMS-1407340. All precipitation data were provided by the Global Historical Climatology Network and is publicly available. All data created by the authors for use in this study can be requested by emailing Branden Olson at brando6@uw.edu.

References

Ailliot, P., C. Thompson, and P. Thomson (2009), Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions, *J. R. Stat. Soc., Ser. C*, *58*(3), 405–426.

Allcroft, D. J., and C. A. Glasbey (2003), A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation, *J. R. Stat. Soc., Ser. C*, *52*(4), 487–498.

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

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

Beaumont, M. A., W. Zhang, and D. J. Balding (2002), Approximate Bayesian computation in population genetics, *Genetics*, *162*(4), 2025–2035.

Brissette, F. P., M. Khalili, and R. Leconte (2007), Efficient stochastic generation of multi-site synthetic precipitation data, *J. Hydrol.*, *345*(3–4), 121–133.

Brown, P. E., P. J. Diggle, M. E. Lord, and P. C. Young (2001), Space-time calibration of radar rainfall data, *J. R. Stat. Soc., Ser. C*, *50*(2), 221–241.

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.

Cannon, A. J. (2008), Probabilistic multisite precipitation downscaling by an expanded Bernoulli-Gama density network, *J. Hydrometeorol.*, *9*(6), 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.

Chib, S., and E. Greenberg (1995), Understanding the Metropolis-Hastings algorithm, *Am. Stat.*, *49*(4), 327–335.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley Intersci., New York.

Cressie, N., and D. L. Zimmerman (1992), On the stability of the geostatistical method, *Math. Geol.*, *24*(1), 45–59.

Dean, T. A., S. S. Singh, A. Jasra, and G. W. Peters (2011), Parameter estimation for hidden Markov models with intractable likelihoods, *Scand. J. Stat.*, *41*(4), 970–987.

Deidda, R. (2000), Rainfall downscaling in a space-time multifractal framework, *Water Resour. Res.*, *36*(7), 1779–1794.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998), Model-based geostatistics, *J. R. Stat. Soc., Ser. C*, *47*(3), 299–325.

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

Fearnhead, P., and D. Prangle (2011), Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC, *J. R. Stat. Soc., Ser. B*, *74*(3), 419–474.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi (2007), Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling, *Int. J. Climatol.*, *27*, 1547–1578.

Friend, A. D., A. K. Stevens, R. G. Knox, and M. G. R. Cannell (1997), A process-based, terrestrial biosphere model of ecosystem dynamics, *Ecol. Modell.*, *95*, 249–287.

Fu, Y. X., and W. H. Li (1997), Estimating the age of the common ancestor of a sample of DNA sequences, *Mol. Biol. Evol.*, *14*, 195–199.

Gelman, A., G. O. Roberts, and W. R. Gilks (1996), Efficient Metropolis jumping rules, *Bayesian Stat.*, *5*, 599–608.

Gupta, V. K., and E. C. Waymire (1993), A statistical analysis of mesoscale rainfall as a random cascade, *J. Appl. Meteorol.*, *32*, 251–267.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, *57*(1), 97–109.

Hewitson, B. C., and R. G. Crane (1996), Climate downscaling: Techniques and application, *Clim. Res.*, *7*, 85–95.

Hughes, J. P., P. Guttorp, and S. P. Charles (1999), A non-homogeneous hidden Markov model for precipitation occurrence, *J. R. Stat. Soc., Ser. C*, *48*(1), 15–30.

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

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.

Kustas, W., A. Rango, and R. Uijlenhoet (1994), A simple energy budget algorithm for the snowmelt runoff model, *Water Resour. Res.*, *30*(5), 1515–1527.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for resampling hydrological time series, *Water Resour. Res.*, *32*(3), 679–693.

Langousis, A., and D. Veneziano (2007), Intensity-duration-frequency curves from scaling representations of rainfall, *Water Resour. Res.*, *43*, W02422, doi:10.1029/2006WR005245.

Langousis, A., and V. Kaleris (2014), Statistical framework to simulate daily rainfall series conditional on upper-air predictor variables, *Water Resour. Res.*, *50*, 3907–3932, doi:10.1002/2013WR014936.

Langousis, A., A. Mamalakis, R. Deidda, and M. Marrocu (2016), Assessing the relative effectiveness of statistical downscaling and distribution mapping in reproducing rainfall statistics based on climate model results, *Water Resour. Res.*, *52*, 471–494, doi:10.1002/2015WR017556.

Lovejoy, S., and D. Schertzer (1995), Multifractals and rain, in *New Uncertainty Concepts in Hydrology and Hydrological Modelling*, edited by A. W. Kundzewicz, Cambridge Press, Cambridge, U. K.

Mauran, 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.

Mearns, L. O., C. Rosenzweig, and R. Goldberg (1997), Mean and variance change in climate scenarios: Methods, agricultural applications, and measures of uncertainty, *Clim. Change*, *35*, 367–396.

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.

Menabde, M., D. Harris, A. Seed, G. Austin, and D. Stow (1997), Multiscaling properties of rainfall and bounded random cascades, *Water Resour. Res.*, *33*(12), 2823–2830.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, *J. Chem. Phys.*, *21*(6), 1087–1092.

Nott, D. J., L. Marshall, and J. Brown (2012), Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?, *Water Resour. Res.*, *48*, W12602, doi:10.1029/2011WR011128.

Olson, B. (2016), Stochastic weather generation with approximate Bayesian computation, Master's thesis, Univ. of Colorado, Boulder, Colo.

Peterson, T. C., and R. S. Vose (1997), An overview of the Global Historical Climatology Network temperature database, *Bull. Am. Meterol. Soc.*, *78*, 2837–2849.

Pritchard, J., M. Seielstad, A. Perez-Lezaun, and M. Feldman (1999), Population growth of human Y chromosomes: A study of Y chromosome microsatellites, *Mol. Biol. Evol.*, *16*, 1791–1798.

Racsko, P., L. Szeidl, and M. Semenov (1991), A serial approach to local stochastic weather models, *Ecol. Modell.*, *57*, 27–41.

Rajagopalan, B., and U. Lall (1999), A *k*-nearest-neighbor simulator for daily precipitation and other weather variables, *Water Resour. Res.*, *35*(10), 3089–3101.

Richardson, C. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resour. Res.*, *17*(1), 182–190.

Rubin, D. (1984), Bayesianly justifiable and relevant frequency calculations for the applied statistician, *Ann. Stat.*, *12*(4), 1151–1172.

Sadegh, M., and J. A. Vrugt (2013), Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation, *Hydrol. Earth Syst. Sci.*, *17*, 4831–4850.

Sadegh, M., and J. A. Vrugt (2014), Approximate Bayesian computation using Markov chain Monte Carlo simulation: DREAM-ABC, *Water Resour. Res.*, *50*, 6767–6787, doi:10.1002/2014WR015386.

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

Semenov, M. A., and E. M. Barrow (1997), Use of a stochastic weather generator in the development of climate change scenarios, *Clim. Change*, *35*.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2007), Sequential Monte Carlo without likelihoods, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(6), 1760–1765.

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

Strandman, H., H. Väisänen, and S. Kellomaki (1993), A procedure for generating synthetic weather records in conjunction of climate scenario for modelling of ecological impacts of changing climate in boreal conditions, *Ecol. Modell.*, *70*, 195–220.

Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz (2013), Approximate Bayesian computation, *PLoS Comput. Biol.*, *9*(1), e1002803.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997), Inferring coalescence times from DNA sequence data, *Genetics*, *145*, 505–518.

Thompson, C. S., P. J. Thomson, and X. Zheng (2007), Fitting a multisite daily rainfall model to New Zealand data, *J. Hydrol.*, *340*(1–2), 25–39.

Veneziano, D., and A. Langousis (2005), The areal reduction factor: A multifractal analysis, *Water Resour. Res.*, *41*, W07008, doi:10.1029/2004WR003765.

Veneziano, D., and A. Langousis (2010), Scaling and Fractals in Hydrology, in *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting*, edited by B. Sivakumar and R. Berndtsson, 145 pp., World Sci., Singapore.

Veneziano, D., A. Langousis, and P. Furcolo (2006), Multifractality and rainfall extremes: A review, *Water Resour. Res.*, *42*, W06D15, doi:10.1029/2005WR004716.

Weiss, G., and A. von Haeseler (1998), Inference of population history using a likelihood approach, *Genetics*, *149*(3), 1539–1546.

Wilks, D. S. (1988), Estimating the consequences of CO_2-induced climatic change on North American grain agriculture using general circulation model information, *Clim. Change*, *13*(1), 19–42.

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

Wilks, D. S. (1999), Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain, *Agric. For. Meteorol.*, *96*, 85–101.

Wilks, D. S. (2010), Use of stochastic weather generators for precipitation downscaling, *Wiley Interdisciplinary Rev.*, *1*, 898–907.

Wilks, D. S., and R. Wilby (1999), The weather generation game: A review of stochastic weather models, *Prog. Phys. Geogr.*, *23*(3), 875–884.

Wilby, R. L., S. P. Charles, E. Zorita, B. Timbal, P. Whetton, and L. O. Mearns (2004), Guidelines for use of climate scenarios developed from statistical downscaling methods, in *Supporting material of the Intergovernmental Panel on Climate Change*, 27 pp., Cambridge University Press, Cambridge, U. K. [Available from the DDC of IPCC TGCIA.]

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.

Zimmerman, D. L., and M. B. Zimmerman (1991), A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors, *Technometrics*, *33*(1), 77–91.

Zorita, E., and H. von Storch (1997), A survey of statistical downscaling techniques, *GKSS Rep. 97/E/20*, GKSS Res. Cent, Geesthacht, Germany.