The Annals of Applied Statistics
2019, Vol. 13, No. 2, 683–712
https://doi.org/10.1214/18-AOAS1218

CLIMATE INFERENCE ON DAILY RAINFALL ACROSS THE AUSTRALIAN CONTINENT, 1876–2015

BY MICHAEL BERTOLACCI[*,1], EDWARD CRIPPS[*,1], ORI ROSEN[‡,2], JOHN W. LAU[*] AND SALLY CRIPPS[†,3,4]

University of Western Australia[*], University of Sydney[†] and University of Texas at El Paso[‡]

Daily precipitation has an enormous impact on human activity, and the study of how it varies over time and space, and what global indicators influence it, is of paramount importance to Australian agriculture. We analyze over 294 million daily rainfall measurements since 1876, spanning 17,606 sites across continental Australia. The data are not only large but also complex, and the topic would benefit from a common and publicly available statistical framework. We propose a Bayesian hierarchical mixture model that accommodates mixed discrete-continuous data. The observational level describes site-specific temporal and climatic variation via a mixture-of-experts model. At the next level of the hierarchy, spatial variability of the mixture weights' parameters is modeled by a spatial Gaussian process prior. A parallel and distributed Markov chain Monte Carlo sampler is developed which scales the model to large data sets. We present examples of posterior inference on the mixture weights, monthly intensity levels, daily temporal dependence, offsite prediction of the effects of climate drivers and long-term rainfall trends across the entire continent. Computer code implementing the methods proposed in this paper is available as an R package.

1. Introduction. Australia is the world's driest inhabited continent and is subject to highly variable rainfall patterns. Rainfall classifications by the Australian Government Bureau of Meteorology (BOM) appear in Figure 1(a) and show extreme spatial variability.[5] In Australia's large arid interior, which extends to the central west and south coasts, some stations receive as little as 150 mm of annual median rainfall, while in Australia's tropical north some stations receive as much as 4000 mm. Rainfall in the southern coastal fringe is winter dominant,

Received April 2018; revised October 2018.

[1]Supported by the ARC Industrial Transformation Research Hub for Offshore Floating Facilities which is funded by the Australian Research Council, Woodside Energy, Shell, Bureau Veritas and Lloyds Register Grant IH140100012, and by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

[2]Supported in part by NSF Grant DMS-1512188.

[3]Supported by Centre for Translational Data Science, University of Sydney.

[4]Supported by Australian Research Council Australian Future Fellowship 140101266.

Key words and phrases. Climate, rainfall, Australia, mixture-of-experts, Gaussian processes, parallel and distributed computing.

[5]www.bom.gov.au/jsp/ncc/climate_averages/climate-classifications/index.jsp?maptype=seasb.

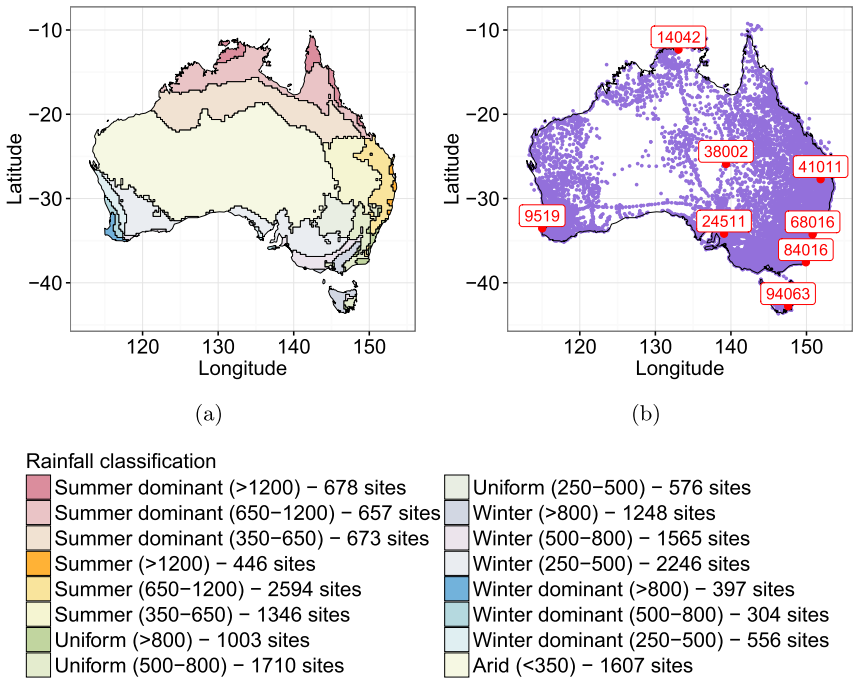


FIG. 1. *Australian Bureau of Meteorology* (*BOM*) *rainfall categories with median annual rainfall levels in millimeters* (a) *and the locations of the* 17,606 *rainfall observation sites analyzed in this article* (b). *Eight sites, chosen to reflect the rainfall categories and used as examples throughout this paper, are marked in red in* (b).

while rainfall in the mid-east coast is summer dominant. The cycle of *droughts and flooding rains*[6] is a common feature of many locations and, indeed, both droughts and floods often occur simultaneously in different regions (see Risbey et al. (2009), Ummenhofer et al. (2009) and references therein). Although this spatial and temporal variability in rainfall is a natural part of Australia's climate, these patterns cause economic and environmental problems, most notably in agriculture, ecosystems and civic water supplies (see Ummenhofer et al. (2015), van Dijk et al. (2013)). As a result, the Australian Government is increasingly relying on the scientific understanding of the drivers of Australian rainfall to inform policy making (Stone (2014)).

Oceanic and atmospheric interactions in the Pacific, Indian and Southern Oceans are thought to be the main climate drivers of the variability in Australian rainfall. There is debate surrounding the boundaries and overlap of the influences of these climate drivers on rainfall across different regions of Australia (see, among others, Ummenhofer et al. (2009, 2011), Feng, Li and Li (2010), Cai et al. (2012),

[6]Dorothea Mackellar, *My Country*.

King, Alexander and Donat (2013) and Pepler et al. (2014)), but, broadly speaking, they are most influential in northern and eastern Australia via the Pacific Ocean, across central and southern Australia via the Indian Ocean and the higher latitudes of Australia via the Southern Ocean (see Risbey et al. (2009) for an overview).

Despite significant advances in rainfall modeling, there are two shortcomings in the climate science literature. First, it is common to aggregate or smooth daily rainfall measurements. This practice ignores the aspects of incidence (rainfall versus no rainfall days) and missingness, and masks nearby spatial/temporal dependencies, which results in an underestimation of uncertainty. Second, there is a scarcity of coherent probabilistic models, which are required to develop a statistical understanding of the evolution of rainfall across both time and space, and the dependence of this evolution on major climate drivers. This article makes progress on these issues by proposing a general Bayesian method to jointly assess the spatial variability of the impact of climate drivers and other covariates on the evolution of daily rainfall across the Australian continent.

Probabilistic modeling of Australian daily rainfall has two competing challenges. First, datasets are often large; our method is applied to approximately 2.94×10^8 time series measurements on 17,606 sites distributed across the entire Australian continent, an area of 7.7 million square kilometers, for the years 1876–2015 inclusive. Figure 1(b) shows the number and location of these sites. Second, daily rainfall measurements have several nonstandard features which are tricky to model: they contain many zeros and have a heavy tailed nonzero rainfall component; they exhibit short, medium and long term dependencies; they vary spatially; and they often have many missing observations—in our application there are 4.4×10^7 missing observations. Thus, a model needs to be sufficiently complex to capture these features, yet parsimonious so that estimation, inference and prediction are computationally feasible for very large datasets.

Statistical modeling of the temporal and spatial evolution of daily rainfall measurements has a rich history. Richardson (1981) introduces a model of rainfall incidence with a two-state Markov chain for rain and no rain, with an exponential distribution for nonzero rainfall days and a state transition probability that is time-varying. This model has seen many extensions over the following decades, including replacing the exponential distribution with a gamma distribution (Stern and Coe (1984)), with a mixture of two exponentials (Wilks (1999)), and with a mixture of gamma and Generalized Pareto distributions (Vrac and Naveau (2007)). Furrer and Katz (2007) extend Richardson's approach to a generalized linear modeling (GLM) framework. Kleiber, Katz and Rajagopalan (2012) extend the model in Furrer and Katz (2007) to a spatial setting, with a mixture of two components, where the mixture weights depend on covariates, and the spatial variability of the regression parameters is modeled via a latent Gaussian process. More recently Naveau et al. (2016) have developed a class of methods for modeling marginal daily rainfall, based on extreme value theory, that jointly accommodates low, moderate and heavy rainfall.

Most relevant to our work are two recent articles (Holsclaw et al. (2016, 2017)) which take a Bayesian approach. Holsclaw et al. (2016) propose a hidden Markov model with a state for zero rainfall and a mixture of an exponential and a gamma distribution to allow for nonzero rainfall days. They apply their method to daily rainfall measurements in northern India and Pakistan, and the upper Yangtze River basin in China, using an ordered multinomial probit model for the probabilities of rainfall states. These state transition probabilities are allowed to vary in time, but the spatial dependence of site-specific parameters is ignored. Holsclaw et al. (2017) extend the work of Holsclaw et al. (2016) to a region-wide mixture model for analyzing the temporal evolution of rainfall states for the subcontinent of India. A contribution of their work is the use of the Pólya–Gamma data augmentation algorithm proposed by Polson, Scott and Windle (2013), to facilitate the MCMC scheme needed for the high dimensional integration. One aim of their article is to provide a sampler that requires little tuning and is capable of handling about 6.9×10^5 observations (63 sites over 30 years).

Our article makes three main contributions. The first is a systematic climate modeling of the effects of the aforementioned oceanic/atmospheric interactions on rainfall that accounts for many sources of uncertainty and produces a spatially varying posterior distribution on a scale as large as the Australian continent. To this end, we develop a hierarchical Bayesian mixture model which can be tailored to answer many research questions regarding daily rainfall, constituting the second contribution of this article.

At the observational level, we describe the temporal evolution of daily rainfall via a mixture-of-experts model (Jacobs et al. (1991), Rosen, Stoffer and Wood (2009), Wood, Rosen and Kohn (2011)). The mixture components are probability density/mass functions for daily rainfall. The mixture weights are parameterized to depend upon covariates that include long-term climate trends, short term autocorrelation and seasonality, and factors which model the impact of external climate drivers and their interactions. Although spatial dependencies exist at the observational level, the size of the problem necessitates a parsimonious model. We do so by modeling spatial variability not at the observational level but rather by allowing the parameters of the mixture weights to vary, using a two-dimensional spatial Gaussian process ($\mathcal{GP}$) prior (Wahba (1990), Wood (2013)). The posterior distributions of the $\mathcal{GP}$ mean parameters are used for climate inference.

The third contribution is the construction of an MCMC algorithm which scales to genuinely large datasets. There are computational challenges in applying any model to such a large data set, parsimonious or not. Similar to Holsclaw et al. (2017), we use the latent variable approach of Polson, Scott and Windle (2013), such that the conditional distributions of these latent variables and the mixture weight parameters permit Gibbs sampling. A conjugate prior for the gamma distribution, as described in Damsleth (1975), also allows the full conditional distributions of the mixture component parameters to be updated via a Gibbs step. We

show how to make model estimation efficient, using distributed and parallel computing, and provide software in the form of an R package (R Core Team (2016)).

The paper proceeds as follows. Section 2 describes the data collection and the construction of the indices of the external climate drivers. Section 3 outlines the general model and how it is tuned for this application. Section 4 details the implementation of the MCMC scheme and how this implementation uses distributed computing to handle the large quantity of data. Section 5 presents some diagnostics for the model and posterior inference on the mixture weights, monthly intensity levels, daily temporal dependence, out-of-sample prediction of the effects of climate drivers and long-term rainfall trends on daily rainfall. Most inferences show broad agreement with previous analyses but some do not, particularly those concerning the oceanic/atmospheric interactions over the Southern Ocean. Section 6 concludes the article.

Three supplements are also available: Bertolacci et al. (2019a) presents the derivation of the conditional distributions required in the MCMC algorithm; Bertolacci et al. (2019b) contains a comparison of two variants of our model; Bertolacci et al. (2019c) presents model diagnostics including metrics of spatial dependency and simulation studies that demonstrate our model's ability to perform climate inference in the presence of spatially correlated data and excessively heavy-tailed data.

2. The data.

2.1. *Rainfall data.* We use daily rainfall measurements on 17,606 Australian observational sites listed by BOM that lie above latitude 50°S and between longitudes 110°W to 155°W[7] (see Figure 1(b)). BOM describes rainfall measurements at its observational sites as:[8]

> Rainfall includes all forms of water particles, whether liquid (for example, rain or drizzle) or solid (hail or snow), that fall from clouds and reaches the ground. The rain gauge is the standard instrument for recording rainfall, which is measured in millimetres. Rainfall is generally observed daily at 9 am local time—this is a measure of the total rainfall that has been received over the previous 24 hours.

For the purposes of this study we take the rainfall measurement for a given date and site to be the rainfall value for that day, and make no distinction as to whether it might be snow, hail or otherwise. The analysis is from February 1st, 1876 to December 31st, 2015, resulting in more than 294 million days of rainfall. The start date is the earliest available recording of the Southern Oscillation Index (SOI) provided by BOM.

[7] Available from http://www.bom.gov.au/climate/data/.

[8] http://www.bom.gov.au/climate/cdo/about/about-rain-data.shtml.

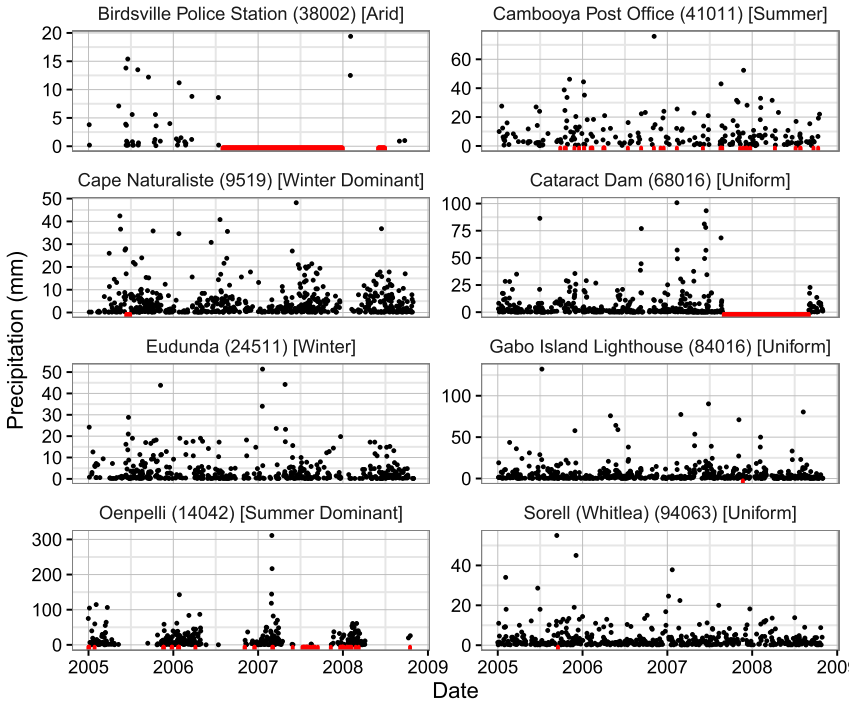


FIG. 2. *Daily rainfall measurements from January* 1*st*, 2005 *to December* 31*st*, 2009, *for eight sites. BOM rainfall classifications are indicated in* [*brackets*]. *Days with zero rainfall are left blank and missing data are colored red.*

Figure 2 presents rainfall measurements from January 1st, 2005, to December 31st, 2009, for eight sites. These sites are chosen to represent the eight most common rainfall classifications in Figure 1(a), and are marked in Figure 1(b). Days with zero rainfall are left blank, and missing data are colored red. Figure 2 indicates that there are substantial gaps in the dataset which take two forms: gaps before a site opens or after it closes, and gaps within the lifetime of a site. The former do not enter the analysis, so that individual sites may contain as few as 89 and as many as 51,102 observations. We treat the latter gaps as missing values, and by these criteria, there are 44,000,301 missing observations.

Rainfall incidence plays an important role in Australian rainfall. Table 1 reports the proportion of zero-rainfall days for the same eight sites across the entire observational period for January (in the Australian summer) and July (in the Australian winter). Oenpelli has 99% of days in winter and 34% of days in summer recording zero rainfall. In comparison, Cape Naturaliste has 25% of days in winter and 88% of days in summer recording zero rainfall. Finally, some sites are classified as uniform, which means that they show little difference between January and July rainfall incidence. Across all sites and over the entire observational period, approx-

TABLE 1
Proportion of days with no rain in January and July for eight sites over the entire observational period

Site	January	July
Birdsville Police Station (38002) [Arid]	0.92	0.95
Cambooya Post Office (41011) [Summer]	0.73	0.81
Cape Naturaliste (9519) [Winter Dominant]	0.88	0.25
Cataract Dam (68016) [Uniform]	0.66	0.73
Eudunda (24511) [Winter]	0.86	0.53
Gabo Island Lighthouse (84016) [Uniform]	0.69	0.54
Oenpelli (14042) [Summer Dominant]	0.34	0.99
Sorell (Whitlea) (94063) [Uniform]	0.73	0.62

imately two thirds of the rainfall measurements in the dataset are identically equal to zero.

2.2. *Climate indices*. We now outline the construction of indices used as proxies for the oceanic/atmospheric drivers of rainfall. These include:

1. the SOI, which measures the difference in surface air pressure between Tahiti and Darwin (Troup (1965)), for the Pacific Ocean;
2. the Dipole Mode Index (DMI), which measures the change in sea surface temperature gradients between the tropical western Indian Ocean and the tropical south-eastern Indian Ocean (Saji et al. (1999), Rayner et al. (2011)), for the Indian ocean; and
3. the Southern Annular Mode (SAM), which measures the difference in zonal mean sea level pressure at 40°S and 65°S (Gong and Wang (1999), Compo et al. (2011)), for the Southern Ocean.

We use monthly values of the SOI sourced from the BOM,[9] which are calculated as

$$\mathrm{SOI}_M = \frac{\Delta \bar{P}_M - \overline{\Delta P_M}}{\mathrm{sd}(\Delta P_M)},$$

where $\Delta \bar{P}_M$ denotes the average difference in Mean Sea Level Pressure (MSLP) between Tahiti and Darwin in month M, and $\overline{\Delta P_M}$ and $\mathrm{sd}(\Delta P_M)$ denote the long-term average and standard deviation of the same quantity, respectively, for the month in question. The long-term averages are derived from the reference period 1933 to 1992.

[9] http://www.bom.gov.au/climate/current/soi2.shtml.

For the DMI, we adopt the procedure as calculated by Japan Agency for Marine-Earth Science and Technology.[10] This is defined as the difference between the Sea Surface Temperature (SST) anomalies between the tropical western Indian Ocean (50°E–70°E, 10°S–10°N) and the tropical south-eastern Indian Ocean (90°E–110°E, 10°S–0°). The values used in this article are calculated using the Hadley Centre Sea Ice and Sea Surface Temperature dataset (Saji et al. (1999), Rayner et al. (2011)).

Finally, for the SAM we use the empirical definition given by Gong and Wang (1999) as,

$$\mathrm{SAM}_M = P_{40,M} - P_{65,M},$$

where $P_{40,M}$ and $P_{65,M}$ are the normalized monthly zonal MSLP at 40°S and 65°S, respectively, using the Hadley Centre Sea Level Pressure dataset (Compo et al. (2011)) to calculate the SAM at a monthly time scale.[11]

All the software used to collect the data has been made available online by the authors. The software required for data acquisition is structured in two R packages. The first R package retrieves the daily rainfall measurements from BOM, with facilities to efficiently bulk-download data.[12] The second R package retrieves and calculates the climate indices.[13]

3. Model and priors.

3.1. *General model for spatially varying marginal temporal processes.* For each site s, $s = 1, \ldots, S$, we model the marginal temporal evolution of daily rainfall at times $t = 1, \ldots, T$ via the following mixture model:

$$y_{t,s} \sim \sum_{k=0}^{K} \pi_{t,s,k}(\cdot|\boldsymbol{\delta}_{s,k,\cdot})\, f_k(y_{t,s}|\boldsymbol{\theta}_k), \tag{3.1}$$

where the f_k's are mixture components comprising continuous and discrete probability distributions parameterized by $\boldsymbol{\theta}_k$, for $k = 0, 1, \ldots, K$. The mixture weights, $\pi_{t,s,k}$, satisfying $0 \le \pi_{t,s,k} \le 1$ and $\sum_{k=0}^{K} \pi_{t,s,k} = 1$, are parameterized by $\boldsymbol{\delta}_{s,k,\cdot} = (\delta_{s,k,1}, \ldots, \delta_{s,k,P})'$, where P is the number of covariates. Note that the "dot" notation in $\boldsymbol{\delta}_{s,k,\cdot}$, used subsequently, indicates a "slice" down the chosen dimension of the three-dimensional array with values $\delta_{s,k,p}$, $s = 1, \ldots, S$, $k = 0, 1, \ldots, K$, $p = 1, \ldots, P$. The model for the mixture weights is as follows; let $z_{t,s}$ be a latent mixture component indicator such that $(y_{t,s}|z_{t,s} = k) \sim f_k(y_{t,s}|\boldsymbol{\theta}_k)$. We encode

[10] Obtained from http://www.jamstec.go.jp/frsgc/research/d1/iod/iod/dipole_mode_index.html.

[11] Obtained from https://www.esrl.noaa.gov/psd/gcos_wgsp/Gridded/data.hadslp2.html.

[12] https://github.com/mbertolacci/bomdata.

[13] https://github.com/mbertolacci/climatedata.

the site-specific temporal dependence through multinomial logits (Jacobs et al. (1991)) of the form

$$\pi_{t,s,k} = P(z_{t,s} = k|\boldsymbol{x}_{t,s}, \Delta_{s,\cdot,\cdot}) = \frac{\exp(\boldsymbol{x}'_{t,s}\boldsymbol{\delta}_{s,k,\cdot})}{\sum_{k=0}^{K}\exp(\boldsymbol{x}'_{t,s}\boldsymbol{\delta}_{s,k,\cdot})}, \tag{3.2}$$

where $\boldsymbol{x}_{t,s}$ is a $P \times 1$ vector of time-varying covariates, $\Delta_{s,\cdot,\cdot} = (\boldsymbol{\delta}_{s,1,\cdot}, \ldots, \boldsymbol{\delta}_{s,K,\cdot})$, and for identifiability, $\boldsymbol{\delta}_{s,0,\cdot}$ is set to $\mathbf{0}$. The construction of the mixture components in equation (3.1) and covariates in equation (3.2) for our application are discussed in Sections 3.2 and 3.3 below.

3.1.1. *Model for spatially varying inference.* We assume that the site-specific parameters, $\delta_{s,k,p}$, which prescribe the mixture weights, satisfy

$$\delta_{s,k,p} \sim N(\mu_{s,k,p}, \sigma^2_{k,p}) \qquad \text{where} \tag{3.3}$$

$$\boldsymbol{\mu}_{\cdot,k,p} \sim \mathcal{GP}(W\boldsymbol{\beta}_{k,p}, \tau^2_{k,p}\Omega) \tag{3.4}$$

for $p = 1, \ldots, P$ and $k = 0, 1, \ldots, K$, so that for each mixture component, k, and coefficient, p, $\mu_{s,k,p}$ is a site-specific scalar mean, and $\sigma^2_{k,p}$ is a common variance across all sites, whose goal is to capture nonspatial variation between sites. In (3.4), the spatial dependence among the $\mu_{s,k,p}$ is induced by a Gaussian process ($\mathcal{GP}$) prior (Wahba (1990)) with mean function $W\boldsymbol{\beta}_{k,p}$ and covariance matrix $\tau^2_{k,p}\Omega$. Section 5.2 reports inference on the influence of SOI, DMI and SAM on daily rainfall via the spatially varying posterior distributions of the $\boldsymbol{\mu}_{\cdot,k,p}$.

The sth row of the $S \times Q$ matrix W, $\boldsymbol{w}_{s,\cdot}$, contains an intercept plus measurements which identify the sth location, the vector $\boldsymbol{\beta}_{k,p}$ contains the corresponding unknown coefficients, and $\tau^2_{k,p}$ is a smoothing parameter. In the application considered in this article, the location of a site is specified by its latitude and longitude, so that $\boldsymbol{w}_{s,\cdot} = (1, \text{lat}_s, \text{lon}_s)'$ and $Q = 3$. We use the reproducing kernel Hilbert space defined by a two-dimensional thin-plate Gaussian process prior to construct Ω in equation (3.4), expressed with a linear combination of basis functions, as described in Wood (2013) (see also Bertolacci et al. (2019a)). To achieve computational feasibility, we truncate the basis expansion to the first 100 (out of 17,606) basis vectors. In our application, this truncation explains 95% of the variation represented by Ω.

The spatial level of the hierarchical model is completed by placing independent normal priors on $\boldsymbol{\beta}_{k,p}$, and independent inverse gamma (IG) priors on $\tau^2_{k,p}$ and $\sigma^2_{k,p}$, for $p = 1, \ldots, P$ and, $k = 0, 1, \ldots, K$. In particular,

$$\begin{aligned} \boldsymbol{\beta}_{k,p} &\sim N(\mathbf{0}, 100\mathrm{I}_3), \\ \sigma^2_{k,p} &\sim \text{IG}(1.1, 0.5), \\ \tau^2_{k,p} &\sim \text{IG}(1.1, 0.5). \end{aligned} \tag{3.5}$$

3.2. *Temporal variability.* We model all temporal variation through equation (3.2). The covariates include an intercept term, a linear trend over the century for long-term change, a periodic harmonic following the solar year for seasonality and a first-order Markovian structure for short-term daily dependence. The external climate drivers SOI, DMI and SAM, measured monthly, and their interactions, are also included as covariates, because their impact is the main object of scientific interest.

To define the relationship between rainfall on day t and the values of climate drivers in month M, let n_M be the number of days in month M and define $N_M = \sum_{m=1}^{M} n_m$ to be the total number of days from January 31st, 1876 until the end of month M. Define the set of days corresponding to month M to be $D_M = \{t : N_{M-1} < t \le N_M\}$. The values SOI_t, DMI_t, and SAM_t in equation (3.6) are constant and equal to SOI_M, DMI_M, and SAM_M, if day $t \in D_M$.

Let N_t be the number of days since January 31st, 1876, and define year_t to be the number of solar years since January 1st, 1900. Thus, for January 1st, 1900, $\mathrm{year}_t = 0$ and $N_t = 8736$, so that $\mathrm{year}_t = (N_t - 8736)/365.25$. Note that year_t is a fractional number, which is negative for days before January 1st, 1900. For example, for February 28th, 1878, $\mathrm{year}_t = -21.84$, because $N_t = 366 + 365 + 28 = 759$. The systematic component, $\boldsymbol{x}'_{t,s}\boldsymbol{\delta}_{s,k,\cdot}$, governing $\pi_{t,s,k}$ in equation (3.2) becomes

$$
\begin{aligned}
\boldsymbol{x}'_{t,s}\boldsymbol{\delta}_{s,k,\cdot} = {} & \delta_{s,k,1} + \delta_{s,k,2}\mathrm{Trend}_t \\
& + \delta_{s,k,3}\cos(2\pi\,\mathrm{year}_t) + \delta_{s,k,4}\sin(2\pi\,\mathrm{year}_t) \\
& + \delta_{s,k,5}\mathrm{SOI}_t + \delta_{s,k,6}\mathrm{DMI}_t + \delta_{s,k,7}\mathrm{SAM}_t \\
& + \delta_{s,k,8}(\mathrm{SOI}_t \times \mathrm{DMI}_t) \\
& + \delta_{s,k,9}(\mathrm{SAM}_t \times \mathrm{DMI}_t) + \delta_{s,k,10}(\mathrm{SOI}_t \times \mathrm{SAM}_t) \\
& + \delta_{s,k,11}(\mathrm{SOI}_t \times \mathrm{DMI}_t \times \mathrm{SAM}_t) \\
& + \sum_{k'=1}^{K} \delta_{s,k,11+k'} I\big(z_{t-1,s} = k'\big),
\end{aligned}
\tag{3.6}
$$

where $\mathrm{Trend}_t = \mathrm{year}_t/100$ is a linear trend, measured over a century.

3.3. *Mixture components model.* We assume that daily rainfall is a mixture of a point mass at zero, for days with zero rainfall, and K gamma distributions, for days with rain. Specifically,

$$
y_{t,s} \sim \pi_{t,s,0}\delta(0) + \sum_{k=1}^{K} \pi_{t,s,k}\mathrm{Ga}(a_k, b_k),
\tag{3.7}
$$

where $\delta(0)$ is a Dirac delta function at 0, and $\mathrm{Ga}(a_k, b_k)$ is a gamma distribution with density $f_k(y) \propto y^{a_k-1}e^{-y/b_k}$ and support $y > 0$. Model (3.7) is similar to the

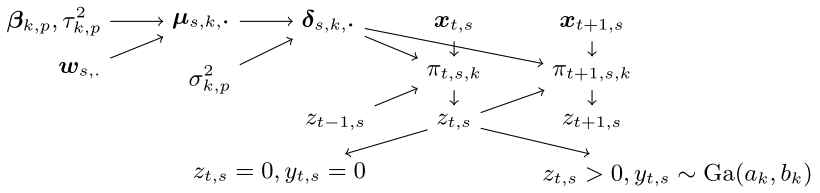


FIG. 3. *A directed acyclic graph representation of the dependence between the model inputs, parameters, and outputs for the Australian daily rainfall application.*

models of Holsclaw et al. (2016, 2017), who use a mixture of one exponential and one gamma component (Holsclaw et al. (2016)), and a mixture of two exponentials (Holsclaw et al. (2017)). To address label switching, we order the means of the gamma components by $a_k b_k < a_{k'} b_{k'}$ for $k < k'$. This ordering also leads to meaningful inference. For example, if $K = 2$, equation (3.7) specifies one state for zero-rainfall days and two states for the nonzero rainfall days, which may be interpreted as no, low and high rainfall days. If $K = 3$, the components may be interpreted as no, low, medium and high rainfall days. Figure 3 displays a graphical summary of the model for the Australian daily rainfall described in equations (3.1)–(3.7).

The hierarchy is completed by placing independent priors on the gamma density parameters (a_k, b_k). Damsleth (1975) shows that a conjugate class of prior distributions for i.i.d. gamma random variables results from $b_k \sim \mathrm{IG}(u, v)$ and $p(a_k|b_k) \propto \rho^{a_k - 1}[b_k^{a_k q}\Gamma(a_k)^r]^{-1}$. In our case, this is useful because the full conditional distributions of (a_k, b_k) can now be sampled without the need for tuning parameters. The values of u, v, ρ, q, r are chosen to be noninformative; for details, see Bertolacci et al. (2019a).

4. Computing. We now outline the MCMC scheme and show how we exploit the structure of equations (3.1)–(3.4) to construct a parallel and distributed algorithm that efficiently estimates the model parameters.

4.1. *MCMC algorithm.* As mentioned in Section 2, there are many days for which daily rainfall measurements are missing. For each site s, denote by T_s^{mis} the set of indices for the missing values, by $y_{t,s}^{\mathrm{mis}}$ (collectively $\boldsymbol{y}_s^{\mathrm{mis}}$) the missing observations, by $y_{t,s}^{\mathrm{obs}}$ (collectively $\boldsymbol{y}_s^{\mathrm{obs}}$) the observed values, and let $\boldsymbol{y}_s$ be the $T \times 1$ vector $\boldsymbol{y}_s = (\boldsymbol{y}_s^{\mathrm{mis}'}, \boldsymbol{y}_s^{\mathrm{obs}'})'$ containing, temporally ordered, all missing and observed values.

Some further notation is required to describe the algorithm. Let $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_S')'$, $\boldsymbol{z}_s = (z_{1,s}, \ldots, z_{T,s})'$, $\boldsymbol{z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_S')'$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_0', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')'$, $\boldsymbol{\mu}_{\cdot,k,p} = (\mu_{1,k,p}, \ldots, \mu_{S,k,p})'$, $\boldsymbol{\sigma}^2_{k,\cdot} = (\sigma^2_{k,1}, \ldots, \sigma^2_{k,P})'$, $\boldsymbol{\delta}_{\cdot,k,p} = (\delta_{1,k,p}, \ldots, \delta_{S,k,p})'$, and let $\Delta_{s,\setminus\{k\},\cdot}$ be the matrix $\Delta_{s,\cdot,\cdot}$ with the kth column omitted. Then, initializing $\boldsymbol{\theta}$, $\boldsymbol{z}$ and the missing values, the MCMC scheme iteratively draws from the following distributions (details are provided in Bertolacci et al. (2019a)).

1. $p(\boldsymbol{\theta}_k|\boldsymbol{y}, \boldsymbol{z})$, for $k = 0, 1, \ldots, K$;
2. For $s = 1, \ldots, S$
   2(a) $p(\boldsymbol{z}_s|\boldsymbol{y}_s^{\text{obs}}, \Delta_{s,\cdot,\cdot}, \boldsymbol{\theta})$;
   2(b) $p(y_{t,s}^{\text{mis}}|\boldsymbol{\theta}, \boldsymbol{z}_s)$ for $t \in T_s^{\text{mis}}$;
   2(c) $p(\boldsymbol{\delta}_{s,k,\cdot}|\boldsymbol{z}_s, \boldsymbol{\mu}_{s,k,\cdot}, \boldsymbol{\sigma}_{k,\cdot}^2, \Delta_{s,\setminus\{k\}}, \cdot)$, for $k = 0, 1, \ldots, K$;
3. For $k = 0, 1, \ldots, K$ and $p = 1, \ldots, P$
   3(a) $p(\sigma_{k,p}^2|\boldsymbol{\delta}_{\cdot,k,p}, \boldsymbol{\mu}_{\cdot,k,p})$;
   3(b) $p(\boldsymbol{\mu}_{\cdot,k,p}|\boldsymbol{\delta}_{\cdot,k,p}, \boldsymbol{\beta}_{k,p}, \sigma_{k,p}^2, \tau_{k,p}^2)$;
   3(c) $p(\boldsymbol{\beta}_{k,p}|\boldsymbol{\mu}_{\cdot,k,p}, \tau_{k,p}^2)$;
   3(d) $p(\tau_{k,p}^2|\boldsymbol{\mu}_{\cdot,k,p}, \boldsymbol{\beta}_{k,p})$.

Step 2(a) is implemented using the forward-backward Gibbs step proposed by Chib (1996) and is conditioned only on the observed values $\boldsymbol{y}_s^{\text{obs}}$ to avoid an absorbing state. Drawing from $p(\boldsymbol{\delta}_{s,k,\cdot}|\boldsymbol{z}_s, \boldsymbol{\mu}_{s,k,\cdot}, \boldsymbol{\sigma}_{k,\cdot}^2, \Delta_{s,\setminus\{k\}}, \cdot)$ in 2(c) could be performed via a Metropolis–Hastings step, but instead we use the data augmentation approach of Polson, Scott and Windle (2013), which greatly simplifies the sampler, given the size and complexity of our data.

4.2. *Parallel implementation and running time analysis*. The dataset described in Section 2 contains around 300 million measurements and takes about 16 GB of disk space. For the case where $K = 2$, the matrix $\Delta_{s,\cdot,\cdot}$ contains 26 entries at each s, yielding 457,756 parameters, and there are about 3000 more parameters to be estimated, corresponding to the higher levels of the hierarchical model. Additionally, each measurement has $K + 1$ latent variables associated with it: a $z_{t,s}$ for the current state, and K latent variables for the Pólya–Gamma-based sampling scheme mentioned in the previous section, resulting in about 900 million latent variables in total. The time required to sample these random variables is prohibitive. For example, if $K = 2$, 40,000 iterations of the algorithm in standard serial computing would take about 400 days on a modern processor. If $K = 3$, 40,000 iterations are estimated to take about 600 days, but in practice it takes even longer, because we require 60,000 iterations to ensure convergence and adequate mixing. To make the estimation feasible, we exploit the model structure to create a distributed implementation of the algorithm. As a result, eight hours are required for $K = 2$, and 17 hours for $K = 3$.

The degree to which a parallel algorithm affords a speed-up over a serial algorithm depends on the proportion of the total running time that can be split between available processing units (Amdahl (1967)). When an algorithm is running in a distributed manner across multiple computing nodes, an additional significant consideration is the time spent communicating between the nodes, which is mostly determined by the total size of the messages sent (Lynch (1996)). Here, $y_{t,s}^{\text{mis}}$, $z_{t,s}$ and $\boldsymbol{\delta}_{s,k,\cdot}$ are conditionally independent across dimension s, and so can be sampled simultaneously. Even on a single computer with multiple processor cores,

this speeds up the sampling in proportion to the number of cores. To distribute the algorithm, the dimension s can be split between computing nodes so that each node sees only a fraction of the total data and must hold and sample only a fraction of the 900 million latent parameters. As we describe below, communication between the nodes is needed only to sample the parameters in the higher levels of the hierarchical model, so the size of the messages is proportional only to the number of sites rather than to the number of time periods. These two ideas jointly yield an efficient distributed parallel implementation of the algorithm, which we describe below.

To parallelize the sampling scheme, let $N \geq 1$ be the number of computing nodes with C cores each, designate node 1 the master, and divide the sites s evenly between the N nodes. Then the sampler runs in parallel as follows.

- Each node performs steps 2(a) through 2(c) on each of the sites s allocated to it, dividing the sites among the C cores within a node. Once complete, each node sends node 1 its new samples of $\Delta_{s,\cdot,\cdot}$ and the summary statistics $\sum_{t=1}^{T} I(z_{t,s} = k) y_{t,s}$ and $\sum_{t=1}^{T} I(z_{t,s} = k) \log y_{t,s}$ for each $k = 1, \ldots, K$.
- Node 1 performs step 1 using the summary statistics and the conditional distributions of the component parameters, and steps 3(a) through 3(d) (the latter in parallel for each pair (k, i) across the C cores of node 1) for the upper levels of the hierarchy using the samples $\Delta_{s,\cdot,\cdot}$. The results are sent as messages to the other $N - 1$ nodes.

We now analyze the runtime and message size per iteration of this algorithm. Suppose $Q' = Q + B$, where B is the number of basis vectors chosen for the GP regression as described in Section 3.1 ($B = 100$ in this work). Then, with the above parallelizations, and omitting the time taken for step 1, the running time of an iteration of the sampler is of order

$$O\left(\frac{S}{NC}\left(K^2 T + K P^2 T + K P^3\right) + \frac{K P}{C}\left(Q' S + Q'^3\right)\right). \tag{4.1}$$

The first term corresponds to steps 2(a) through 2(c) and the second to steps 3(a) through 3(d). The total size of the messages sent is proportional to $K P S$, which is the number of $\delta_{s,k,i}$ parameters. The key features of (4.1) are that both the time and number of messages are linear in the number of time periods T and the number of sites S, and that, for the terms proportional to T, a linear speedup can be achieved by using N computing nodes.

The implementation is written as an R package in the R language and in C++ (using RcppArmadillo), and is distributed across multiple computing nodes using the Rmpi package (R Core Team (2016), Eddelbuettel and Sanderson (2014), Yu (2002)). Multithreading is implemented with OpenMP.[14] The R package for estimating the parameters of the model presented in this article is available online.[15]

[14] http://openmp.org/wp/.

[15] https://github.com/mbertolacci/storm/.

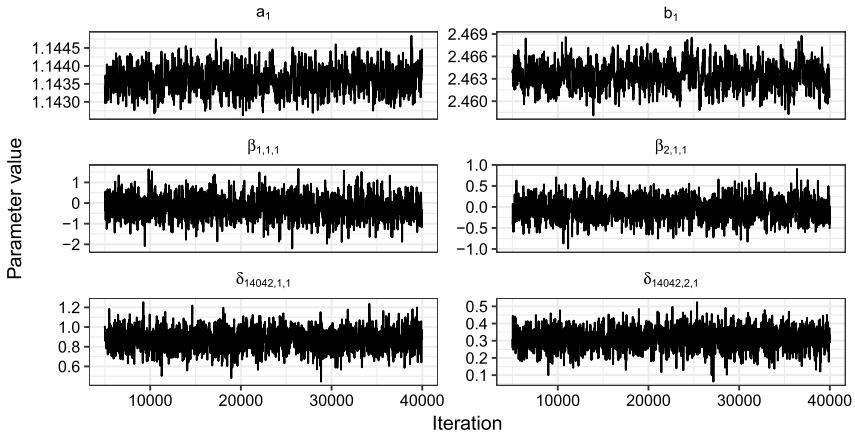


FIG. 4. *Trace plots for a selection of parameters from the MCMC samples giving rise to the results in Section 5.*

Using the sampling scheme in Section 4.1, 40,000 posterior samples corresponding to the model with $K = 2$ were generated using a Cray XC40 supercomputer with nodes of 64 GB of RAM, with two 2.6 GHz Intel Xeon E5-2690 v3 CPUs with 12 cores each for a total of 24 cores per node. We used 52 nodes with 339 sites per node, for a total of 1248 cores. It took around eight hours to generate the 40,000 samples, the first 5000 of which were discarded as burn-in. Convergence was assessed using trace plots, a selection of which is shown in Figure 4.

5. Results. We present the results for the model with a point mass at zero and $K = 2$ gamma components, which is interpretable as corresponding to zero, light and heavy rainfall days. Figure 5 displays the posterior mean density estimates of the two gamma components and shows that one of the densities (f_1) has probability mass concentrated between 0 and 6 mm, while the other has higher probability mass for daily rainfall greater than 6 mm. Section 5.1 assesses the model performance for $K = 2$ by its posterior predictive coverages (PPC), predictive quantiles and daily temporal dependencies. Section 5.2 reports posterior inference regard-

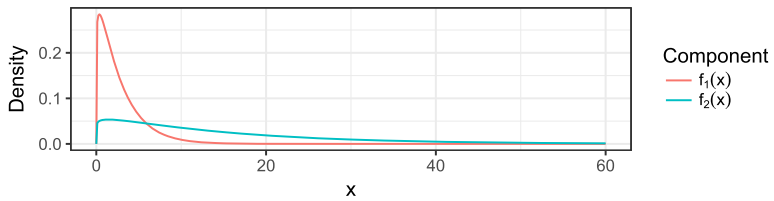


FIG. 5. *Posterior density estimates for $K = 2$ gamma components show clear separation. The posterior mean estimates (and their standard errors), of the density parameters are: $\hat{a}_1 = 1.144$ (0.0004), $\hat{b}_1 = 2.463$ (0.0017) and $\hat{a}_2 = 1.100$ (0.0005), $\hat{b}_2 = 14.13$ (0.006).*

ing the association of the oceanic/atmospheric interactions with Australian daily rainfall, as well as its long term trend.

As indicated in Section 1, much work has been done investigating the types and number of mixture components required to adequately capture the tail behavior of temporally evolving daily rainfall. For inferential interpretability, the main article reports results on $K = 2$ gamma components, but the model is easily extended to more components, and Bertolacci et al. (2019b) present results corresponding to $K = 3$ which show no significant improvement in PPC.

5.1. *Model performance.* The model's performance is evaluated by comparing the observed data to the posterior predictive distribution

$$p(\boldsymbol{y}_s^*|\boldsymbol{y}^{\text{obs}}) = \int p(\boldsymbol{y}_s^*|\boldsymbol{y}^{\text{obs}}, \Delta_{s,\cdot,\cdot}, \boldsymbol{\theta}) p(\Delta_{s,\cdot,\cdot}, \boldsymbol{\theta}|\boldsymbol{y}^{\text{obs}})\, d(\Delta_{s,\cdot,\cdot}, \boldsymbol{\theta}), \tag{5.1}$$

where $\boldsymbol{y}_s^*$ denotes a predicted value and $\boldsymbol{y}^{\text{obs}} = (\boldsymbol{y}_1^{\text{obs}'}, \ldots, \boldsymbol{y}_S^{\text{obs}'})'$. Conditional on the latent component indicators, the first expression in the integrand of equation (5.1) can be evaluated as follows:

$$p(\boldsymbol{y}_s^*|\boldsymbol{y}^{\text{obs}}, \Delta_{s,\cdot,\cdot}, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}_s^*} \left(\prod_{t=1}^{T} f_{z_{t,s}^*}(y_{t,s}^*|\boldsymbol{\theta}_{z_{t,s}^*}) p(z_{t,s}^*|z_{t-1,s}^*, \Delta_{s,\cdot,\cdot}) \right),$$

where the sum is taken over all possible indicator vectors $\boldsymbol{z}_s^*$. The MCMC scheme of Section 4 is used to draw samples from equation (5.1) and obtain PPC of monthly rainfall. In particular, equation (5.1) is approximated by

$$\frac{1}{L} \sum_{j=1}^{L} p(\boldsymbol{y}_s^*|\boldsymbol{y}^{\text{obs}}, \Delta_{s,\cdot,\cdot}^{[j]}, \boldsymbol{\theta}^{[j]}),$$

where $\Delta_{s,\cdot,\cdot}^{[j]}$ and $\boldsymbol{\theta}^{[j]}$ are the jth draws from $p(\boldsymbol{\theta}, \Delta_{s,\cdot,\cdot}|\boldsymbol{y}^{\text{obs}})$, and L is the number of draws after burn-in.

5.1.1. *Posterior predictive coverage (PPC) of monthly rainfall.* For month M and site s, let $\bar{y}_{M,s}$ be the empirical average rainfall, that is, the average of recorded rainfall in month M at site s, and let $\bar{y}_{M,s}^0$ be the proportion of days with a recorded rainfall of zero in month M at site s. Let $\hat{p}(\bar{y}_{M,s}^*|\boldsymbol{y}^{\text{obs}})$ be the estimated posterior predictive distribution of average rainfall and let $\hat{p}(\bar{y}_{M,s}^{0*}|\boldsymbol{y}^{\text{obs}})$ be the estimated posterior predictive distribution of the proportion of days without rain.

We define the $100(1-\alpha)\%$ PPC interval for site s in month M to be the interval $[\bar{y}_{\alpha/2}^*, \bar{y}_{1-\alpha/2}^*]$ which satisfies $\hat{p}(\bar{y}_{\alpha/2}^* < \bar{y}^*|\boldsymbol{y}^{\text{obs}}) = 1-\alpha/2$ and $\hat{p}(\bar{y}_{1-\alpha/2}^* < \bar{y}^*|\boldsymbol{y}^{\text{obs}}) = \alpha/2$, with intervals for $\bar{y}^{0*}$ defined analogously. These quantities vary across sites, s, and months, M, but the subscripts s and M are suppressed for clarity. However, due to the large number of zero-rainfall days in the dataset, the

definition of these intervals must be modified slightly. If $\bar{y}^*_{\alpha/2} = 0$ for a particular site s and month M, we take the $100(1-\alpha)\%$ PPC interval to be $[\bar{y}^*_0, \bar{y}^*_{1-\alpha}]$. Similarly, if $\bar{y}^{0*}_{1-\alpha/2} = 1$, we take the interval to be $[\bar{y}^{0*}_{\alpha}, \bar{y}^{0*}_1]$.

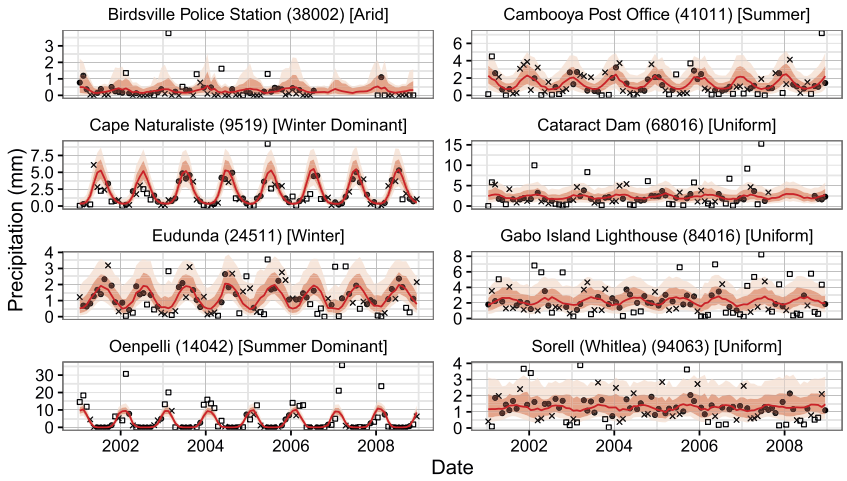
Figure 6 illustrates the model's ability to capture the marginal temporal variability of Australian daily rainfall. Figure 6(a) plots pointwise quantiles of the posterior predictive distribution $\hat{p}(\bar{y}^*_{M,s}|\boldsymbol{y}^{\text{obs}})$ for the eight sites discussed in Sections 1 and 2. In particular, we plot the rainfall, $\bar{y}^*_{0.5}$ (red solid line), and the band consisting of PPC intervals $[\bar{y}^*_{\alpha/2}, \bar{y}^*_{1-\alpha/2}]$, corresponding to $\alpha = 0.5$ (shaded dark pink band), and $\alpha = 0.2$ (shaded light pink band), against time, along with the observed data. Figure 6(b) displays the analogous plots for $\bar{y}^{0*}_{M,s}$.

These plots illustrate that the model given in equation (3.7), which incorporates the dependence of time and other covariates via the mixture weights, can accommodate spatially varying rainfall patterns. For example, the time series plots of Cape Naturaliste and Oenpelli show that the model can accommodate phase shifts in seasonality (winter dominance vs. summer dominance). Cape Naturaliste (classified as Winter Dominant rainfall) reports high probabilities of zero rainfall/low monthly averages in the summer months and vice-versa for winter months. Oenpelli (classified as Summer Dominant rainfall) reports high probabilities of zero rainfall/low monthly averages in the winter months and vice-versa for summer months. Sites classified as Uniform rainfall (Cataract Dam, Gabo Island Lighthouse and Sorell) show less pronounced seasonality.

Figure 7 illustrates the predictive performance of the model across all sites. The top panel displays histograms of the proportion of observed monthly averages for site s which lie in the interval $[\bar{y}^*_{0.25,s}, \bar{y}^*_{0.75,s}]$, (left) and $[\bar{y}^*_{0.1,s}, \bar{y}^*_{0.9,s}]$, (right), for sites $s = 1, 2, \ldots, 17{,}606$. To emphasize the dependence of these intervals on the site we have included the subscript s. These histograms show that the model-based quantiles of $\bar{y}^{\text{obs}}_{M,s}$ agree with the data; the average across all sites and months of the proportion of observed monthly averages, which lie in the model-based PPC 50% interval, is also 50% (left), and the corresponding proportion for the PPC 80% interval is 78% (right). The bottom panel displays analogous histograms for the proportion of dry days, $\bar{y}^{0,\text{obs}}_{M,s}$ which fall within the PPC 50% interval (left) and 80% interval (right). The average proportion of values which lie within the PPC intervals is 56% for the PPC 50% intervals, so that these intervals are only slightly wide on average, and 78% for the PPC 80% intervals, so that these intervals are only slightly narrow on average.

Finally, Figure 8 shows the estimated predictive mean of the probability that daily rainfall belongs to components $k = 0, 1$ and 2 (i.e., $p(\pi^*_{t,s,k}|\boldsymbol{y}^{\text{obs}})$) for the eight illustrative sites. Figure 8 demonstrates that the model captures expected heterogeneity on many aspects: opposite peak rainfall times for the Summer Dominant and Winter Dominant sites, and the appropriate weighting of the components at different sites and different times. For example, the model assigns higher weights to the heavy rainfall component, $k = 2$, for the tropical site Oenpelli than

(a) *Empirical monthly average rainfall* ($\bar{y}_{M,s}$).



(b) *Monthly proportion of days without rain* ($\bar{y}^0_{M,s}$).

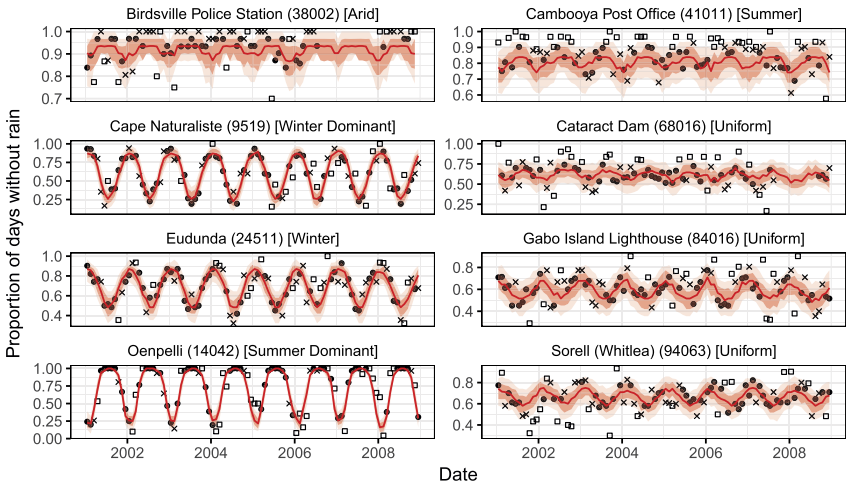


FIG. 6. *Plots of the posterior predicted median rainfall and proportion of days without rain,* $\bar{y}^*_{0.5}$ *and* $\bar{y}^{0*}_{0.5}$ (*red solid line*), *PPC intervals* $[\bar{y}^*_{0.25}, \bar{y}^*_{0.75}]$ *and* $[\bar{y}^{0*}_{0.25}, \bar{y}^{0*}_{0.75}]$ (*shaded dark pink band*), *and* $[\bar{y}^*_{0.1}, \bar{y}^*_{0.9}]$ *and* $[\bar{y}^{0*}_{0.1}, \bar{y}^{0*}_{0.9}]$ (*shaded light pink band*), *against time, along with the observed data for the eight sites, based on* $K = 2$. *Open squares mark observed data values which lie outside the* 80% *PPC band, crosses mark those inside the* 80% *PPC band but outside the* 50% *band, and filled circles mark those within the* 50% *band.*

for the Winter Dominant site Cape Naturaliste, higher weights for both the light and heavy components, indicating more rainfall in the appropriate seasons for sites with seasonally varying rainfall, and variation among years due to the climate co-

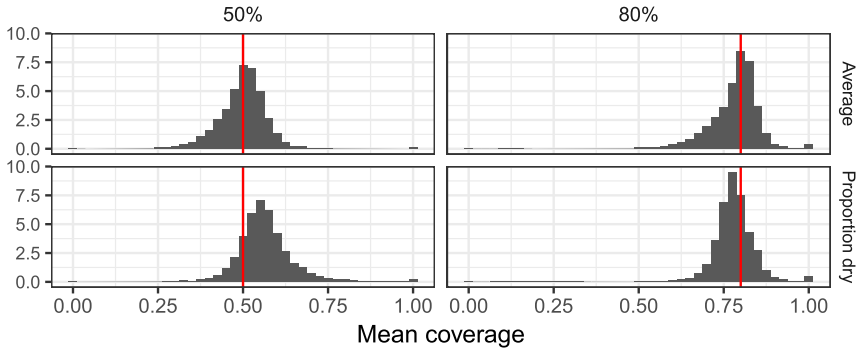


FIG. 7. *The proportion of observed monthly averages,* $\bar{y}^{\text{obs}}_{M,s}$, *(top panel), which fall within the PPC* 50% *interval (left) and* 80% *interval (right) for all* 17,606 *sites, and the proportion of observed monthly proportion of dry days,* $\bar{y}^{0\text{obs}}_{M,s}$ *(bottom panel), which fall within the PPC* 50% *interval (left) and* 80% *interval (right), for the model with* $K = 2$.

variates. These results, combined with accurate overall PPC of rainfall intensity and the probability of zero rainfall discussed above, indicate good model performance.

5.1.2. *Predictive quantiles of nonzero daily rainfall.* In this section, we examine the agreement between empirical quantiles of nonzero rainfall and quantiles obtained from our model's predictive densities. Metrics are computed for each of

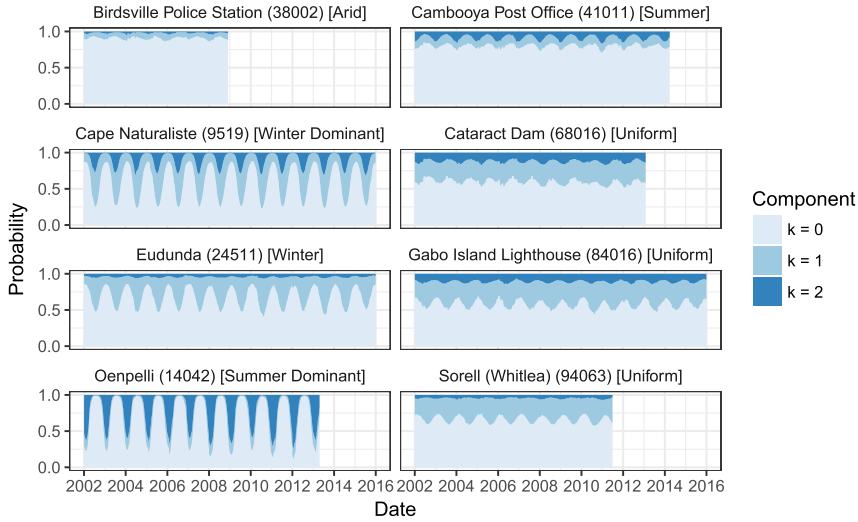


FIG. 8. *Mean predictive probability of being in each component for each day for the model with* $K = 2$. *The zero rainfall component is* $k = 0$, *and* $k = 1, 2$ *are the small and large rainfall components, respectively.*

March–May (autumn), September–November (spring), December–February (summer) and June–August (winter), using the sites that have at least 100 days of nonzero rainfall in that season. Let $\hat{q}_{p,\mathcal{T}}(s)$ and $q^*_{p,\mathcal{T}}(s)$ be, respectively, the empirically estimated and posterior predictive pth quantile at site s in season $\mathcal{T}$. As a measure of agreement between $\hat{q}_{p,\mathcal{T}}(s)$ and $q^*_{p,\mathcal{T}}(s)$, we define

$$D_{p,\mathcal{T}}(s) = \hat{E}\left[\log \frac{q^*_{p,\mathcal{T}}(s)}{\hat{q}_{p,\mathcal{T}}(s)} \Big| y^{\text{obs}}\right], \tag{5.2}$$

the estimated posterior predictive mean of the log ratio of $q^*_{p,\mathcal{T}}(s)$ and $\hat{q}_{p,\mathcal{T}}(s)$. When $D_{p,\mathcal{T}}(s)$ is zero, the posterior predictive quantile matches the empirical one, while if the model over/underestimates the empirical quantile, $D_{p,\mathcal{T}}(s)$ is positive/negative, with larger absolute values indicating poorer performance.

For each $\mathcal{T}$ = autumn, spring, summer, and winter and each percentile $p = 1\%, 2\%, \ldots, 99\%$, Figure 9 shows the median value, $\tilde{D}_{p,\mathcal{T}}$, and the interquartile range (IQR) of $D_{p,\mathcal{T}}(s)$ across all sites. We define nominal performance for the percentile and season pair $(p, \mathcal{T})$ as the IQR of $D_{p,\mathcal{T}}(s)$ containing zero. For all seasons, performance is nominal for $7\% \le p \le 86\%$. For autumn and summer, performance is also nominal when $p > 86\%$. The percentiles corresponding to $p > 86\%$ in spring and winter are overestimated, and the degree of overestimation increases towards $p = 99\%$. In autumn, spring, summer and winter, the percentiles corresponding to $p < 5\%, 7\%, 6\%$ and 6%, respectively, are underestimated, and the degree of underestimation increases towards $p = 1\%$. To investigate the effect of the number of components on the model fit in the tails of the distribution, the performance for $K = 3$ is reported in the supplementary material (Bertolacci et al. (2019b)). The addition of another component expands the range of percentiles showing nominal performance in all seasons to $4\% \le p \le 96\%$.

For illustrative purposes, we examine Q–Q plots of the empirical quantiles of nonzero daily rainfall for summer and winter against quantiles obtained from the predictive densities at the eight illustrative sites. Figure 10 presents these Q–Q plots, where top rows correspond to summer months and bottom rows to winter months. As in the case of $D_{p,\mathcal{T}}(s)$, percentiles up to around 90% show good agreement for most sites. In Figure 9, the lowest percentiles were underestimated, but this is difficult to see in the Q–Q plots, which suggests that, on the scale of millimeters, underestimation of the low percentiles is small. For the 90% and greater percentiles, the three sites that show the worst performances are: summer (tropical wet season) rainfall at Oenpelli is underestimated, winter (wet season) rainfall at Cape Naturaliste is overestimated and summer and winter rainfall at Cataract Dam are underestimated. Bertolacci et al. (2019b) present Q–Q plots for the model with $K = 3$; the addition of another component improves the estimation, especially at Oenpelli, although for the other seven sites, the improvement from increasing K from 2 to 3 is marginal.

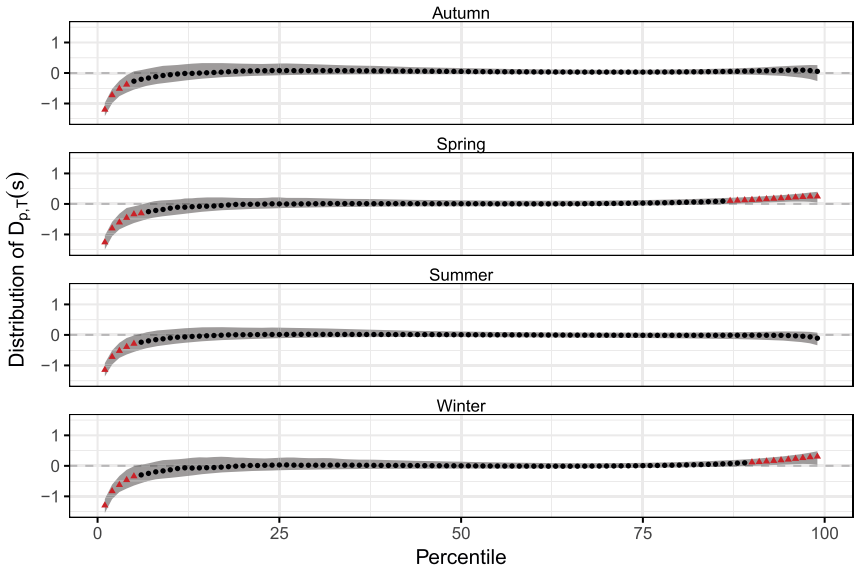


FIG. 9. *Median values (points) and interquartile range (shaded band) of $D_{p,\mathcal{T}}(s)$ across all sites, s, for each $\mathcal{T}$ = autumn, spring, summer, and winter (top panel to bottom panel, respectively) and each $p = 1\%, 2\%, \ldots, 99\%$, for the model with $K = 2$. A black circle indicates that the corresponding percentile/season is nominal (i.e., the IQR of $D_{p,\mathcal{T}}(s)$, $s = 1, \ldots, S$, contains zero), while a red triangle indicates that it is not.*

Taken together, Figures 9 and 10 show that $K = 2$ performs well across the majority of sites, but that the tail of the distribution at some sites, such as Oenpelli and Cape Naturaliste, remain difficult to estimate. While $K = 3$ improves the model's ability to capture tail behavior, future work into $K > 3$ or the inclusion of heavier tailed distributions (such as the Generalized Pareto or Extended Generalized Pareto distributions) would be of interest. For the application in this paper, Bertolacci et al. (2019c) presents a simulation study demonstrating that the model's ability to perform climate inference is not compromised by the presence of excessively heavy-tailed data.

5.1.3. *Daily temporal dependencies.* This section presents metrics of daily temporal dependence for the data and the model fits. The dataset contains substantial dependencies in time, as well as substantial variation in the strengths of those dependencies across sites. The left-hand panel of Figure 11 shows a histogram of the empirical first-order autocorrelations of rainfall amount at the daily level, calculated using Spearman's rank correlation, and denoted by $R_s(1)$ for site s, for each of the 17,606 sites.

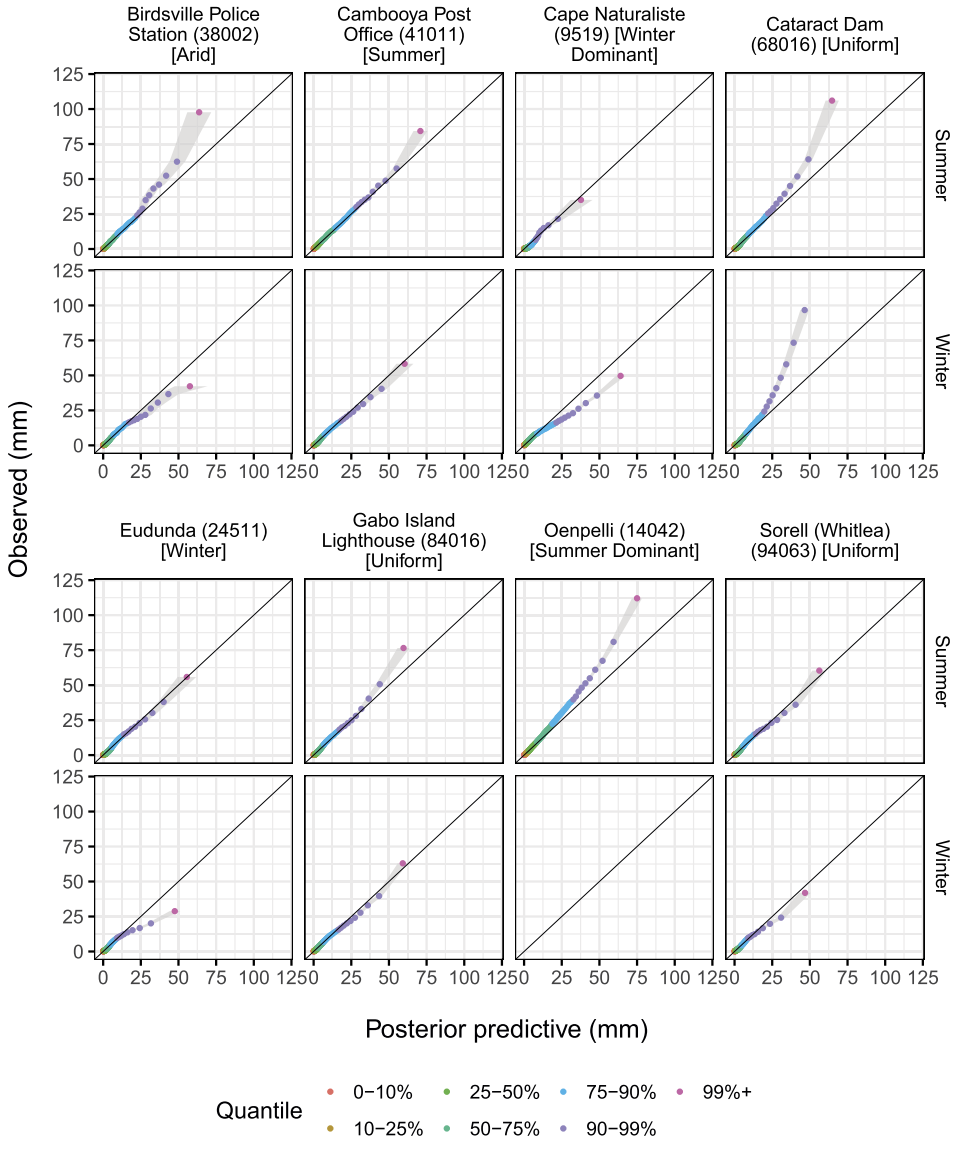


FIG. 10. *Q–Q plots for days with nonzero rainfall for summer (top rows) and winter (bottom rows) for the model with* $K = 2$. *80% posterior intervals for each quantile shown by grey shaded bands. Oenpelli in winter is omitted as it has had fewer than* 100 *days with nonzero rainfall.*

The log odds that subsequent days within a site have the same rainfall occurrence are defined as

$$\mathcal{LO}_s(1) = \log \frac{p(y_{t,s} = 0, y_{t+1,s} = 0) + p(y_{t,s} > 0, y_{t+1,s} > 0)}{p(y_{t,s} = 0, y_{t+1,s} > 0) + p(y_{t,s} > 0, y_{t+1,s} = 0)}, \tag{5.3}$$

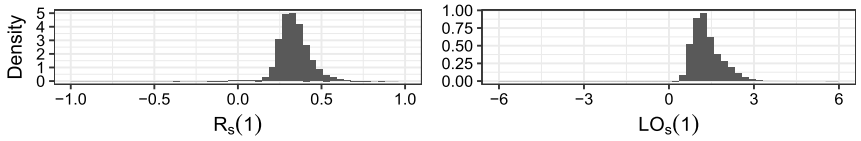


FIG. 11. *Histograms of the estimated Spearman autocorrelation of rainfall amount for each site,* $R_s(1)$ (*left*), *and the log odds of having the same rainfall occurrence on subsequent days* $\mathrm{LO}_s(1)$ (*right*). *Metrics are for each site, measured between the pairs* $(y_{t,s}, y_{t+1,s})$.

and are estimated using empirical counts of each event for each site (see Charles, Bates and Hughes (1999)). The right-hand panel of Figure 11 displays a histogram of the resulting empirical log odds. The histograms in Figure 11 show that the vast majority of sites have positive dependence between days, and that the strength of the dependence varies from site to site. These dependencies may have many components, including seasonal and climate factors, the influence of climate drivers, and local atmospheric effects that persist between days.

Using the posterior predictive distribution, we assess the model's ability to capture the temporal dependencies shown in Figure 11. Using samples from the posterior predictive distribution of equation (5.1), we estimate the posterior predictive distributions of $R_s(1)$ and $\mathrm{LO}_s(1)$, denoted by $\hat{p}(\rho_s^*(1)|\boldsymbol{y}^{\mathrm{obs}})$ and $\hat{p}(\mathcal{LO}_s^*(1)|\boldsymbol{y}^{\mathrm{obs}})$, respectively. These posterior predictive estimates are plotted against the empirical estimates in Figure 12. The posterior median estimates for Spearman's autocorrelation of rainfall amount, $R_s(1)$ (left plot), broadly match the empirical estimates, with most points close to the diagonal. For the log odds of having the same rainfall occurrence, $\mathrm{LO}_s(1)$ (right plot), the points are again close to the diagonal. These

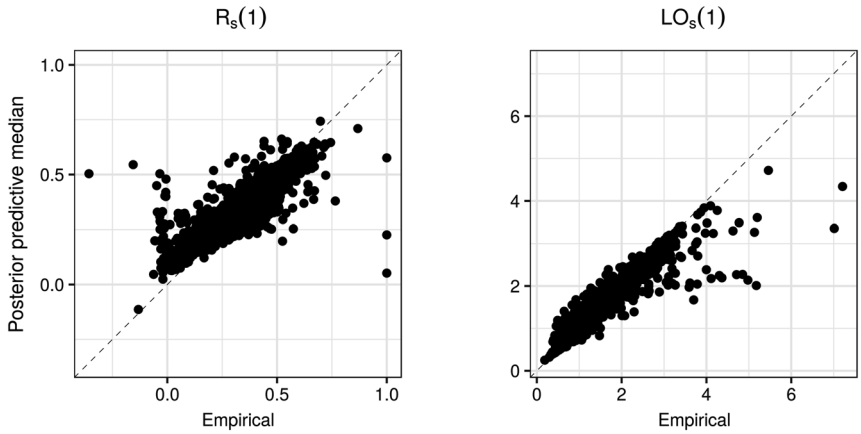


FIG. 12. *Scatter plots of posterior predictive medians versus empirical estimates for Spearman's autocorrelation* $R_s(1)$ (*left*) *and the log odds of having the same rainfall occurrence on subsequent days* $\mathrm{LO}_s(1)$ (*right*) *for each site.*

plots suggest that the first-order Markovian structure of equation (3.6), along with the covariates, are sufficient to capture within-site temporal dependencies.

5.2. *Climate inference*. As mentioned in Section 1, explicit temporal dependencies are present at the observation level of the model hierarchy, including trends, seasonality, covariates and short-term dependencies, but for the sake of parsimony and inferential interpretability, spatial dependence is modeled only through the mixture weights' parameters—not at the observational level. Section 5.1.3 provides metrics indicating that the temporal dependencies of daily rainfall at individual sites are modeled satisfactorily. Bertolacci et al. (2019c) presents similar metrics for contemporaneous spatial dependencies, which are, as expected, not fully captured by the model. Although capturing these dependencies is not the purpose of this work, their presence could potentially impact the estimation of the influence of climate drivers. However, Bertolacci et al. (2019c) demonstrate via a simulation study that the model is able to detect statistically significant and nonsignificant effects of external covariates in the presence of spatially correlated data.

In this section, we first examine the long-term trend of Australian daily rainfall over the past 139 years. Second, we examine the dependence between the oceanic and atmospheric interactions, as encoded by the covariates SOI_t, DMI_t and SAM_t, and Australian daily rainfall, as well as the way in which this dependence varies across the Australian continent.

We examine the spatially varying long-term evolution in the distribution of daily rainfall in Australia by looking at changes in the mixture weights. Equation (3.6) describes the spatially varying parameters in the multinomial logit weights, where $\delta_{s,k,2}$ corresponds to Trend_t. Positive values of $\delta_{s,k,2}$ indicate that the mixture weight assigned to component $k = 1, 2$ has increased over time relative to the zero-rainfall component, $k = 0$ (recall that $\boldsymbol{\delta}_{s,0,\cdot} \equiv \mathbf{0}$). We present our inference on $\delta_{s,k,2}$ via maps of the estimated posterior 10%, 50% and 90% quantiles of their means, $\mu_{s,k,2}$, across the entire spatial field of Australia, evaluated using a 60×60 grid. Thus, when both the 10% and 90% maps at a given location are blue (or red), zero is not included in the corresponding 80% credible interval, in which case we say that $\mu_{s,k,2}$ is significantly different from zero. On the other hand, a change from red to blue (or blue to red) between the 10% and 90% maps at the same location, indicates lack of significance.

Figure 13 shows maps of these estimated posterior quantiles. The upper row corresponds to $\mu_{s,1,2}$, where the left, middle and right maps display the 10%, 50% and 90% quantiles, respectively. The middle and bottom rows are analogous maps for $\mu_{s,2,2}$ and $\mu_{s,2,2} - \mu_{s,1,2}$. The top row, $k = 1$, shows that the low-rainfall component of the distribution of daily rainfall in the southwest and northeast corners of Australia has significantly decreased over time relative to the zero-rainfall component, while in the northwest of Australia it has significantly increased. The middle row shows that the heavy-rainfall component of daily rainfall in the southwestern corner of Australia and the mid-northeast coast of Australia has decreased,

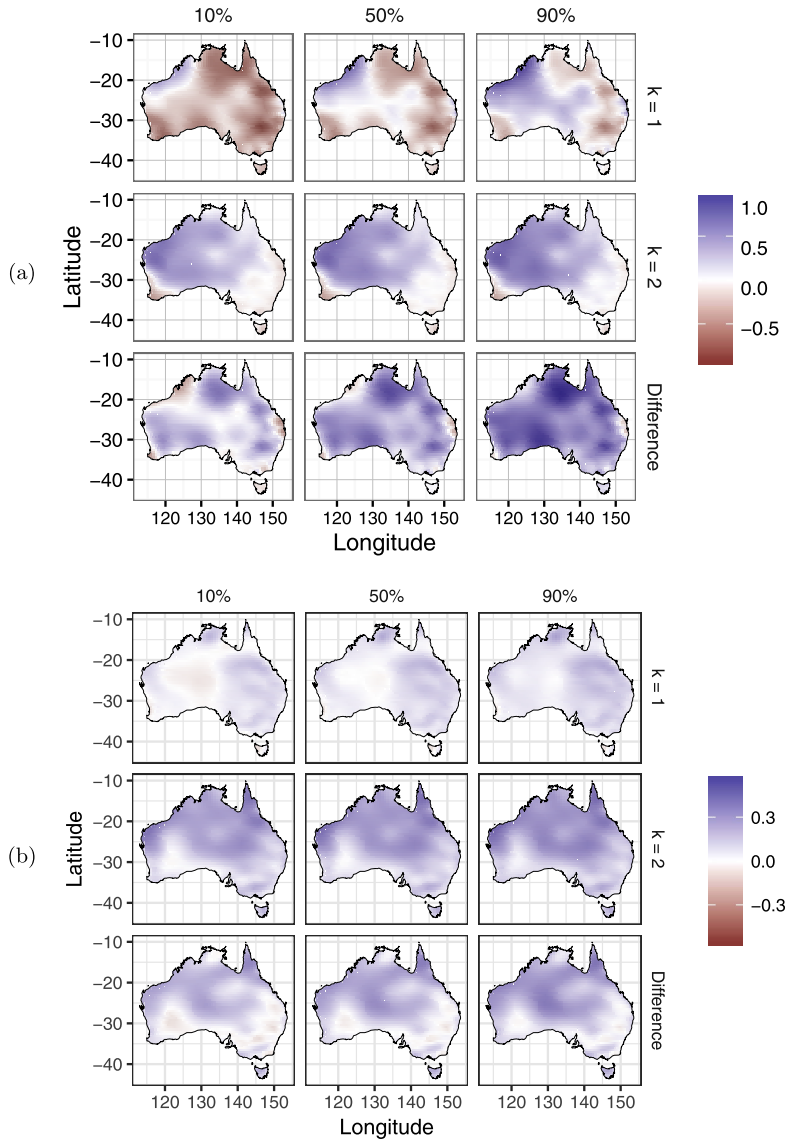


FIG. 13. *Maps of posterior* 10%, 50% *and* 90% (*columns*) *quantiles of* $\mu_{s,k,p}$ ($k = 1, 2$ *in* 1*st and* 2*nd row, respectively, of each panel*) *and of* $\mu_{s,2,p} - \mu_{s,1,p}$ (*bottom row of each panel*) *for the model with* $K = 2$, *where* $p = 2$ (Trend_t) *in panel* (a) *and* $p = 5$ (SOI_t) *in panel* (b).

relative to the zero-rainfall component, while in northwestern Australia it has increased. The top and middle rows in Figure 13(a) considered together suggest that the southwest and mid-northeastern coast of Australia are experiencing more dry days, while the northwest of Australia is experiencing fewer dry days. The bottom row of Figure 13(a) shows that on days on which rain occurs, it is more extreme in

the northeastern regions of Australia, while the reverse is true for the southwestern corner of Australia and mid-southeastern coast of Australia. Taken all together, Figure 13(a) indicates the southwestern corner has experienced a significant decline in rainfall levels. This result agrees with the climate science literature that speculates the decline in southwestern Australia rainfall levels is a consequence of land clearing (e.g., Andrich and Imberger (2013), Kala, Lyons and Nair (2011), Pitman et al. (2004)).

Similarly, Figures 13(a), 14(a) and 14(b) show the posterior quantiles for $\mu_{s,k,5}$, $\mu_{s,k,6}$ and $\mu_{s,k,7}$, respectively, corresponding to the covariates SOI_t, DMI_t, and SAM_t. Most locations display the same color for both the 10% and 90% quantiles, indicating significant relationships over much of Australia. In addition, inference regarding SOI_t and DMI_t matches previous work: positive SOI_t and negative DMI_t values are associated with more rainfall across most of the continent, with the opposite effect for negative SOI_t and positive DMI_t (Ummenhofer et al. (2011)). These results also match those for regions that have been investigated elsewhere in detail, so that, for instance, the eastern seaboard receives more rain for positive DMI_t (Pepler et al. (2014)). Furthermore, it appears that for most of the continent, positive SOI_t values are associated with rainfall which is more intense and variable, while for negative DMI_t values, the effect is ambiguous.

Our results for SAM_t are mixed. The top and middle rows of Figure 14(b) show that positive values of SAM_t are related to increases of rainfall for the southwest tip of Australia and to decreases in rainfall for the northwest coast of Australia. The bottom row shows that on days when rain is recorded, positive values of SAM_t are related to less heavy rain for southern Australia. This is in contrast with the climate literature that discusses the association of positive SAM_t values with decreased rainfall levels in southern Australia (Risbey et al. (2009)), in particular, the southwestern and southeastern tips of Australia (e.g., Hendon, Thompson and Wheeler (2007)). Also in contrast to the climate literature is the aforementioned decrease in rainfall for the northwest coast of Australia, a key feature of Figure 14(b) that does not, to our knowledge, appear in other research. Risbey et al. (2009) found no significant association between the values of SAM_t and rainfall in this region. Possible explanations for this discrepancy include differences in model construction, SAM index construction, the observational period (1957 to 2009 in Risbey et al. (2009)) and the granularity of the data. For example, the aggregation of daily to monthly data may mask the impact of SAM_t on rainfall.

6. Discussion and conclusion. We have presented a Bayesian mixture model for high dimensional nonstationary time series that accommodates nonstandard measurements and provides spatially varying inference. The effects of external covariates, and short and long-term temporal dependencies, are modeled through a mixture-of-experts model. A Gaussian process prior models the spatial dependencies of the model's mixture weights' parameters, the posterior distribution of which provides spatially varying inference.

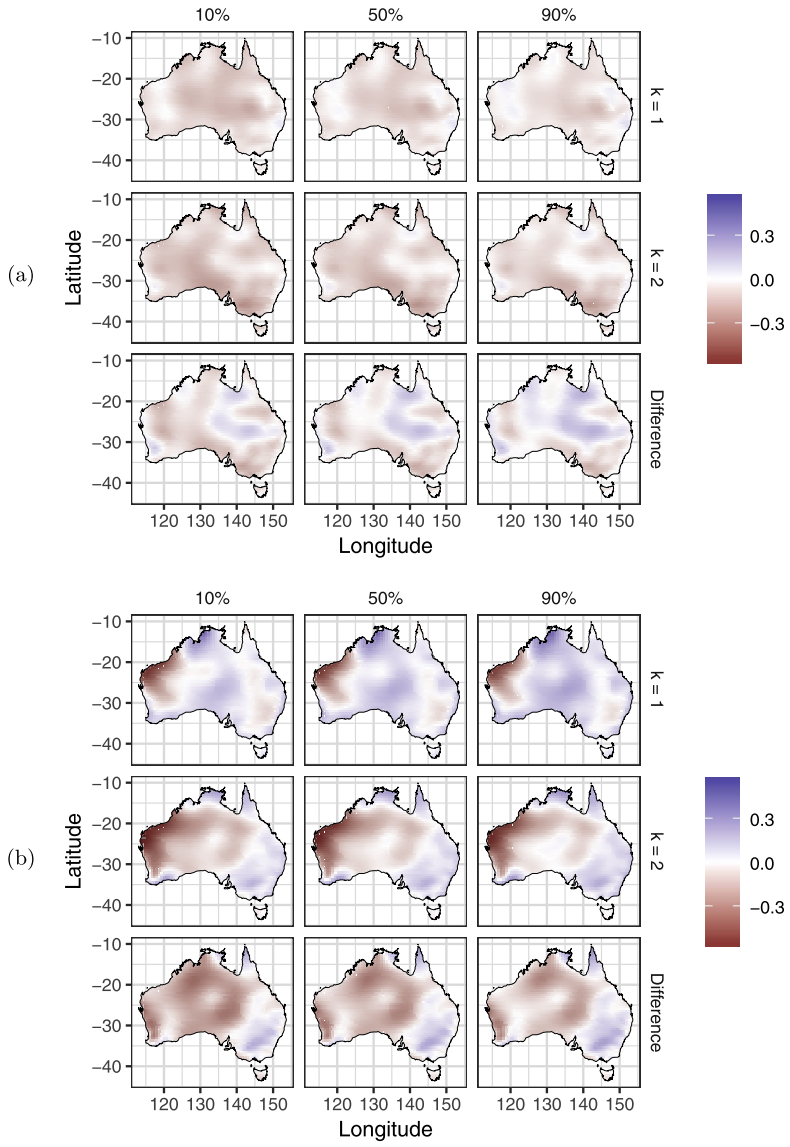


FIG. 14. *Maps of posterior* 10%, 50% *and* 90% (*columns*) *quantiles of* $\mu_{s,k,p}$ ($k = 1, 2$ *in* 1*st and* 2*nd row, respectively, of each panel*) *and of* $\mu_{s,2,p} - \mu_{s,1,p}$ (*bottom row of each panel*) *for the model with* $K = 2$, *where* $p = 6$ (DMI_t) *in panel* (a) *and* $p = 7$ (SAM_t) *in panel* (b).

The model has been tailored to understand the evolution of Australian daily rainfall over time and space, as well as the dependence between Australian daily rainfall and climate drivers via the choice of covariates and the specification of interpretable mixture components in a fully probabilistic framework. Extensions to more components and to other covariates are available to future users via publicly

available software. Work is underway to include an explicit spatial structure at the observational level that still scales to data sets such as the one presented in this article.

The data set used in this study is very large, with the observational period from February 1st, 1876 to December 31st, 2015, spanning 17,606 sites, resulting in more than 294 million observations. Analysis at this scale requires a suite of computational innovations. The innovations presented in this paper are encapsulated in an R package available for use in either a parallel and distributed environment for large data sets such as ours, or in standard serial computing environment for smaller data sets.

SUPPLEMENTARY MATERIAL

Supplement A: Model comparison supplement for "Climate inference on daily rainfall across the Australian continent, 1876–2015" (DOI: 10.1214/18-AOAS1218SUPPA; .pdf). We fit the model with $K = 3$ gamma components and compare the results to those corresponding to $K = 2$ gamma components.

Supplement B: Conditional distributions for the sampling scheme in "Climate inference on daily rainfall across the Australian continent, 1876–2015" (DOI: 10.1214/18-AOAS1218SUPPB; .pdf). We derive the conditional distributions used by the sampling scheme described in Section 4.1 of this paper.

Supplement C: Temporal and spatial diagnostics for "Climate inference on daily rainfall across the Australian continent, 1876–2015" (DOI: 10.1214/18-AOAS1218SUPPC; .pdf). We present log-odds and Spearman correlation diagnostics for the application to Australian daily rainfall, 1876–2015, along with a simulation study to assess the model's ability to perform spatially varying inference in the presence of spatially correlated observations.

REFERENCES

AMDAHL, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April* 18–20, 1967, *Spring Joint Computer Conference*. *AFIPS '*67 (*Spring*) 483–485. ACM, New York.

ANDRICH, M. A. and IMBERGER, J. (2013). The effect of land clearing on rainfall and fresh water resources in western Australia: A multi-functional sustainability analysis. *J. Appl. Econometrics* **20** 549–563.

BERTOLACCI, M., CRIPPS, E., ROSEN, O., LAU, J. and CRIPPS, S. (2019a). Conditional distributions for the sampling scheme in "Climate inference on daily rainfall across the Australian continent, 1876–2015." DOI:10.1214/18-AOAS1218SUPPA.

BERTOLACCI, M., CRIPPS, E., ROSEN, O., LAU, J. and CRIPPS, S. (2019b). Model comparison supplement for "Climate inference on daily rainfall across the Australian continent, 1876–2015." DOI:10.1214/18-AOAS1218SUPPB.

BERTOLACCI, M., CRIPPS, E., ROSEN, O., LAU, J. and CRIPPS, S. (2019c). Model diagnostics for "Climate inference on daily rainfall across the Australian continent, 1876–2015." DOI:10.1214/18-AOAS1218SUPPC.

CAI, W., VAN RENSCH, P., COWAN, T. and HENDON, H. (2012). An asymmetry in the IOD and ENSO teleconnection pathway and its impact on Australian climate. *J. Climate* **25** 6318–6329.

CHARLES, S. P., BATES, B. C. and HUGHES, J. P. (1999). A spatiotemporal model for downscaling precipitation occurrence and amounts. *J. Geophys. Res., Atmos.* **104** 31657–31669.

CHIB, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *J. Econometrics* **75** 79–97. MR1414504

COMPO, G. P., WHITAKER, J. S., SARDESHMUKH, P. D., MATSUI, N., ALLAN, R. J., YIN, X., GLEASON, B. E., VOSE, R. S., RUTLEDGE, G., BESSEMOULIN, P. et al. (2011). The twentieth century reanalysis project. *Q. J. R. Meteorol. Soc.* **137** 1–28.

DAMSLETH, E. (1975). Conjugate classes for gamma distributions. *Scand. J. Stat.* **2** 80–84. MR0378169

EDDELBUETTEL, D. and SANDERSON, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Statist. Data Anal.* **71** 1054–1063. MR3132026

FENG, J., LI, J. and LI, Y. (2010). Is there a relationship between SAM and southwest western Australia winter rainfall. *J. Climate* **23** 6082–6089.

FURRER, E. M. and KATZ, R. W. (2007). Generalized linear modeling approach to stochastic weather generators. *Climate Research* **34** 129–144.

GONG, D. and WANG, S. (1999). Definition of Antarctic oscillation index. *Geophysical Research Letters* **26** 459–462.

HENDON, H., THOMPSON, D. and WHEELER, M. (2007). Australian rainfall and surface temperature variations associated with the southern hemisphere annular mode. *J. Climate* **20** 2452–2467.

HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2016). A Bayesian hidden Markov model of daily precipitation over South and East Asia. *Journal of Hydrometeorology* **17** 3–25.

HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2017). Bayesian nonhomogeneous Markov models via Pólya-gamma data augmentation with applications to rainfall modeling. *Ann. Appl. Stat.* **11** 393–426. MR3634329

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87.

KALA, J., LYONS, T. J. and NAIR, U. S. (2011). Numerical simulations of the impacts of land-cover change on cold fronts in South-West western Australia. *Boundary-Layer Meteorology* **138** 121–138.

KING, A., ALEXANDER, L. and DONAT, M. (2013). Asymmetry in the response of eastern Australia extreme rainfall to low-frequency Pacific variability. *Geophysical Research Letters* **40** 1–7.

KLEIBER, W., KATZ, R. W. and RAJAGOPALAN, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res.* **48**.

LYNCH, N. A. (1996). *Distributed Algorithms. The Morgan Kaufmann Series in Data Management Systems*. Morgan Kaufmann, San Francisco, CA. MR1388778

NAVEAU, P., HUSER, R., RIBEREAU, P. and HANNART, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resour. Res.* **52** 2753–2769.

PEPLER, A., TIMBAL, B., RAKICH, C. and COUTTS-SMITH, A. (2014). Indian Ocean dipole overrides ENSO's influence on cool season rainfall across the eastern seaboard of Australia. *J. Climate* **27** 3816–3826.

PITMAN, A. J., NARISMA, G. G., PIELKE, R. A. and HOLBROOK, N. J. (2004). The impact of land cover change on the climate of southwest western Australia. *J. Geophys. Res.* **109** 1–12.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

R CORE TEAM (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RAYNER, N. A., PARKER, D. E., HORTON, E. B., FOLLAND, C. K., ALEXANDER, L. V., ROWELL, D. P., KENT, E. C. and KAPLAN, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res., Atmos.* **108**.

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

RISBEY, J., POOK, M., MCINTOSH, P., WHEELER, M. and HENDON, H. (2009). On the remote drivers of rainfall variability in Australia. *Mon. Weather Rev.* **137** 3233–3253.

ROSEN, O., STOFFER, D. S. and WOOD, S. (2009). Local spectral analysis via a Bayesian mixture of smoothing splines. *J. Amer. Statist. Assoc.* **104** 249–262. MR2504376

SAJI, N. H., GOSWAMI, B. N., VINAYACHANDRAN, P. N. and YAMAGATA, T. (1999). A dipole mode in the tropical Indian Ocean. *Nature* **401** 360–363.

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

STONE, R. (2014). Constructing a framework for national drought policy: The way forward—The way Australia developed and implemented the national drought policy. *Weather and Climate Extremes* **3** 117–125.

TROUP, A. J. (1965). The "southern oscillation". *Q. J. R. Meteorol. Soc.* **91** 490–506.

UMMENHOFER, C. C., ENGLAND, M. H., MCINTOSH, P. C., MEYERS, G. A., POOK, M. J., RISBEY, J. S., GUPTA, A. S. and TASCHETTO, A. S. (2009). What causes southeast Australia's worst droughts?. *Geophysical Research Letters* **36** L04707.

UMMENHOFER, C. C., GUPTA, A. S., BRIGGS, P. R., ENGLAND, M. H., MCINTOSH, P. C., MEYERS, G. A., POOK, M. J., RAUPACH, M. R. and RISBEY, J. S. (2011). Indian and Pacific Ocean influences on southeast Australian drought and soil moisture. *J. Climate* **24** 1313–1336.

UMMENHOFER, C. C., GUPTA, A. S., ENGLAND, M. H., TASCHETTO, A. S., BRIGGS, P. R. and RAUPACH, M. R. (2015). How did ocean warming affect Australian rainfall extremes during the 2010/2011 La Niña event. *Geophysical Letters* **42** 9942–9951.

VAN DIJK, A., BECK, H., CROSBIE, R., DE JEU, R., LIU, G., PODGER, Y., TIMBAL, B. and VINEY, N. (2013). The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resour. Res.* **49**.

VRAC, M. and NAVEAU, P. (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resour. Res.* **43**.

WAHBA, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

WILKS, D. S. (1999). Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology* **93** 153–169.

WOOD, S. (2013). Applications of Bayesian smoothing splines. In *Bayesian Theory and Applications* (P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens, eds.) 309–335. Oxford Univ. Press, Oxford. MR3221170

WOOD, S., ROSEN, O. and KOHN, R. (2011). Bayesian mixtures of autoregressive models. *J. Comput. Graph. Statist.* **20** 174–195. MR2816544

YU, H. (2002). Rmpi: Parallel statistical computing in R. *R News* **2** 10–14.

M. BERTOLACCI
E. CRIPPS
J. LAU
SCHOOL OF MATHEMATICS AND STATISTICS
UNIVERSITY OF WESTERN AUSTRALIA
PERTH, WESTERN AUSTRALIA
AUSTRALIA
E-MAIL: michael.bertolacci@research.uwa.edu.au
edward.cripps@uwa.edu.au
john.lau@uwa.edu.au

S. CRIPPS
CENTRE FOR TRANSLATIONAL DATA SCIENCE
SCHOOL OF MATHEMATICS AND STATISTICS
UNIVERSITY OF SYDNEY
SYDNEY, NEW SOUTH WALES
AUSTRALIA
E-MAIL: sally.cripps@sydney.edu.au

O. ROSEN
DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF TEXAS AT EL PASO
EL PASO, TEXAS 79968-0514
USA
E-MAIL: orosen@utep.edu