



# A probabilistic gridded product for daily precipitation extremes over the United States

Mark D. Risser<sup>1</sup> · Christopher J. Paciorek<sup>2</sup> · Michael F. Wehner<sup>1</sup> · Travis A. O'Brien<sup>1</sup> · William D. Collins<sup>1</sup>

Received: 16 July 2018 / Accepted: 16 January 2019 / Published online: 28 February 2019  
© The Author(s) 2019

## Abstract

Gridded data products, for example interpolated daily measurements of precipitation from weather stations, are commonly used as a convenient substitute for direct observations because these products provide a spatially and temporally continuous and complete source of data. However, when the goal is to characterize climatological features of extreme precipitation over a spatial domain (e.g., a map of return values) at the native spatial scales of these phenomena, then gridded products may lead to incorrect conclusions because daily precipitation is a fractal field and hence any smoothing technique will dampen local extremes. To address this issue, we create a new “probabilistic” gridded product specifically designed to characterize the climatological properties of extreme precipitation by applying spatial statistical analysis to daily measurements of precipitation from the Global Historical Climatology Network over the contiguous United States. The essence of our method is to first estimate the climatology of extreme precipitation based on station data and then use a data-driven statistical approach to interpolate these estimates to a fine grid. We argue that our method yields an improved characterization of the climatology within a grid cell because the probabilistic behavior of extreme precipitation is much better behaved (i.e., smoother) than daily weather. Furthermore, the spatial smoothing innate to our approach significantly increases the signal-to-noise ratio in the estimated extreme statistics relative to an analysis without smoothing. Finally, by deriving a data-driven approach for translating extreme statistics to a spatially complete grid, the methodology outlined in this paper resolves the issue of how to properly compare station data with output from earth system models. We conclude the paper by comparing our probabilistic gridded product with a standard extreme value analysis of the Livneh gridded daily precipitation product. Our new data product is freely available on the Harvard Dataverse (<https://bit.ly/2CXdnuj>).

**Keywords** Extreme value analysis · Precipitation · Spatial statistics · Nonparametric bootstrap · Global Historical Climatology Network · Gaussian processes · Gridded daily precipitation

## 1 Introduction

Gridded data products of precipitation are a popular substitute for daily measurements of rainfall from weather stations. The products are generated by interpolating the station measurements onto a uniform space-time grid to create a spatially

and temporally complete, continuous, and homogeneous version of the raw data. For this reason, gridded products are often used to summarize the climatological properties of extreme precipitation, for example maps of return values, and then to evaluate the performance of climate models with respect to extremes. In traditional analysis of precipitation extremes using gridded products (see, e.g., Wehner 2013; Sylla et al. 2013, and many others), the extreme climatology is estimated separately for each grid cell using a univariate extreme value analysis. However, we assert that gridded daily precipitation products are problematic data sources for constructing these extreme climatologies because daily precipitation is well-known to exhibit fractal scaling (e.g., Lovejoy et al. 2008; Maskey et al. 2016, and numerous references therein), and therefore any spatial smoothing or averaging during the gridding process will diminish variability

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00382-019-04636-0>) contains supplementary material, which is available to authorized users.

---

✉ Mark D. Risser  
mdrisser@lbl.gov

<sup>1</sup> Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

<sup>2</sup> University of California, Berkeley, CA, USA

and extreme values. Additionally, a recent thread of research (King et al. 2013; Gervais et al. 2014; Timmermans et al. 2019) explicitly questions the appropriateness of using gridded products as a substitute for observed extremes.

As an alternative to using gridded data products, irregularly observed weather station measurements can be used to obtain spatially-complete summaries of extreme precipitation by utilizing the diverse statistics literature on spatial extreme value analysis. These statistical tools collectively analyze extremes over space for processes such as precipitation, specifically allowing one to estimate the distribution of extreme precipitation even for locations where no data is available. The methods broadly fall into one of four categories: max-stable processes (Haan 1984; Smith 1990; Schlather 2002; Kabluchko et al. 2009), which provide mathematically-founded models for characterizing spatial dependence among extremes; copula-based approaches (Hüsler and Reiss 1989; Demarta and McNeil 2005; Sang and Gelfand 2010; Fuentes et al. 2013; Krupskii et al. 2018), which construct a joint multivariate distribution for spatial extremes via careful modeling of transformed marginal (univariate) distributions; Bayesian methods (Reich and Shaby 2012; Shaby and Reich 2012; Morris et al. 2017), which use a hierarchical framework to build numerically tractable models based upon the mathematics of max-stable processes; and nonparametric mixture models (Gelfand and Kottas 2002; Wang et al. 2011; Kottas et al. 2012), which simultaneously analyze both the average and extreme values of precipitation. See Davison et al. (2012) for an excellent review of these approaches.

In this paper, we seek to characterize the climatology of extreme precipitation based on measurements from approximately five thousand Global Historical Climatology Network (GHCN) stations (see Sect. 2) over the contiguous United States (CONUS). Unfortunately, practically speaking, many of the aforementioned statistical methods are only suitable for a small number of weather stations. For example, Davison et al. (2012) use a data set with just 36 stations, Saunders et al. (2017) employ 173 stations to derive a max-stable process to model extreme precipitation in Australia, and Shaby and Reich (2012) analyze the largest data set considered to date (to our knowledge) comprised of approximately one thousand stations. Regardless of the number of weather stations of interest, a second limitation of the aforementioned methods is that they are only appropriate for homogeneous spatial domains. CONUS, on the other hand, is a highly heterogeneous spatial domain with variable topography interacting with a diverse set of physical phenomena that produce extreme precipitation. These phenomena include atmospheric rivers on the west coast, tropical cyclones on the east coast, and mesoscale convective systems in the Great Plains.

In light of these shortcomings in existing methodologies for spatial extremes, note that we do not need to simulate

realistic fields of seasonal maxima, which is a complicating factor in the aforementioned max-stable and Bayesian approaches to spatial extreme value analysis (furthermore, note that we do not intend to model the full space-time distribution of daily precipitation, as in Serinaldi and Kilsby 2014). Instead, we simply wish to characterize the statistics of extreme precipitation over CONUS. In this case, one apparently viable option would be to employ latent variable or conditional independence models (Cooley et al. 2007; Craigmile and Guttorp 2013; Mannshardt et al. 2013), which provide flexible methodologies suitable for larger data sets and heterogeneous spatial domains. These methods assume that daily precipitation totals over space occur independently of one another and are conditional on latent processes that characterize spatial dependence in extremes. The assumption of independence invalidates the application of latent variable methods to modeling extremes of an atmospheric process such as precipitation because daily precipitation measurements at nearby stations are certainly not independent. This dependence is induced by the spatial coherence of storm systems so that if one weather station experiences an extreme on a specific day it is more likely that a nearby station will also experience an extreme on the same day. Failing to account for this so-called “storm dependence” means that the resulting maps of extreme statistics will be misleading since the true spatial signals will be commingled with noise.

In order to create a viable alternative analytical technique, we first conduct an extreme value analysis to estimate the extreme statistics of precipitation for each station and then interpolate these statistics to a fine grid. In order to account for spatial coherence in these extreme statistics, we use second-order nonstationary spatial Gaussian process models that account for heterogeneities in the climatology of extreme precipitation and enable statistical inference at locations where we do not have observations of daily precipitation. By construction, this framework accounts for both first-order heterogeneities in the expected values due to topography and second-order heterogeneities in the co-variability. Then, we use a nonparametric, or “block”, bootstrap approach to characterize uncertainty in the extreme statistics of precipitation while accounting for storm dependence. Intuitively, our approach can be seen as commuting the order of operations relative to existing analytical methods that in general first apply spatial and temporal interpolation to station data onto a regular grid and then calculate extreme statistics from that gridded product. We demonstrate that our method yields a larger signal-to-noise ratio in estimates of extreme statistics relative to the ratios yielded by more traditional analytical methods that do not borrow strength spatially.

Focusing on daily measurements of precipitation from the Global Historical Climatology Network (GHCN) over the contiguous United States (CONUS), the result of our analysis is a probabilistic gridded product that describes the

spatial climatological distribution of extreme precipitation. To illustrate why our methodology provides an improved characterization of extreme precipitation statistics, consider two nearby towns in a homogeneous region, e.g., the Great Plains. While these two towns might experience different storms for any given day or month, our general hypothesis is that these two towns will exhibit similar statistics of extreme precipitation because of their similar climatology. The same argument holds over the rest of CONUS, especially after correcting for the effects of elevation. As a result, it makes more sense to borrow strength over space and conduct smoothing on the statistics of extremes rather than daily weather itself. A traditional extreme value analysis using gridded daily products does not incorporate this important borrowing of strength over space.

Two threads of research should be mentioned in relation to the novel methodology described in this paper. The first is Diaconescu et al. (2015) and references therein, which explores the order of operations when regridding data products and climate model output for comparison. In this paper, they compare a “first-step” procedure (which first interpolates daily weather fields and then computes a derived index on the common grid) and a “last-step” procedure (which first computes the derived index and then interpolates to a common grid), finding that the last-step procedure yields much smaller errors from the regridding process (a result that is robust to the interpolation method). The intuition for our approach supports a preference for the last-step procedure and is indeed quite similar, except that (1) we propose regridding station data instead of preprocessed data products, and (2) we interpolate parameters from a statistical model instead of data summaries (e.g., the largest seasonal daily precipitation total): hence, our results immediately yield a variety of gridded climatological summaries (e.g., return levels or return probabilities for any return period) instead of requiring a separate interpolation scheme for each desired summary. Second, our methodology is related to regional frequency analysis (RFA) (Wallis et al. 2007; see also Norbiato et al. 2007; Yang et al. 2010) which, as with the work at hand, seeks to provide spatially-complete maps of the climatology of extreme precipitation based on the interpolation of pointwise summary statistics calculated for each of a set of weather stations. However, the methodologies are quite different: the cornerstone of RFA is that “data from sites within a homogeneous region can be pooled to improve the reliability of the magnitude-frequency estimates” (Wallis et al. 2007), and hence RFA requires the specification of such regions. This is a nontrivial task, particularly for a large, heterogeneous domain such as CONUS. Our approach, on the other hand, allows the data itself to determine the degree of local homogeneity in the climatology of extreme precipitation across a set of weather stations, and is hence more easily applied to a generic spatial domain.

Finally, we note that the methodology outlined in this paper resolves an outstanding problem for evaluation of Earth System Models (ESMs) with respect to extremes: namely, how to properly compare irregularly observed station data with output from ESMs. Comparison of ESMs against station data poses several major challenges. First, the volume of the atmosphere sampled by station data is typically orders of magnitude smaller than the volumes represented by a typical ESM grid cell. While ESM sub-grid parameterizations are designed to emulate the parent distributions of measurable quantities and to report averages across these distributions, stations can record samples from any part of the parent distribution of their measurable quantities, including the “tails”. Second, the typical integrations of ESMs forced just with external boundary conditions are designed to mimic the climatological statistics of typical weather conditions for each surface station, but not the precise deterministic time-series of the weather actually recorded by those stations. Thus, the question is how best to compare deterministic time series of point measurements against a statistical characterization of the mesoscale climate conditions consistent with those measurements. The methodology developed in this paper addresses this question by framing the model-data intercomparison in terms of the extreme statistics from the outset, and by developing a method for interpolating statistical properties between stations that produces mesoscale and synoptic-scale statistics directly comparable to those from an ESM.

The paper proceeds as follows: in Sect. 2 we introduce the GHCN data used to create our probabilistic gridded product, and in Sect. 3 we describe the extreme value analysis, spatial smoothing, and block bootstrap approach. In Sect. 4, we present the results of our analysis and give a summary of the probabilistic gridded product, and in Sect. 4.3 we conduct a comparative analysis using the Livneh gridded daily precipitation product. Section 5 concludes the paper.

## 2 Data

The data used for the following analysis consist of daily measurements of total precipitation (in millimeters) obtained from the Global Historical Climatology Network (GHCN; Menne et al. 2012a, b) over the contiguous United States (CONUS). The GHCN is quite extensive over CONUS, consisting of over twenty thousand weather stations with measurements dating back to the late nineteenth century, although of course the length and quality of individual records are highly variable. In addition to daily precipitation measurements, the database includes three quality control flags, providing for each day an indication of the data quality (“QFLAG”), source quality (“SFLAG”), and measurement quality (“MFLAG”). Quality control for the raw (nonmissing) daily data values

involved the following criteria: (1) values were kept only if the QFLAG field was blank, meaning the measurement did not fail any quality assurance checks; (2) values were kept only if the SFLAG field was not equal to “S” (which implies that the measurement may differ significantly from “true” daily data); (3) any values with MFLAG equal to “T” were set to a measurement of 0 mm (“T” indicates a “trace of precipitation”). After processing the daily values based on the quality control flags, we selected the subset of stations that had a minimum of 66.7% nonmissing daily precipitation measurements over December 1, 1949 through November 30, 2017. This procedure resulted in a high-quality set of daily precipitation measurements for  $n = 5202$  stations (see Supplementary Figure D.1) covering 68 years.

To establish some notation, define  $Z_{tk}(s_i)$  to be the precipitation measurement, in mm, for day  $k = 1, \dots, m_t$  in a fixed 3-month season (e.g., December, January, and February or DJF) of year  $t = 1950, \dots, 2017$  at station  $s_i \in S = \{s_1, \dots, s_n\} \subset G$ , where  $m_t$  is the number of daily observations in a fixed season of year  $t$ ,  $S$  denotes the  $n = 5202$  stations shown in Supplementary Figure D.1, and  $G$  denotes the contiguous United States. Note that the year represents a “season year” such that, for example, the 1950 DJF is defined as December, 1949 to February, 1950.

### 3 Statistical methods

Recall from Sect. 1 that the essence of our method is to first obtain estimates of the climatological features of extreme precipitation based on station data (Sect. 3.1) and then use a spatial statistical approach to infer the true underlying climatology over a fine grid (Sect. 3.2), accounting for uncertainty and storm dependence using a nonparametric block bootstrap approach (Sect. 3.3). We now provide specific details on each step of this analysis.

#### 3.1 Stage 1: extreme value analysis for individual stations

The following provides a framework for modeling extreme value statistics of daily precipitation in a fixed 3-month season (i.e., DJF, MAM, JJA, or SON) over CONUS, first considering each station individually. While there are several different ways to characterize the extreme values of a stochastic process (see, e.g., Coles 2001), we opt for the generalized extreme value (GEV) family of distributions, which is a modeling framework for the maxima of a process over a pre-specified “block”, here, each 3-month season. For an arbitrary station  $s$ , define the seasonal maximum in year  $t$  as  $Y_t(s)$ , that is,  $Y_t(s) = \max_k \{Z_{tk}(s) : k = 1, \dots, m_t\}$ . Coles (2001) (Theorem 3.1.1, page 48) shows that (for large  $m_t$ ) the cumulative

distribution function (CDF) of  $Y_t(s)$  is a member of the GEV family

$$G_{s,t}(y) \equiv \mathbb{P}(Y_t(s) \leq y) = \exp \left\{ - \left[ 1 + \xi_t(s) \left( \frac{y - \mu_t(s)}{\sigma_t(s)} \right) \right]^{-1/\xi_t(s)} \right\}, \quad (1)$$

defined for  $\{y : 1 + \xi_t(s)(y - \mu_t(s))/\sigma_t(s) > 0\}$ . The GEV family of distributions (1) is characterized by three space-time specific parameters: the location parameter  $\mu_t(s) \in \mathcal{R}$  (which describes the center of the distribution), the scale parameter  $\sigma_t(s) > 0$  (which describes the spread of the distribution), and the shape parameter  $\xi_t(s) \in \mathcal{R}$ . The shape parameter  $\xi_t(s)$  is the most important in terms of determining the qualitative behavior of the distribution of daily rainfall at a given location: if  $\xi_t(s) < 0$ , the distribution has a finite upper bound; if  $\xi_t(s) > 0$ , the distribution has no upper limit; if  $\xi_t(s) = 0$ , the distribution is again unbounded and the CDF (1) is interpreted as the limit  $\xi_t(s) \rightarrow 0$  (Coles 2001).

The GEV framework is commonly referred to as the “block maxima” approach to extreme value analysis. The point process or “peaks over threshold” (POT) approach is often preferred to the block maxima approach because, as in the threshold excess model, estimates of the climatological coefficients are obtained from all extreme values (i.e., those that exceed a high threshold) instead of the single maximum over a block of time. But, as discussed in Coles (2001) Section 4.3.1, the main challenge is identifying a threshold for what is considered “extreme”: too small a threshold violates the mathematical (asymptotic) basis of the POT approach, leading to biased coefficient estimates, while too large a threshold results in a very small number of exceedances, yielding large variance. Also, in practice, when conducting station-specific extreme value analysis, the POT approach resulted in a larger number of numerical optimization errors. Finally, in this case where we wish to characterize trends in extremes over time, it is not clear that a temporally-constant threshold is appropriate. Therefore, we opted to use the GEV framework for this analysis. However, as a sensitivity analysis we also conducted the full analysis using the POT approach and found that results were similar (see Supplementary Figure D.9 for a comparison).

As the notation in (1) suggests, we can specify flexible time-varying models for the spatially-varying climatological coefficients. For the analysis in this paper, we use a simple temporal trend, where the location parameter varies linearly with time and the other coefficients are constant in time:

$$\mu_t(s) = \mu_0(s) + \mu_1(s)t, \quad \sigma_t(s) \equiv \sigma(s), \quad \xi_t(s) \equiv \xi(s), \quad (2)$$

(following, e.g., Westra et al. 2013 and others). We henceforth refer to  $\mu_0(s)$ ,  $\mu_1(s)$ ,  $\sigma(s)$ , and  $\xi(s)$  as the *climatological coefficients* for location  $s$ , as these values describe the climatological distribution of extreme precipitation in each year. Note that the trend model (2) averages over both inter-annual variability (e.g., the El Niño/Southern Oscillation or ENSO)

and lower frequency modes of variability (e.g., the Pacific Decadal Oscillation or PDO), such that we only attempt to characterize temporally smooth trends in extremes. Some authors (e.g., Cooley et al. 2007; Risser and Wehner 2017) have also permitted the scale parameter to contain covariates allowing the width of the GEV distribution to co-vary. Uncertainty in the magnitude of the shape parameter is generally quite large, negating any benefits of allowing it to co-vary. In this paper, however, we consider only (2) because in

(Coles 2001). Furthermore, while the return value is the extreme quantile of the extreme value distribution in each year, we can equivalently calculate the return period for a particular magnitude event  $x$  in year  $t$ , denoted  $\rho_t^{(x)}(\mathbf{s})$ , which indicates that there is a one in  $\rho_t^{(x)}(\mathbf{s})$  chance that an event at least as large as  $x$  will occur in year  $t$  at location  $\mathbf{s}$ . In other words,  $\rho_t^{(x)}(\mathbf{s})$  is the inverse probability of the seasonal maximum daily precipitation total in year  $t$  exceeding  $x$ , i.e.,  $P(Y_t(\mathbf{s}) > x) = 1/\rho_t^{(x)}(\mathbf{s})$ . The return period is available in closed form by inverting (4):

$$\rho_t^{(x)}(\mathbf{s}) = \begin{cases} (1 - \exp\{-[1 - \xi(\mathbf{s})([\mu_0(\mathbf{s}) + \mu_1(\mathbf{s})t] - x)/\sigma(\mathbf{s})]^{-1/\xi(\mathbf{s})}\})^{-1}, & \xi(\mathbf{s}) \neq 0 \\ (1 - \exp\{-\exp\{([\mu_0(\mathbf{s}) + \mu_1(\mathbf{s})t] - x)/\sigma(\mathbf{s})\}\})^{-1}, & \xi(\mathbf{s}) = 0 \end{cases} \tag{5}$$

a statistical sense it performed as well as more sophisticated trend models for individual stations.

Considering all years, the statistical model (or log-likelihood) for all of the observed seasonal maxima at station  $\mathbf{s}$ , defined as  $\mathbf{y}(\mathbf{s}) = \{y_t(\mathbf{s}) : t = 1950, \dots, 2017\}$ , conditional on the climatological coefficients  $\mu_0(\mathbf{s}), \mu_1(\mathbf{s}), \sigma(\mathbf{s})$ , and  $\xi(\mathbf{s})$ , is

$$\begin{aligned} \mathcal{L}(\mu_0(\mathbf{s}), \mu_1(\mathbf{s}), \sigma(\mathbf{s}), \xi(\mathbf{s}); \mathbf{y}(\mathbf{s})) = & - \sum_{t=1950}^{2017} \log \sigma(\mathbf{s}) \\ & - [1 + 1/\xi(\mathbf{s})] \sum_{t=1950}^{2017} \log \left[ 1 + \xi(\mathbf{s}) \left( \frac{y_t(\mathbf{s}) - [\mu_0(\mathbf{s}) + \mu_1(\mathbf{s})t]}{\sigma(\mathbf{s})} \right) \right] \\ & - \sum_{t=1950}^{2017} \left[ 1 + \xi(\mathbf{s}) \left( \frac{y_t(\mathbf{s}) - [\mu_0(\mathbf{s}) + \mu_1(\mathbf{s})t]}{\sigma(\mathbf{s})} \right) \right]^{-1/\xi(\mathbf{s})}. \end{aligned} \tag{3}$$

The fact that (3) involves a sum indicates that we assume independence across years, which is a reasonable assumption given the time-varying statistical model in (2).

While the spatially-varying climatological coefficients in (2) are of interest themselves, we are often more interested in a summary of the climatological coefficients known as the  $r$ -year return value (sometimes called the return level). The  $r$ -year return value, denoted  $\phi_t^{(r)}(\mathbf{s})$ , is defined as the seasonal maximum daily precipitation total (i.e.,  $Y_t(\mathbf{s}) = \max_k \{Z_{tk}(\mathbf{s})\}$ ) that is expected to be exceeded on average once every  $r$  years. In other words,  $\phi_t^{(r)}(\mathbf{s})$  is an estimate of the  $1 - \frac{1}{r}$  quantile of the distribution of seasonal maximum daily precipitation in year  $t$  at station  $\mathbf{s}$ , i.e.,  $P(Y_t(\mathbf{s}) > \phi_t^{(r)}(\mathbf{s})) = 1/r$ . Because of how the year-specific distribution (1) has been defined,  $\phi_t^{(r)}(\mathbf{s})$  can be written in closed form in terms of the climatological coefficients:

$$\phi_t^{(r)}(\mathbf{s}) = \begin{cases} [\mu_0(\mathbf{s}) + \mu_1(\mathbf{s})t] - \frac{\sigma(\mathbf{s})}{\xi(\mathbf{s})} [1 - \{-\log(1 - 1/r)\}^{-\xi(\mathbf{s})}], & \xi(\mathbf{s}) \neq 0 \\ [\mu_0(\mathbf{s}) + \mu_1(\mathbf{s})t] - \sigma(\mathbf{s}) \log\{-\log(1 - 1/r)\}, & \xi(\mathbf{s}) = 0 \end{cases} \tag{4}$$

(Coles 2001).

In summary, this first stage of our method involves applying the GEV analysis defined by (2) and (3) to the observed seasonal maxima, independently for each station. For this step, we use the `climextRemes` software package (Paciorek 2016), which is an R/Python package for conducting extreme value analysis with climate data. We use

`climextRemes` functionality to obtain maximum likelihood estimates of the climatological coefficients, denoted  $\{\hat{\mu}_0(\mathbf{s}), \hat{\mu}_1(\mathbf{s}), \hat{\sigma}(\mathbf{s}), \hat{\xi}(\mathbf{s})\}$  for  $\mathbf{s} \in \mathcal{S}$ , which are estimates of the true climatological coefficients  $\{\mu_0(\mathbf{s}), \mu_1(\mathbf{s}), \sigma(\mathbf{s}), \xi(\mathbf{s})\}$ .

### 3.2 Stage 2: spatial statistical modeling for the climatological coefficients

In order to account for dependence across the climatological coefficients over the spatial domain  $G$ , we use second-order nonstationary spatial Gaussian process models for each of the spatially-varying climatological coefficients in (2). Gaussian processes (GPs) are an extremely popular tool in statistical modeling, and are broadly applied in spatial and environmental statistics as well as machine learning and emulation of complex mathematical models. GPs are a form of “nonparametric”



or nonlinear regression, as they characterize a nonlinear relationship between a set of inputs and an output: in our application, the inputs are geographic coordinates and the output (or response) is one of the estimated climatological coefficients  $\{\mu_0, \mu_1, \log \sigma, \xi\}$ . Intuitively, GPs interpolate or smooth the output variable to infer a nonlinear relationship with the inputs.

The GP is a popular choice for modeling point-referenced, continuously-indexed spatial processes like the climatological coefficients because all finite-dimensional distributions are known to be Gaussian (see Eq. 7 below), and because the GP is completely specified by a characterization of its first- and second-order properties. Furthermore, the second-order properties can be specified using any valid spatial covariance function to describe the degree and nature of spatial dependence (or smoothness) in the environmental process of interest. For example, define  $\{u(\mathbf{s}) : \mathbf{s} \in G\}$  to be a general process defined over a spatial domain  $G \subset \mathbb{R}^2$ ; without loss of generality, suppose  $u(\cdot)$  is mean-zero. The spatial covariance function of  $u(\cdot)$  is defined as

$$C(\mathbf{s}, \mathbf{s}') \equiv \text{Cov}[u(\mathbf{s}), u(\mathbf{s}')] = E[u(\mathbf{s})u(\mathbf{s}')] \quad (6)$$

for all  $\mathbf{s}, \mathbf{s}' \in G$ , where  $E[\cdot]$  is the expected value with respect to the distribution of  $u(\cdot)$ . The covariance function is always symmetric (i.e.,  $C(\mathbf{s}, \mathbf{s}') = C(\mathbf{s}', \mathbf{s})$ ) and must be a nonnegative definite function. In practice,  $C$  is often assumed to be second-order stationary, meaning that the covariance is fully defined by the separation vector  $\mathbf{s} - \mathbf{s}'$ . However, for the application at hand, this is a highly restrictive assumption, because the spatial covariance for environmental processes likely varies across the domain. Therefore, we use a *nonstationary* covariance function, which allows the covariance to depend on location and can account for second-order heterogeneities over the spatial domain of interest. A GP that uses a nonstationary covariance function is often referred to simply as a nonstationary GP. For further details on this subject, we refer the interested reader to Risser (2016).

A nonstationary Gaussian process can be used in a statistical model for the climatological coefficients over CONUS as follows. Let  $\theta$  represent an arbitrary coefficient, i.e.,  $\theta \in \{\mu_0, \mu_1, \log \sigma, \xi\}$  (we model the scale parameter  $\sigma$  on the logarithmic scale because  $\sigma$  must be positive); the Gaussian process model for each  $\theta(\mathbf{s})$  can be framed as the statistical linear model

$$\theta(\mathbf{s}) = \delta_\theta(\mathbf{s}) + u_\theta(\mathbf{s}), \quad \mathbf{s} \in G. \quad (6)$$

In (6),  $\delta_\theta(\mathbf{s}) \equiv E[\theta(\mathbf{s})]$  represents the expected or average behavior of  $\theta(\mathbf{s})$  and  $u_\theta(\mathbf{s})$  represents a mean-zero residual process that characterizes deviations between  $\theta(\mathbf{s})$  and  $\delta_\theta(\mathbf{s})$ . In a more traditional linear regression setting (i.e., ordinary least squares),  $u_\theta(\mathbf{s})$  is assumed to be (spatially) independent, but here we instead model  $u_\theta(\mathbf{s})$  as a spatially correlated nonstationary Gaussian process with a mean fixed at

zero and covariance function  $C_\theta$ . For a fixed set of locations  $S = \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset G$ , (6) implies that

$$\boldsymbol{\theta} \equiv (\theta(\mathbf{s}_1), \dots, \theta(\mathbf{s}_n)) \sim N_n(\boldsymbol{\delta}_\theta, \boldsymbol{\Sigma}_\theta), \quad (7)$$

where  $N_d(\mathbf{a}, \mathbf{B})$  denotes a  $d$ -variate Gaussian distribution with mean vector  $\mathbf{a}$  and covariance matrix  $\mathbf{B}$ . In (7),  $\boldsymbol{\delta}_\theta = (\delta_\theta(\mathbf{s}_1), \dots, \delta_\theta(\mathbf{s}_n))$  and the  $i, j$  element of  $\boldsymbol{\Sigma}_\theta$  is  $C_\theta(\mathbf{s}_i, \mathbf{s}_j)$ .

To complete the model specification in (6), we must specify the form of  $\delta_\theta(\mathbf{s})$  as well as the covariance function  $C_\theta$ ; for full details, see Appendix 2. In short, the mean function  $\delta_\theta(\mathbf{s})$  will be linear in an elevation-based covariate with a spatially-varying effect, and the covariance function for  $C_\theta$  is second-order nonstationary such that  $C_\theta(\mathbf{s}, \mathbf{s}') \neq C_\theta(\mathbf{s} - \mathbf{s}')$ . Again, intuitively, using a nonstationary covariance function allows the second order properties of  $u_\theta(\cdot)$  to vary over space, including both variance and the magnitude and direction of the spatial extent of the smoothing. A statistical treatment of (7) uses a likelihood-based approach (e.g., maximum likelihood estimation) to estimate the statistical parameters associated with the mean and covariance.

### 3.3 Nonparametric bootstrap for uncertainty quantification

Finally, we specify a systematic framework for combining the GEV likelihood in (3) and the Gaussian process priors for the coefficient fields. However, note that we have not yet specified how to interrelate the likelihood (3) across stations, and these relationships are a critical component of a statistical model for daily precipitation extremes over space. As mentioned in Sect. 1, the simplest approach is the conditional independence or latent variable approach (see, e.g., Cooley et al. 2007; Craigmille and Guttorp 2013; Mannshardt et al. 2013), which assumes independence across stations in the likelihood and relies on the Gaussian process priors to capture dependence in extremes. However, while this approach is somewhat feasible for large data sets and heterogeneous spatial domains, the fact that it does not account for storm dependence makes it theoretically incorrect. And, as mentioned in Sect. 1, existing approaches for more appropriately modeling extremes over space are insufficient for daily precipitation measurements from a large network like GHCN over CONUS.

Alternatively, we propose a hierarchical framework that combines the practical benefits of the conditional independence approach with a more appropriate characterization of the uncertainty in the resulting estimates of the climatological coefficients. First, recall from Sect. 3.1 that we have maximum likelihood estimates of the coefficients, denoted  $\hat{\boldsymbol{\theta}} = (\hat{\theta}(\mathbf{s}_1), \dots, \hat{\theta}(\mathbf{s}_n))$  (where again  $\theta \in \{\mu_0, \mu_1, \log \sigma, \xi\}$ ). This vector is an estimate of the true underlying spatial field  $\boldsymbol{\theta} = (\theta(\mathbf{s}_1), \dots, \theta(\mathbf{s}_n))$ . However,

because these estimates are obtained independently for each station,  $\hat{\theta}$  includes true spatial signal, spatial noise from storm dependence, and any additional measurement noise. The following hierarchical model links the estimates  $\hat{\theta}$  with  $\theta$ :

$$\begin{aligned} \hat{\theta} | \theta &\sim N_n(\theta, \mathbf{D}_\theta) \\ \theta &\sim N_n(\hat{\delta}_\theta, \hat{\Sigma}_\theta) \end{aligned} \tag{8}$$

following, e.g., Holland et al. (2000), Tye and Cooley (2015), and Russell et al. (2016). This model specifies that the estimates of  $\hat{\theta}$  conditional on the true field  $\theta$  are unbiased with covariance  $\mathbf{D}_\theta$ , and that the true signal  $\theta$  follows a Gaussian process as defined in Sect. 3.2. In (8),  $\mathbf{D}_\theta$  quantifies the discrepancy between  $\hat{\theta}$  and  $\theta$ , which we model as a diagonal matrix. In other words, we assume that the error in  $\hat{\theta}$  is spatially-independent. The spatial covariance  $\hat{\Sigma}_\theta$ , on the other hand, is a non-diagonal matrix that characterizes spatial coherence in  $\theta$ . Estimates of the process mean  $\hat{\delta}_\theta$ , spatial covariance  $\hat{\Sigma}_\theta$ , and spatially-independent error  $\hat{\mathbf{D}}_\theta$  are obtained via local likelihood techniques (Risser and Calder 2017; for more details see Appendix 2) using the marginalized model

$$\hat{\theta} \sim N_n(\hat{\delta}_\theta, \hat{\Sigma}_\theta + \mathbf{D}_\theta), \tag{9}$$

where we have integrated out the process  $\theta$  from the joint distribution implied by (8).

The Gaussian process assumption allows us to recover an estimate of the true process  $\theta(\cdot)$  at the station locations, denoted  $\tilde{\theta} = (\tilde{\theta}(s_1), \dots, \tilde{\theta}(s_n))$ . Conditional on  $\hat{\theta}$  as well as maximum likelihood estimates  $\hat{\delta}_\theta$ ,  $\hat{\Sigma}_\theta$ , and  $\hat{\mathbf{D}}_\theta$ , our best estimate of the true climatological coefficient process is

$$\tilde{\theta} = \hat{\delta}_\theta + \hat{\Sigma}_\theta [\hat{\Sigma}_\theta + \hat{\mathbf{D}}_\theta]^{-1} (\hat{\theta} - \hat{\delta}_\theta), \tag{10}$$

which is also known as the *kriging predictor* in a traditional geostatistical framework. Here,  $\hat{\Sigma}_\theta [\hat{\Sigma}_\theta + \hat{\mathbf{D}}_\theta]^{-1}$  is the matrix version of [signal]/[signal + noise], so we can see that best estimate  $\tilde{\theta}$  is the sum of the spatial mean ( $\hat{\delta}_\theta$ ) and a spatial residual term ( $\hat{\theta} - \hat{\delta}_\theta$ ) that is re-scaled based on the relative magnitude of the signal and noise. In fact, this relationship can be generalized to a generic location  $s' \in G$  for which we do not have observations of daily precipitation (e.g., on a fine grid). Similar to (10), our best estimate of the true climatological coefficient at  $s'$  is

$$\tilde{\theta}(s') = \hat{\delta}_\theta(s') + \hat{\mathbf{c}}_{\theta, \theta(s')}^\top [\hat{\Sigma}_\theta + \hat{\mathbf{D}}_\theta]^{-1} (\hat{\theta} - \hat{\delta}_\theta). \tag{11}$$

In (11),  $\hat{\delta}_\theta(s')$  is the estimated mean at  $s'$  and  $\hat{\mathbf{c}}_{\theta, \theta(s')}^\top = (\hat{C}_\theta(s_1, s'), \dots, \hat{C}_\theta(s_n, s'))$  is the estimated covariance between  $\theta(\cdot)$  at  $s'$  and values at the station locations. After obtaining best estimates for each of the climatological coefficients, we can then calculate best estimates of the

return values  $\tilde{\phi}_t^{(r)}(s')$  using (4) and the return periods  $\tilde{\rho}_t^{(x)}(s')$  using (5).

However, assuming that  $\mathbf{D}_\theta$  is spatially-independent (i.e., diagonal) implies that the spatial covariance  $\hat{\Sigma}_\theta$  describes both the true spatial signal as well as any spatially-correlated error from storm dependence. Therefore, our best estimates of the climatological coefficients  $\tilde{\theta}(s)$  from (11) contain both real spatial signal and spatially-correlated noise from the storm dependence. To account for this issue, we use the nonparametric or “block” bootstrap to characterize uncertainty in our estimates of the climatological coefficients as well as return values/periods. The block bootstrap requires no assumptions of independence within each year of data and preserves the spatial and temporal features of the data by re-sampling entire years of data, using the same resampled years for all station locations. We rely on the bootstrapping procedure to address storm dependence, since any real spatial signal will show up in most of the bootstrap data sets. The block bootstrap approach proceeds as follows: define  $\mathbf{y}(s_i) = \{y_t(s_i) : t = 1950, \dots, 2017\}$  to be the observed vector of the seasonal maxima for station  $i$ . Then, for  $b = 1, \dots, B$ , the bootstrap sample is obtained by drawing  $T = 68$  years from  $\{1950, \dots, 2017\}$ , denoted  $\{a_1^*, a_2^*, \dots, a_T^*\}$ , so that the  $b$ th bootstrap sample for station  $i$  is

$$\mathbf{y}_{bi}^* = (y_{a_1^*}(s_i), y_{a_2^*}(s_i), \dots, y_{a_T^*}(s_i)).$$

For each bootstrap sample, we use the multi-stage procedure outlined in Sects. 3.1 and 3.2 with (11) to obtain the best estimates of the coefficients  $\tilde{\theta}_b(s)$ , return values  $\tilde{\phi}_{b,t}^{(r)}(s)$ , and return periods  $\tilde{\rho}_{b,t}^{(x)}(s)$ . The resulting field of bootstrap standard errors for any quantity of interest, e.g., the  $r$ -return value estimate in year  $t$ , is calculated as

$$v_j^{(r)}(s) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left( \tilde{\phi}_{b,t}^{(r)}(s) - \frac{1}{B} \sum_{b=1}^B \tilde{\phi}_{b,t}^{(r)}(s) \right)^2}.$$

A potentially cleaner way to handle the storm dependence would be to include spatially-correlated error in  $\mathbf{D}_\theta$ , allowing us to explicitly separate the true spatial signal from the error. Exploratory analysis (see Appendix 1) reveal the presence of spatially-correlated error in the coefficient estimates  $\hat{\theta}$  at non-negligible scales due to storm dependence. As such, we considered using a correlated (non-diagonal) empirical estimate for the covariance  $\mathbf{D}_\theta$  in (8) (for example, Holland et al. 2000 use a bootstrap-based estimate). However, this approach resulted in several problems. First, it is extremely difficult to estimate a high-dimensional covariance matrix (here, a  $5202 \times 5202$  matrix) based on bootstrap sampling from a limited temporal record. Second, in this case we found that the signal and noise have similar spatial scales (again see Appendix 1), making it difficult to appropriately

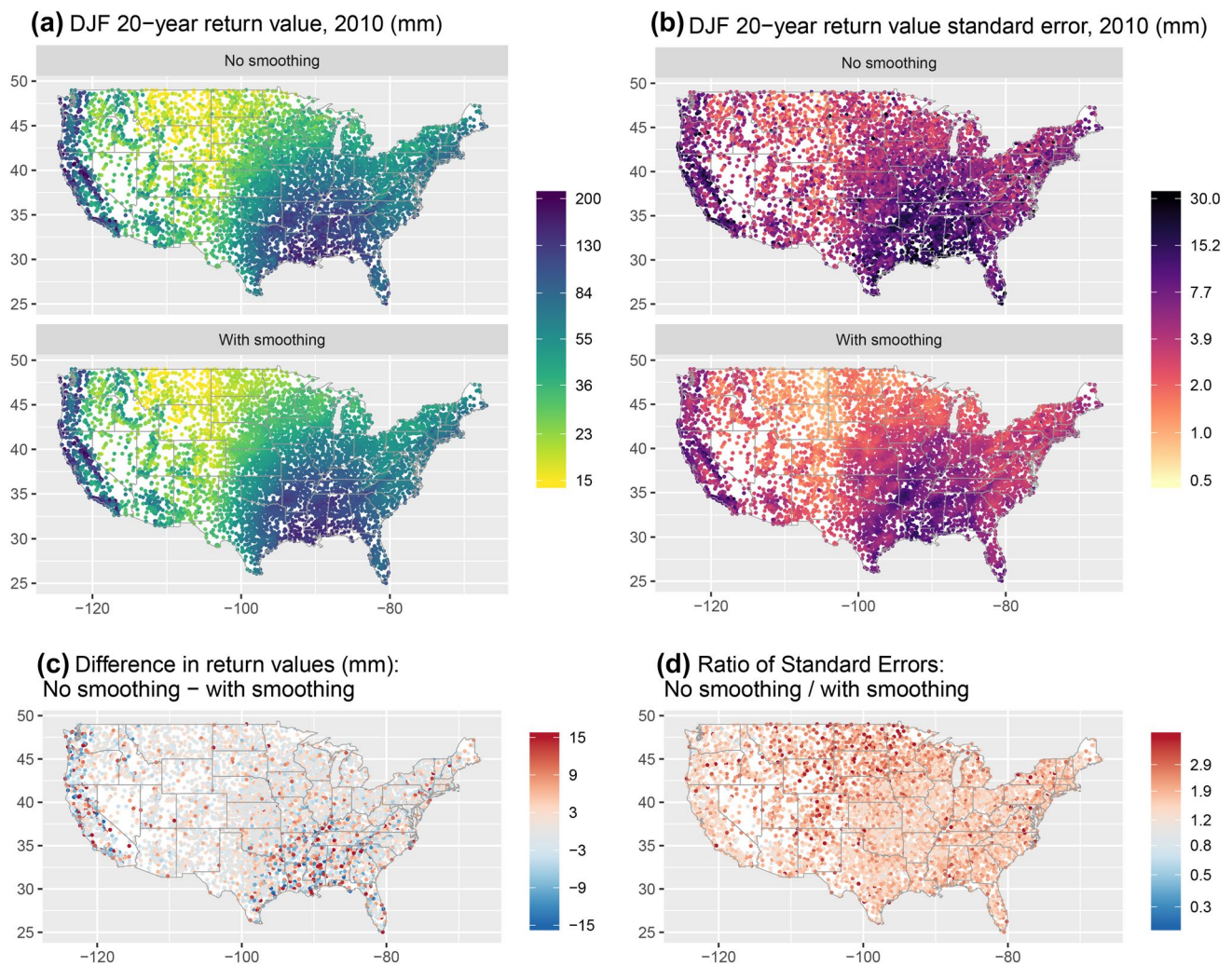
separate signal from noise. Third, we were not confident that including correlated error resulted in the correct amount of shrinkage, because the resulting uncertainty estimates from a test case were unrealistic.

## 4 Results

### 4.1 Spatial statistics and uncertainty quantification

While the original goal of this analysis was to create an improved probabilistic gridded product, the methodology yields an additional benefit of giving increased confidence in the extreme statistics of precipitation at the stations. To illustrate this concept, we first present maps of the estimated 20-year return value for DJF in 2010 at the GHCN station locations (Fig. 1; estimated return values

for the other seasons are shown in Supplementary Figure D.2). The year 2010 was somewhat arbitrarily chosen, although importantly 2010 can be directly compared with existing data products (e.g., Livneh; see Sect. 4.3). For comparison, we present results from a “traditional” extreme value analysis, where the return values are estimated independently for each station, i.e., without spatial smoothing, alongside estimates obtained from our new method. Figure 1 also shows the bootstrap standard errors in the estimated return values, again with and without spatial smoothing, as well as a direct comparison of the difference in return values and ratio of standard errors. The two approaches yield similar return values (i.e., the “signal”), with differences of less than several millimeters (except in areas with large return values, e.g., the southeast US). However, when considering the bootstrap standard errors (an estimate of the “noise”), our new approach



**Fig. 1** Comparison of 20-year return values for DJF in 2010 (mm; panel **a**) and bootstrap standard errors (mm; panel **b**) for a traditional analysis with no spatial smoothing versus our approach with spatial

smoothing. An explicit comparison of the return values and standard errors are shown in panel **c**, **d**, respectively

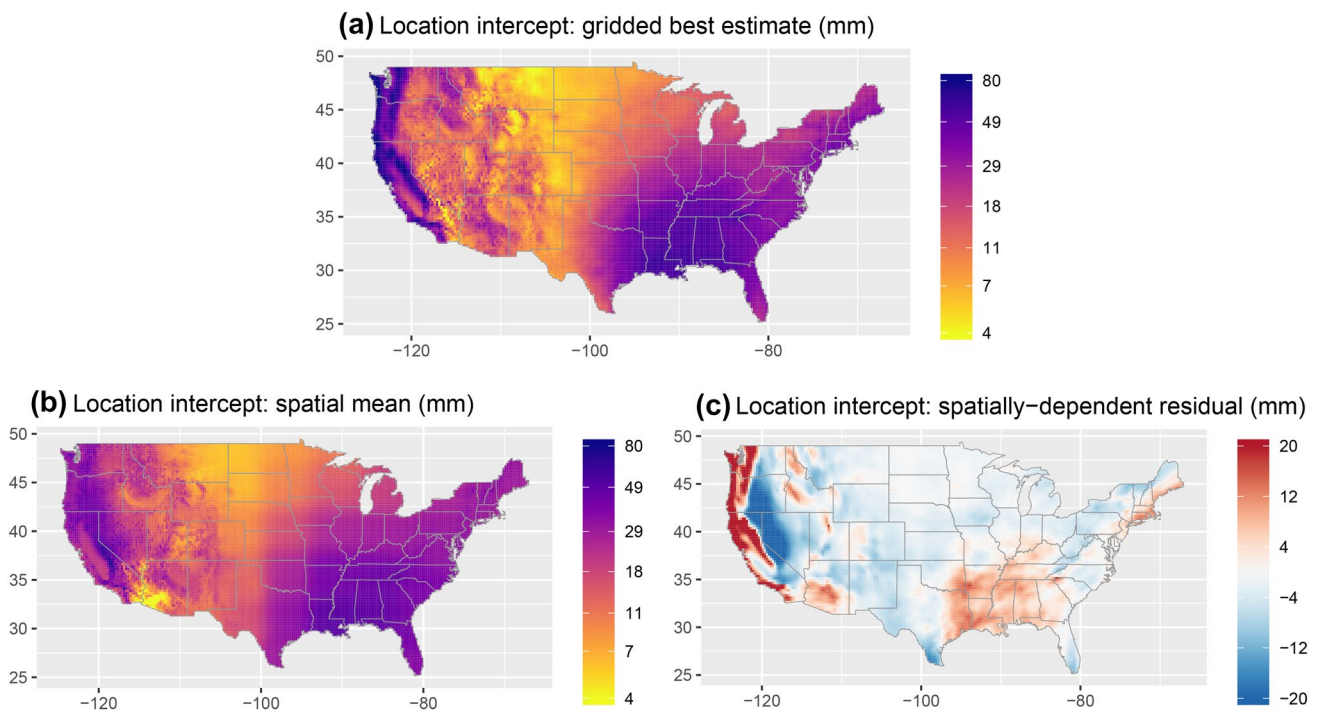


with smoothing yields much smaller standard errors across much of the domain. In other words, spatial smoothing yields approximately the same signal but reduced noise. The bottom right panel of Fig. 1 explicitly highlights this reduction in noise by plotting the ratio of standard errors with and without smoothing: for DJF, smoothing results in uncertainty estimates that are on average approximately half as large, with as much as a three-fold reduction in the upper Great Plains. The results are similar for the other seasons (see Supplementary Figure D.3), with the largest reduction in uncertainty occurring in JJA.

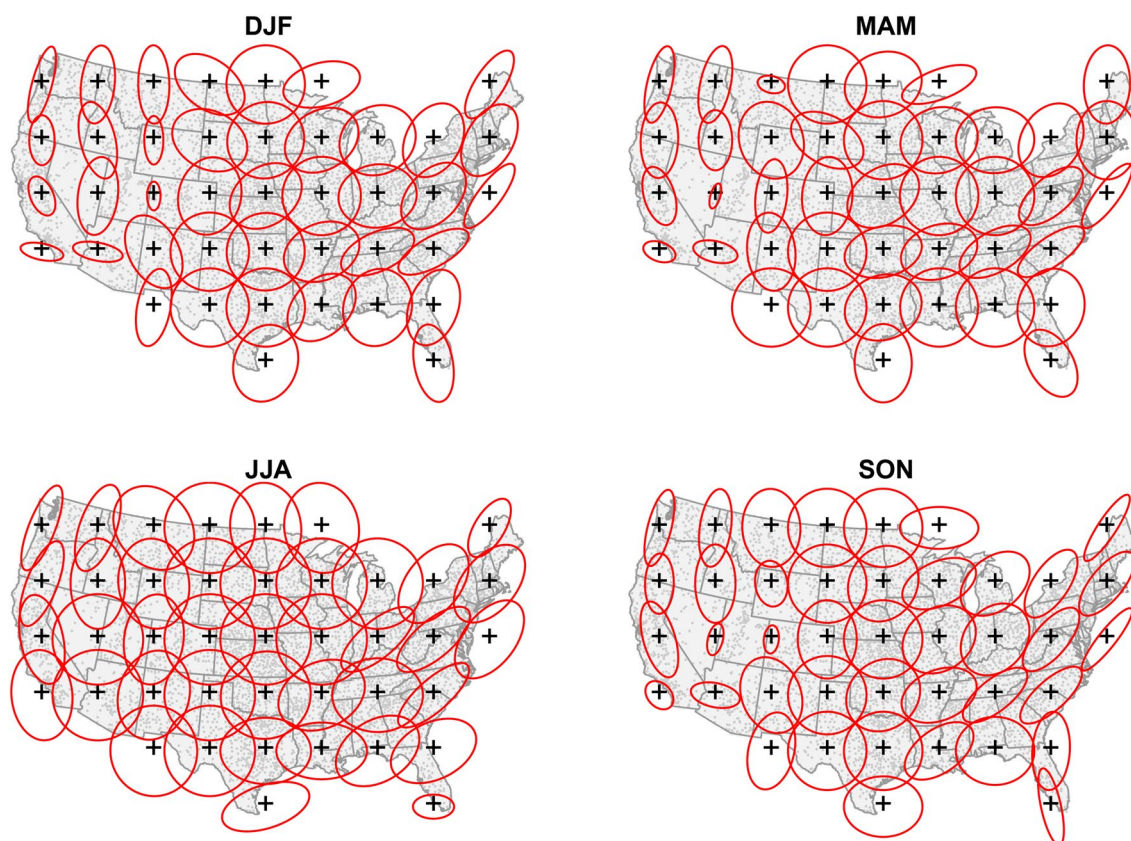
In addition to the increased signal-to-noise ratio, the use of spatial statistics in our novel approach yields insight into the behavior of extreme precipitation. As described in Sect. 3.2, nonstationary Gaussian process models allow the spatial length-scale (i.e., magnitude and direction of spatial dependence) to vary across the spatial domain. As an illustration, we show results for the location intercept coefficient  $\mu_0(\mathbf{s})$  in DJF, which characterizes the center of the extreme value distribution (constant over time) for each spatial location and hence drives the magnitude of the return values. Our best estimates of  $\mu_0(\mathbf{s})$  in DJF are shown in Fig. 2a, where we show statistically-gridded estimates (using Eq. 11) as it is easier to visualize the spatial distribution with a “filled-in” map. Figure 2 also shows the elevation-based spatial mean ( $\hat{\delta}_\theta(s')$  in Eq. (11) in Fig. 2b, as well as the spatially-dependent residual

$(\hat{\mathbf{c}}_{\theta, \theta(s')}^\top [\hat{\Sigma}_\theta + \hat{\mathbf{D}}_\theta]^{-1} (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\delta}}_\theta))$  in Eq. (11) in Fig. 2c. (Corresponding plots for MAM, JJA, and SON are shown in the Supplementary Figures D.4, D.5, and D.6.) The spatial mean in Fig. 2b represents the estimated first-order properties of  $\mu_0(\mathbf{s})$ , and characterizes a linear relationship between elevation and the center of the extreme value distribution. The use of a spatial Gaussian process yields the spatially-dependent residual term, which characterizes additional nonlinear relationships over space in  $\mu_0(\mathbf{s})$  that are not captured by the elevation covariate.

Focusing on the spatially-dependent residual, first note that the residuals are smooth over the southeast United States and upper Great Plains but highly heterogeneous across the Rocky Mountains and along the west coast. This is not surprising, but note that these differences can be explicitly characterized by the nonstationary Gaussian process, which estimates this length-scale directly from the data—and, these differences persist even after accounting for elevation directly via the spatial mean in Fig. 2b. (Note that elevation artifacts can still be seen in Fig. 2c, indicating some inadequacy in our characterization of the elevation-based mean.) To visualize these differences, consider Fig. 3, which shows a heuristic representation of the spatially-varying length scale (both magnitude and direction) via a set of ellipses, which are estimated locally across CONUS at the mixture component locations. Intuitively, a long/skinny ellipse (e.g.,



**Fig. 2** Gridded best estimates of the location intercept  $\mu_0(\mathbf{s})$  over CONUS (mm) for DJF (panel **a**), decomposed into the elevation-based spatial mean (panel **b**) and the spatially-dependent residual (panel **c**). In other words, panel **a** is the sum of panels **b**, **c**



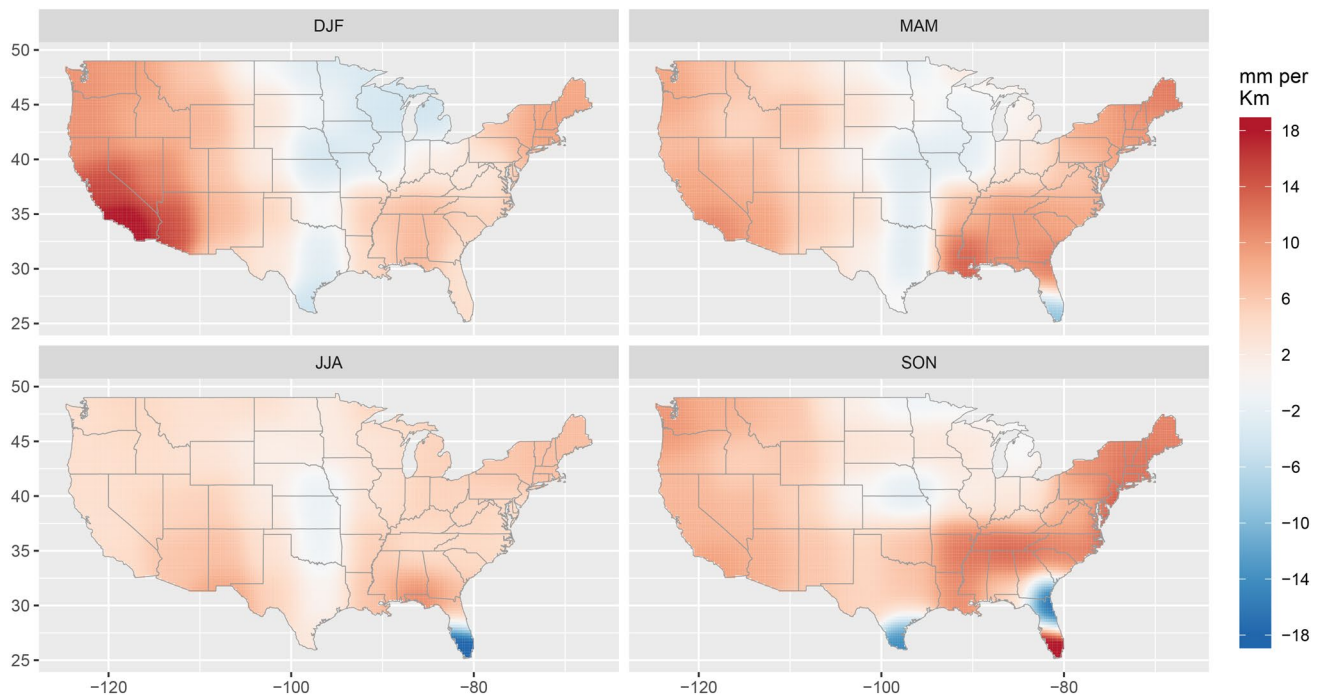
**Fig. 3** Directional spatial length-scale for the location intercept  $\mu_0(\mathbf{s})$  for each season, estimated empirically from the data. The ellipses are a heuristic representation of the magnitude and direction of spatial dependence in  $\mu_0(\mathbf{s})$

coastal Washington state in DJF) means that the length scale of spatial dependence is much larger in one direction than in the orthogonal direction; a circular ellipse (e.g., western Illinois in DJF) means that the length-scale is roughly the same in all directions. Similarly, a large ellipse corresponds to long-range spatial dependence while a small ellipse corresponds to shorter-range spatial dependence. Note that the Gaussian process estimates translate to the variations in heterogeneity noted in the gridded spatial residuals: in DJF, the Rocky mountain range and west coast generally has a shorter length-scale (i.e., highly heterogeneous) while the southeast US and upper Great Plains generally have longer length-scales (i.e., highly smooth or homogeneous). There are other interesting features to these ellipse maps, e.g., coastal effects (both east and west), as well as the southwest-to-northeast orientation of the ellipses across the southeast US (particularly in SON). Also, note the seasonal differences in the spatial length-scale, particularly DJF vs. JJA, which could be due to either the climatology or the different storm types leading to extreme precipitation in winter and summer.

To be clear, estimating a spatially-varying length scale directly from the data reiterates the importance of using a statistical approach to gridding the GEV coefficients and

hence the return values: the data itself informs the degree of smoothing, which varies over the domain. This approach is more physically meaningful than heuristic approaches like bilinear interpolation or inverse-distance weighting, or even “ordinary kriging” using an isotropic Gaussian process, which uses a constant (and circular) spatial length-scale.

As mentioned in Sect. 3.2, we include an orographic correction in the nonstationary Gaussian process mean for each GEV coefficient. Intuitively, this orographic correction involves estimating a linear relationship between elevation and each GEV coefficient empirically from the data, while allowing the magnitude and sign of this relationship to vary smoothly across CONUS. While the mean is linear in elevation, note that this does not completely specify the relationship between extreme precipitation and elevation because of the spatial smoothing induced by the Gaussian process (this is illustrated in Fig. 2). Figure 4 shows the data-driven, spatially-smoothed estimate of the relationship between extreme precipitation and elevation, across each season. Dark red areas indicate a strong positive relationship between elevation and extreme precipitation (i.e., higher elevation corresponds to stronger storms), while blue areas indicate a negative relationship. Across



**Fig. 4** Spatially-varying elevation correction for the Gaussian process mean of  $\mu_0(\mathbf{s})$ , across each season (mm of precipitation per km of elevation). Positive values indicate larger return values for higher elevations; negative values indicate smaller return values for higher elevations

all of the seasons, the relationship between elevation and extreme precipitation is generally positive, with the largest effects showing up in the western United States. Of course, elevation itself is rather homogeneous over much of the central United States. However, in the western United States (where elevation plays an important role in the climatology of precipitation), there are important seasonal differences in the relative elevation/precipitation relationship: for example, especially in California, the effect is larger for DJF than JJA. This is not surprising, as there is a strong wet/dry seasonal cycle in the western United States, but it is encouraging that the nonstationary GP was able to infer this directly from the data.

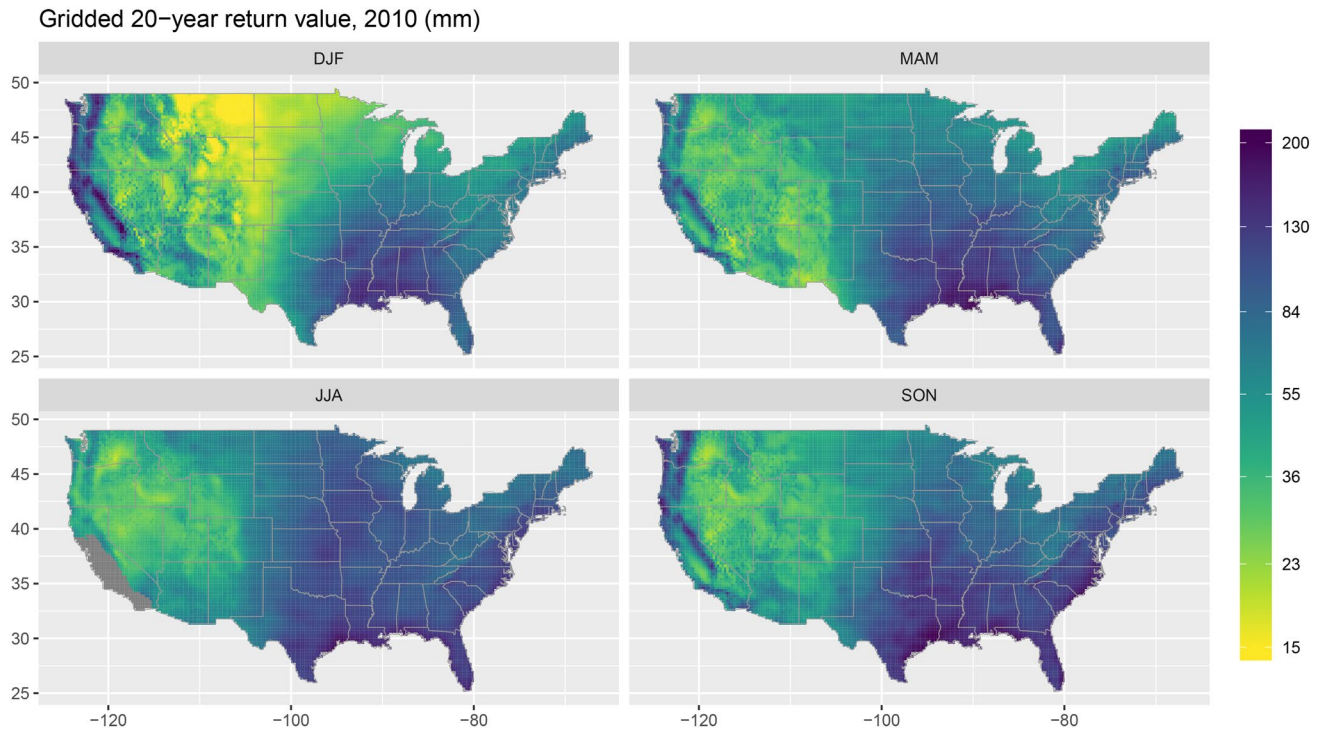
Of course, many gridded daily products also include an orographic correction, most notably the Parameter-elevation Relationships on Independent Slopes Model (PRISM; see Daly et al. 2008 and the references therein). While our orographic correction is applied to the climatological coefficients rather than daily precipitation itself, several notes of comparison should be made. First of all, like PRISM, our approach involves a linear relationship between the climate variable and elevation (compare Eq. 2 in Daly et al. 2008 with Eq. 14 in the Appendix). However, PRISM proceeds to incorporate a distance weighting scheme (see Eq. 3 in Daly et al. 2008) that is fixed *a priori* and spatially constant (albeit elevation-dependent). Our approach with nonstationary Gaussian processes, on the other hand, estimates the

appropriate length scale for distance weighting directly from the data, using a non-constant weighting scheme over space. Furthermore, the Gaussian process prediction (via Eq. 11) implicitly accounts for the over-representation issue in distance-weighted averaging (such as in PRISM) while also explicitly removing noise in the data due to measurement error. Finally, it is worth restating that the most important difference between our approach and PRISM is in *what* is actually being smoothed; whereas PRISM smooths precipitation itself, our method smooths GEV distribution parameters.

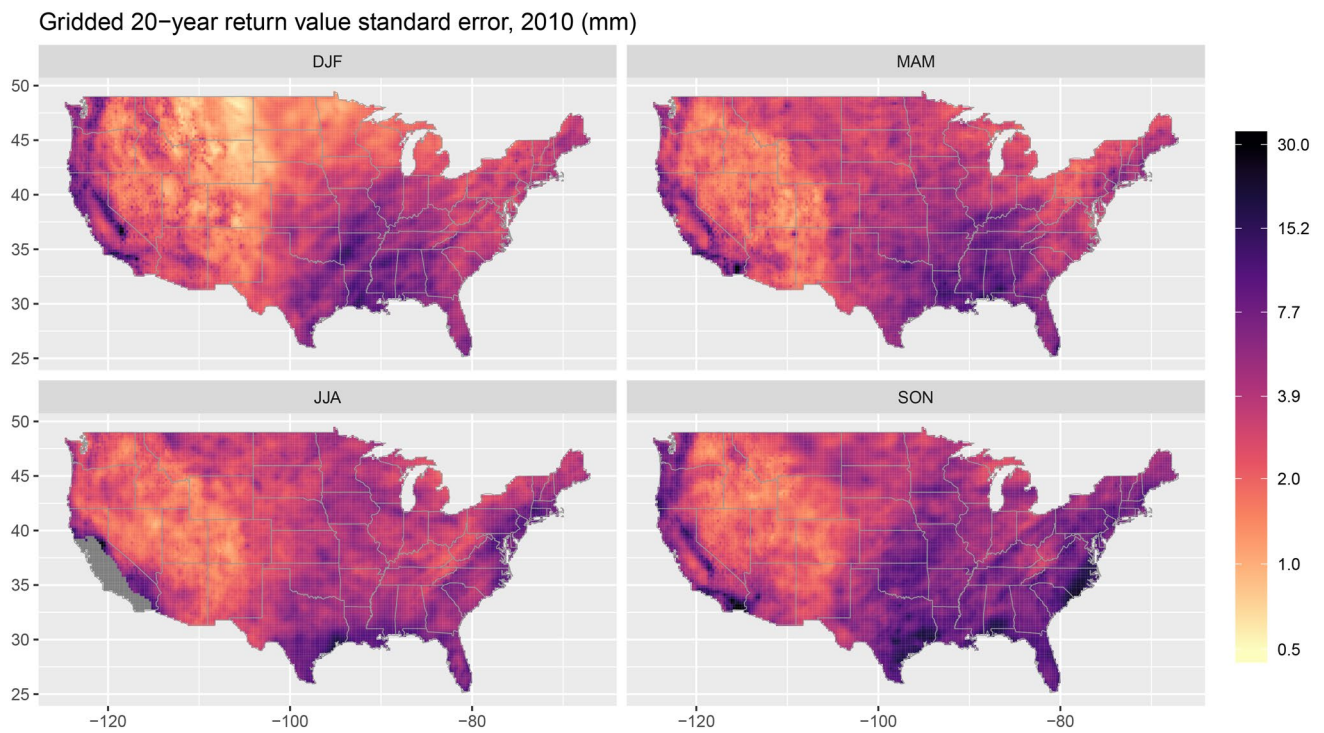
## 4.2 Probabilistic gridded product

Returning to the main goal of the analysis, we now present our new probabilistic gridded product. The product consists of spatially complete maps (on the  $0.25^\circ$  grid) of the four climatological coefficients in (2): a best estimate map calculated from the full data, as well as 250 smoothed maps calculated from the bootstrap data sets. With these gridded data in hand, one can create summary statistics for the entire spatial domain as well as bootstrap sampling distributions for confidence intervals or standard errors. Furthermore, these calculations are possible for any individual year in 1950–2017 (note, however, that we have not explicitly accounted for year-to-year variability due to, e.g., ENSO). The most common summaries would likely be return values (4) or return periods (5), but one could just as easily calculate risk ratios





**Fig. 5** Spatially-complete 20-year return values (mm) for CONUS, across each season. Part of California is masked out for JJA because the dry season in this region means that the stations in this area do not register any “extreme” precipitation measurements



**Fig. 6** Spatially-complete bootstrap standard errors (mm) for the 20-year return value across each season. The standard error provides an estimate of the uncertainty in the return values shown in Fig. 5.

Part of California is masked out in JJA due to seasonal dryness (see the caption of Fig. 5)



for comparing one time point with another (see, e.g., Risser and Wehner 2017) or even trends over time.

As an illustration, Fig. 5 shows spatially complete maps of the 20-year return value for each season, and the bootstrap standard errors are shown in Fig. 6. These maps are unique in that they are high resolution, and they produce return-value estimates consistent with what would be measured at a point (rather than over an area). As a result, the return value estimates from our method are substantially higher than those estimated from gridded datasets. For example, Figure A4b in Volosciuk et al. (2015), shows 20-year DJF return values estimated from the gridded Climate Prediction Center daily precipitation product (the equivalent of the DJF panel in our Fig. 5), and the highest daily return values shown are only about 125 mm, in comparison to values approaching 200 mm estimated using our method. We explore this comparison more deeply in Sect. 4.3.

The influence of topography, which is explicitly included in the spatial model (see Sect. 4.1) is clearly evident throughout Fig. 5, as the Sierra Nevada topographic variability associated with the Basin and Range province is clearly evident: high topography areas generally have larger return values. Notably, however, the influence of topography in the western U.S. is substantially damped during the summer convective season, with the Sierra Nevada and Cascade ranges hardly evident. This variation in orographic influence, which emerges naturally from the data itself, is consistent with what one might physically expect: orographic influence in the western U.S. is most prevalent during the winter season, when storm systems have a strong zonal flow that is roughly perpendicular to the dominant orography, and it is weaker during the season when precipitation is primarily convective. Note that Fig. 3 shows that the spatial dependence of DJF extreme precipitation in the western U.S. is more meridional than zonal and appears to be aligned with the dominant orientation of topography in the vicinity. This would indicate that though winter storms impinging on the western U.S. have a predominantly zonal flow, extremes tend to spatially co-occur along meridionally-oriented (and somewhat perpendicular) mountain ranges that intercept the flow. Intriguingly, however, orographic modulation of extremes is also absent in the southwestern U.S., which is the center of action during the monsoon season when large moisture fluxes from the south bring summer precipitation that one might expect to also cause orographically modulated precipitation. It is plausible that this is because the direction of the incoming moisture flux is roughly parallel, rather than perpendicular, to the dominant direction of ranges in the Basin and Range province, which may weaken the orographic effect.

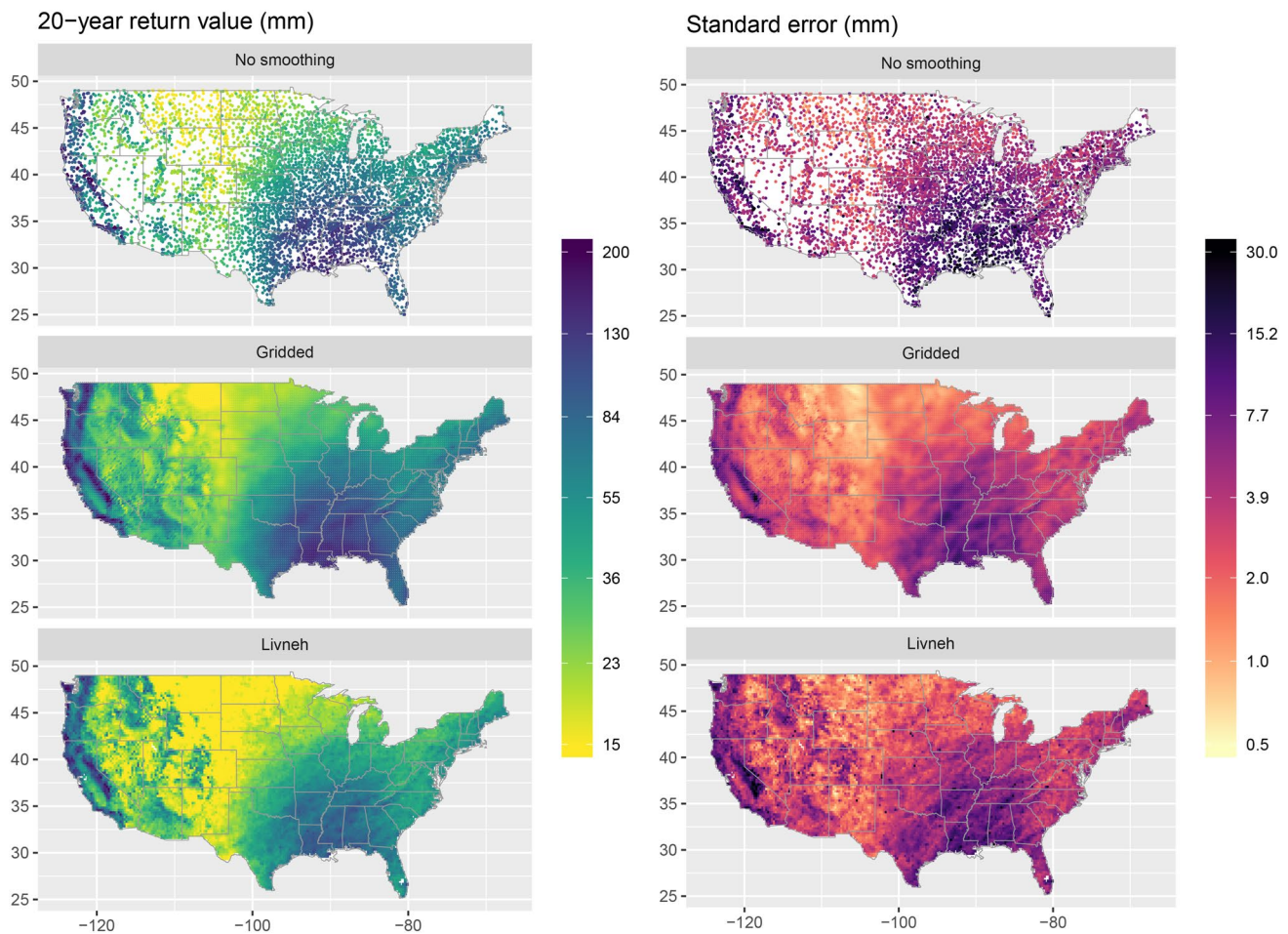
Our probabilistic gridded product is freely available via a public repository on the Harvard Dataverse (Risser et al. 2019; <https://bit.ly/2CXdnuj>). The data are packaged

together into network common data form (netCDF) files. Three separate files are available, with data provided separately for each season: (1) best estimates of the GEV coefficients (e.g., Fig. 2a) as well as bootstrap standard errors, (2) 10-, 20-, 50-, and 100-year return values (with bootstrap standard errors) for 1955, 1965, 1975, 1985, 1995, 2005, and 2015, and (3) smoothed bootstrap estimates (250 total) for each GEV coefficient (these are used to calculate the bootstrap standard errors in the first two files). We also provide code that can be used to calculate return values, return periods, or any function thereof for any year in 1950–2017, as well as the corresponding bootstrap sampling distribution (which can be used to quantify uncertainty via standard errors or confidence intervals).

### 4.3 Comparison to the Livneh data product

As mentioned in Sect. 1, gridded data products are often used in place of station data to calculate the extreme statistics of precipitation. In light of our novel approach for calculating extreme statistics over CONUS, we now compare our results with a more traditional analysis using the Livneh gridded data product (Livneh et al. 2014). Using this data product, we conduct the extreme value analysis described in Sect. 3.1 independently for each grid cell over CONUS, using the block bootstrap (but without spatial smoothing) to calculate uncertainties.

A comparison of the DJF 20-year return values and standard errors for 2010 is shown in Fig. 7: this plot shows estimates without spatial smoothing (also shown in Fig. 1) and our gridded product, as well as estimates based the Livneh data product. Livneh return values and standard errors for other seasons are show in Supplementary Figures D.7 and D.8, which can be directly compared with Figs. 5 and 6. A visual comparison of the plots in Fig. 7 reveals two immediate points: (1) the return values calculated from the Livneh data product are systematically smaller than the return values generated using the station data, both smoothed and raw, and (2) the standard errors from both approaches appear to be approximately the same. In other words, our gridded product produces return values that are consistent with station estimates but higher than estimates from gridded data, while reducing uncertainty relative to independent station estimates. Figure 8 shows a direct comparison of these two quantities by plotting our gridded estimates versus the corresponding Livneh estimates: in the top row, note that the station data return values are uniformly larger than the Livneh return values, and that the bias is worse for the largest return values. In the bottom row of Fig. 8, however, we can see that the scatterplots are clustered around the 45° line, meaning that the standard errors are comparable. There is, however, a tendency for the Livneh standard errors to be larger than



**Fig. 7** DJF 20-year return values for 2010 (left) and standard errors (right) for an analysis with no smoothing (top), our gridded product (middle), and for an analysis using the Livneh data product (bottom)

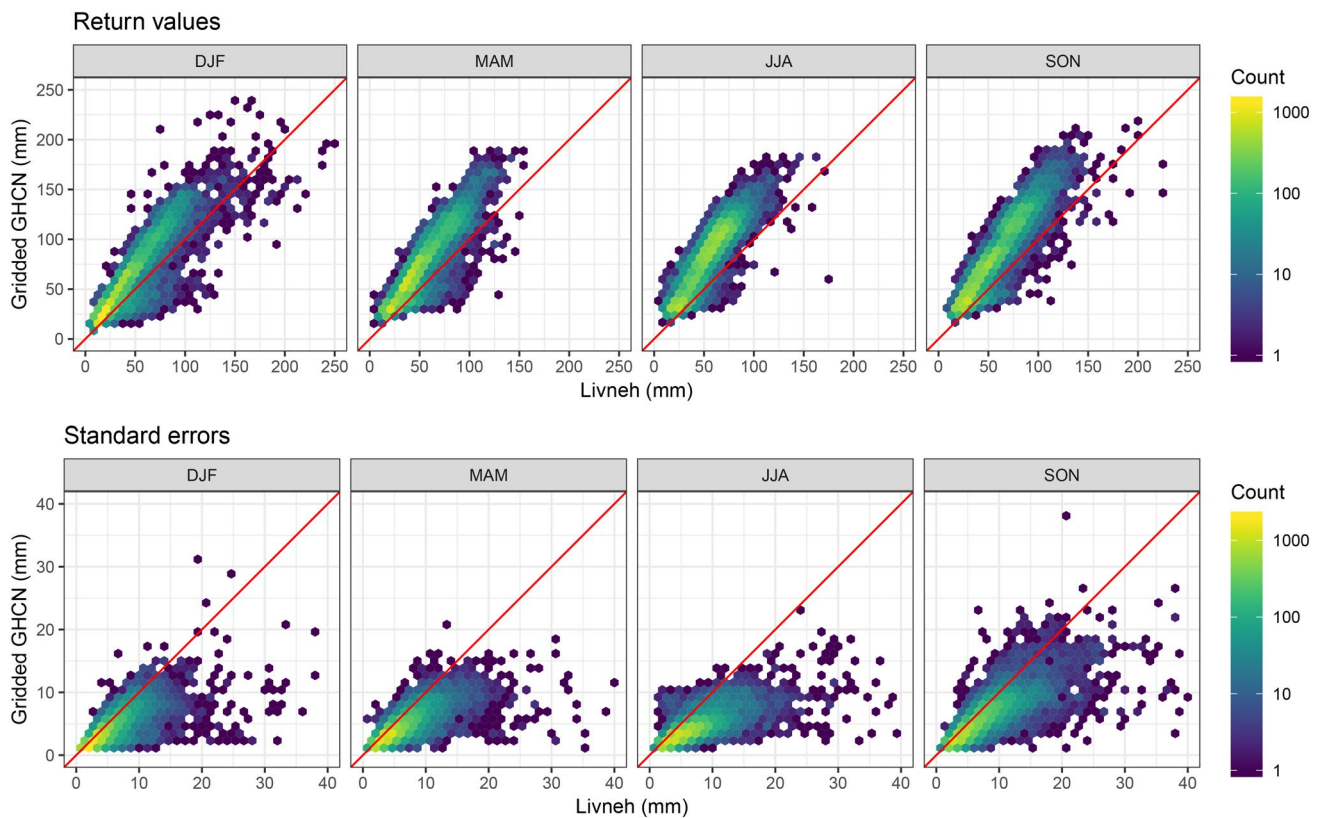
the GHCN standard errors, particularly for locations with greater uncertainty. And, since the Livneh return values are smaller than the GHCN values, this tendency for large Livneh standard errors is increased when they are expressed as percentages.

## 5 Conclusions

In this paper, we have developed novel statistical methodology for conducting a spatial analysis of extreme precipitation for a large network of weather stations over a heterogeneous domain. Using Gaussian processes, our approach uses data-driven smoothing to borrow strength over space, which yields increased confidence in our subsequent estimates of the climatology of extreme precipitation. The result of our analysis is a new “probabilistic” gridded product specifically designed for characterizing extreme precipitation, which will be made freely available in a public data repository.

Compared to traditional gridded products, our results yield important differences in estimates of return values. Furthermore, our methodology is able to produce spatially-complete, high resolution maps of return values based on irregularly observed station data, and we are therefore able to characterize extreme statistics of precipitation at small spatial scales—especially relative to the large, spatially-averaged summaries provided in, for example, Kunkel et al. (2013) or Easterling et al. (2017). As a result, our new gridded product provides important insight into the underlying physics of precipitation over CONUS (both in the spatial length-scale differences and further exploration of the extreme statistics themselves), and the return value maps are of value with respect to impacts of extreme precipitation, resource management, and infrastructure design.

An important extension of the results presented in this paper involves characterizing high-resolution observed trends in extreme precipitation. In future work we plan to explore the presence and statistical significance of any trends over



**Fig. 8** Seasonal density scatter plots of return values and their standard errors, with quantities from the Livneh data product on the  $x$ -axis and corresponding quantities from our GHCN station data analysis on the  $y$ -axis. The red line indicates the 45° line

time in the statistics of extreme precipitation. Such an exploration naturally requires incorporation of more sophisticated trend models where the effects of interannual or decadal variability (e.g., the El Niño/Southern Oscillation or the Pacific Decadal Oscillation) are explicitly accounted for. However, introducing additional time-varying covariate information into a statistical model for the extreme statistics of precipitation necessitates robust methodology for selecting the variables that are relevant for explaining year-to-year variability.

In conclusion, we note that the correct interpretation and usage of gridded products closely depends upon their construction. In the methodology presented here, the long period return values within a grid box are interpreted as representative of any point within that grid box. A simple interpretation is that for a given return period, any point within the grid box has the same return value. On the other hand, modeled daily precipitation is the total precipitation within a grid box integrated at all points within the box as constrained by the conservation properties of the model. Hence, when compared to return values from a climate model, observed gridded return values obtained by our method should be considered an upper bound. In other words, if during an extreme storm, simulated precipitation is such that every point within a grid box is simultaneously precipitating at a

very high rate then we would expect the modeled return values to agree with our observational estimates. This situation could indeed occur if the model's grid cells are small enough and if storm properties are relatively spatially uniform on this scale. Hence, we might expect that our estimates of mid-latitude winter storm extreme precipitation would agree with the very high resolutions of cloud system resolving models, but not for convective summer storms, which would always be simulated lower than our estimates because of these scale considerations.

Alternatively, one might expect that return values based on gridded daily observed precipitation products, as in Wehner (2013), might be more directly comparable to modeled precipitation if the gridding process is conservative. However, because the station density is low compared to model grid boxes, the probability of missing an extreme rainfall event with the grid box is high. Hence, return values calculated from a daily gridded precipitation product provide a lower bound on what climate models should be expected to produce. As a model evaluation strategy, we can utilize these two differing ways of estimating observed long period precipitation return values as model performance metrics by considering them as upper and lower bounds on expected model targets. Typically, models at CMIP5-class horizontal



resolutions produce seasonal return values that are lower than observational estimate, because simulated gradients of temperature and moisture are weaker than observed. As resolution is increased to  $\sim 25$  km, simulated mid-latitude winter extreme precipitation typically compares more favorably with this lower bound, but simulated summer extreme precipitation can be too high, even compared to the upper bounds, due to deficiencies in cumulus convection parameterizations (Wehner et al. 2014). It would be reasonable to expect that as higher resolution multi-decadal simulations become available that simulated mid-latitude winter extreme precipitation will be increased but should be less than the lower bound. Improvements in simulated mid-latitude summer extreme precipitation await convection-permitting resolutions.

**Acknowledgements** The authors would like to thank members of the 2018 Statistical and Applied Mathematical Sciences Institute (SAMSI) working group on extreme precipitation (specifically Daniel Cooley, Brook Russell, Richard Smith, Ken Kunkel, and Whitney Huang) as well as Christina Patricola and Benjamin Timmermans for helpful discussions throughout the development of this work. Furthermore, we would like to thank the reviewers of this manuscript for helpful and constructive feedback that improved the quality of the paper. The data supporting this article are based on publicly available measurements from the National Centers for Environmental Information (at <ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/>). This research was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract no. DE-AC02-05CH11231 and used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the U.S. Department of Energy, under Contract no. DE-AC02-05CH11231. This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

**OpenAccess** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## Appendix 1: Exploratory analysis

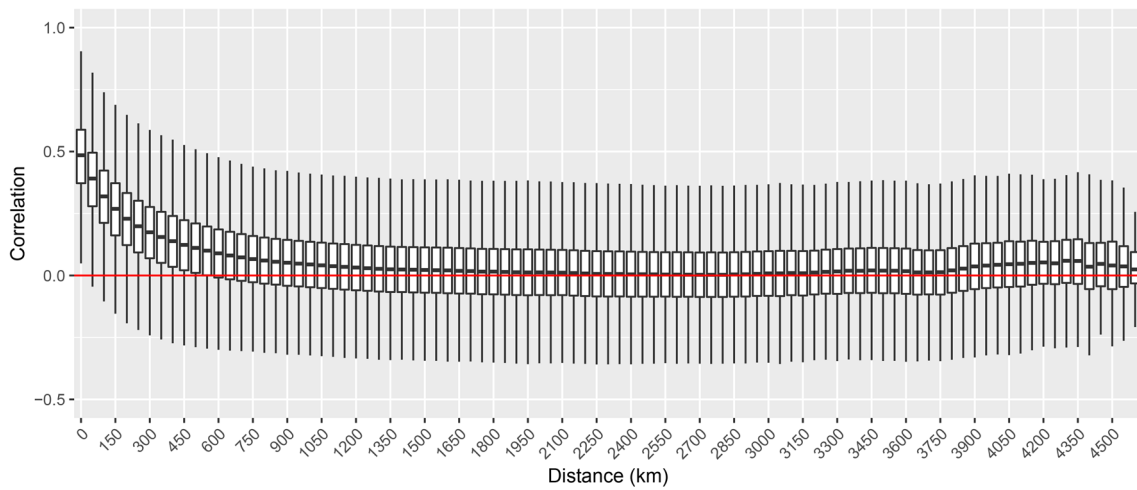
As described in Sect. 1, there is clearly spatial dependence in the underlying fields of daily precipitation (due to storm systems; here referred to as “storm dependence”) that is

not accounted for when conducting the GEV analysis on seasonal maxima separately for each station. As such, fields of the pointwise maximum likelihood estimates contain both signal and noise, because the true underlying signal is contaminated by error due to the unaccounted-for storm dependence (note, however, that simple sampling noise or measurement error would occur even without storm dependence). The block bootstrap approach outlined in Sect. 3.3 can be used to approximate the sampling distribution of parameter estimates at each station and quantify the spatial features of this error. Since we use the same bootstrap samples for all stations, we can also obtain estimates of the pairwise correlations between sampling distributions. Accounting for this correlation allows us to explicitly separate the signal and noise present in the maximum likelihood estimates.

As a concrete example, consider the following exploratory plots for the location parameter intercept  $\mu_0(s)$  for DJF. To explore the co-variance of the pointwise maximum likelihood estimates, we can use the pairwise empirical correlations from the bootstrap samples across all pairs of stations. The easiest way to visualize these correlations simultaneously is via a correlogram, where we bin the pairwise correlations by distance and present box plots for each bin. The correlogram for the location intercept in DJF is shown in Fig. 9. Note that there is non-negligible correlation up to approximately 1000 km, indicating that the spatial range of correlation in the bootstrap errors is quite large at least for some regions of CONUS.

To visualize these correlations more directly for individual stations, consider the plots in Fig. 10, which show spatial maps of the empirical correlations between six selected reference stations and all other stations in CONUS. The plot in the top left of Fig. 10 (reference location 1) is typical of the correlation map for many stations across CONUS: there appears to be moderate correlations (i.e., approximately 0.5) at very short distances and nonzero correlations extending over a large swatch of the country. The reference station in California (reference location 2) shows strong local correlation, but this correlation dies off rather quickly. Reference locations 3–6 show very strong local correlations that extend over large, irregular areas. Given that we are considering DJF and reference locations 3–6 are in the center of the United States, these large clusters of strong correlation are likely the direct result of the winter storm systems that are common in this area. The bottom left plot (reference location 4) is particularly interesting, as this shows a strong yet narrow band of large correlation, oriented in a southwest-to-northeast direction. Again, while some of these correlations may be spurious, there are definitely non-negligible correlations for many reference locations that extend over large spatial domains.

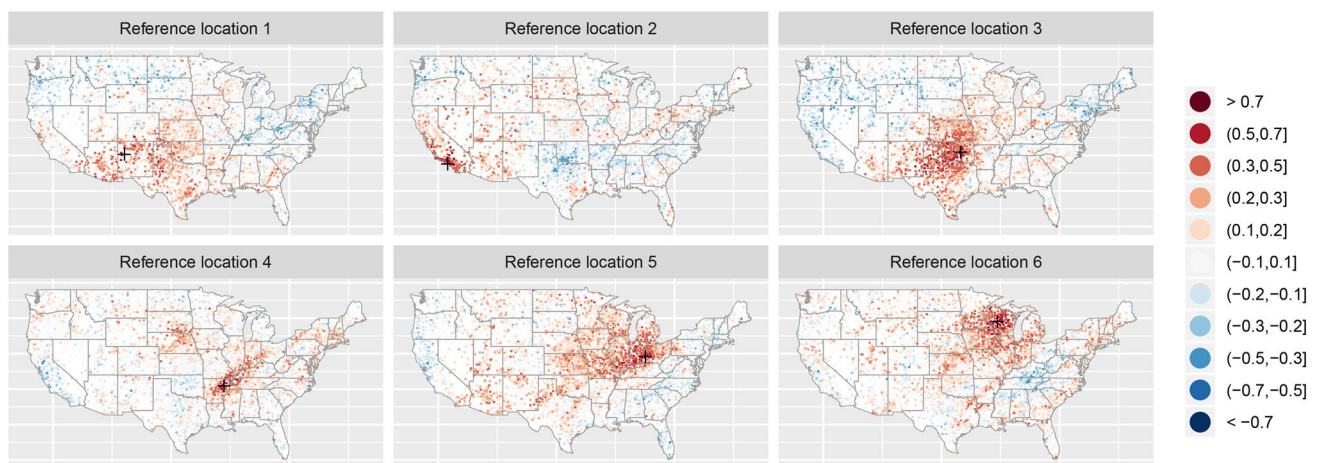




**Fig. 9** Correlogram of bootstrap correlations for the location parameter intercept in DJF

These explorations confirm the presence of spatially-correlated error in the coefficient estimates at non-negligible scales, and that this noise is a result of the unaccounted-for storm dependence. It is therefore important to account for this error in the second step of the analysis where we estimate the first- and second-order properties of the true spatial signal (Sect. 3.2). Unfortunately, it is not entirely clear how to do this, since the spatial coherence of both signal and error appear to be on similar spatial scales. Furthermore, it is obvious that the correlation patterns are highly irregular and long-range (which could be evidence of teleconnections), so we would prefer to use an empirically-based approach rather than complex statistical modeling. The simplest approach would be to use the empirical (sample) covariance matrix of the bootstrap samples as a fixed plug-in estimate for the true error covariance  $\mathbf{D}_{varvec{\theta}}$ , but unfortunately the sample covariance is

known to be a poor estimate of the true covariance. Furthermore, since  $B = 250$  (the number of bootstrap samples) is less than  $n = 5202$  (the number of locations; in practice  $B$  could be made arbitrarily large), the sample covariance is singular. Of course, the statistics literature describes a wide variety of strategies for covariance regularization, including modeling the covariance separately in terms of standard deviations and correlations (Barnard et al. 2000) and a variety of shrinkage-based approaches (e.g., Daniels and Kass 2001; Schäfer and Strimmer 2005; Hannart and Naveau 2014), among many others. However, it is not immediately clear what type of regularization is appropriate, as some of the long-range correlations shown in Fig. 10 are real. Furthermore, when we implemented a test case where  $\mathbf{D}_{varvec{\theta}}$  was fixed as the sample bootstrap covariance, we were not convinced that the spatial smoothing/shrinkage was being done correctly (the fact that the



**Fig. 10** Spatial maps of bootstrap correlations for several reference points for the location parameter intercept in DJF

signal and error have similar spatial scales may be causing problems). Therefore, as mentioned in Sect. 3.3, we used a simple (i.e., diagonal) model for the error covariance estimated directly from the pointwise maximum likelihood estimates, relying on the block bootstrap to account for the storm dependence. Further exploration of this issue is beyond the scope of this paper and will be the subject of future methodological work.

## Appendix 2: Spatial statistical modeling for the coefficient fields

Recall from Sect. 3.2 that we account for dependence across the climatological coefficients over the spatial domain using second-order nonstationary spatial Gaussian process models for each of the spatially-varying climatological coefficients in (2). Given the two-stage nature of the analysis, recall from Sect. 3.3 that we estimate the Gaussian process mean  $\delta_\theta$ , covariance  $\Sigma_\theta$ , and error covariance  $\mathbf{D}_\theta$  using the marginalized model

$$\hat{\theta} \sim N_n(\delta_\theta, \Sigma_\theta + \mathbf{D}_\theta), \quad (12)$$

where  $\delta_\theta = (\delta_\theta(\mathbf{s}_1), \dots, \delta_\theta(\mathbf{s}_n))$ , the  $ij$  element of  $\Sigma_\theta$  is  $C_\theta(\mathbf{s}_i, \mathbf{s}_j)$ , and  $\mathbf{D}_\theta$  is a diagonal matrix of elements  $\tau_\theta^2(\mathbf{s}_i)$ . In other words,  $\tau_\theta^2(\mathbf{s}_i)$  quantifies the variability of the error at station  $\mathbf{s}_i$ .

In general, using a Gaussian process to estimate a second-order nonstationary covariance function  $C_\theta$  for a large data set is computationally demanding. First of all, for a data set of size  $n$ , a single evaluation of the multivariate Gaussian likelihood requires  $\mathcal{O}(n^2)$  storage and  $\mathcal{O}(n^3)$  calculations, which is costly (yet not impossible) for  $n = 5202$ . More seriously, many nonstationary covariance function models are highly parameterized (we refer the interested reader to Risser 2016), and parameter estimation becomes very difficult for large  $n$ . To navigate both of these difficulties, we elect to use the covariance function outlined in Risser and Calder (2017), which uses local-likelihood estimation and a mixture component technique to estimate the high-dimensional covariance function parameters. The essence of the method is that we first estimate the statistical parameters of (12) locally and then use a Gaussian smoothing kernel to interpolate the local estimates to the station locations. Formally, the method starts out by defining a coarse grid of  $K$  “mixture component” locations over the spatial domain, denoted  $\{\mathbf{b}_k : k = 1, \dots, K\}$ . Then, independently for each  $\mathbf{b}_k$ , a set of statistical parameters are estimated from a stationary (anisotropic) Gaussian process with a mean that is linear in an elevation-based covariate (as well as an intercept) using data from all of the stations that lie within a particular radius  $r$  of  $\mathbf{b}_k$ . Let  $\phi$  represent an arbitrary parameter in the stationary Gaussian process model, i.e., one of the mean regression coefficients, spatial variance, and spatial covariance

parameters (for simplicity, we drop the  $\theta$  subscript, but note that these parameters are estimated separately for each climatological coefficient). Then, conditional on mixture component estimates  $\{\hat{\phi}_k : k = 1, \dots, K\}$ , the parameter estimate for an arbitrary location  $\mathbf{s}$  is

$$\hat{\phi}(\mathbf{s}) = \sum_{k=1}^K w_k(\mathbf{s}) \hat{\phi}_k, \quad (13)$$

where  $w_k \propto \exp\{-||\mathbf{s} - \mathbf{b}_k||^2/(2h)\}$  such that  $\sum_{k=1}^K w_k(\mathbf{s}) = 1$ . These estimates are plugged into the nonstationary covariance function outlined in Risser and Calder (2017), which was originally derived in Paciorek and Schervish (2006) and Risser and Calder (2015), as well as the diagonal error covariance. The estimate for each component of the mean of (12) is

$$\hat{\delta}(\mathbf{s}) = \hat{\beta}_0(\mathbf{s}) + \hat{\beta}_1(\mathbf{s})x(\mathbf{s}), \quad (14)$$

where  $x(\mathbf{s})$  is an elevation-based covariate at location  $\mathbf{s}$  and  $\hat{\beta}_0(\mathbf{s})$  and  $\hat{\beta}_1(\mathbf{s})$  are estimated as in (13) from the local coefficient estimates  $\{\hat{\beta}_0^k, \hat{\beta}_1^k : k = 1, \dots, K\}$ .

To implement the above statistical model for a given application, however, one must make several choices:

1. A set of mixture component locations  $\{\mathbf{b}_k : k = 1, \dots, K\}$
2. The local fit radius  $r$
3. The bandwidth parameter  $h$  in the smoothing kernel

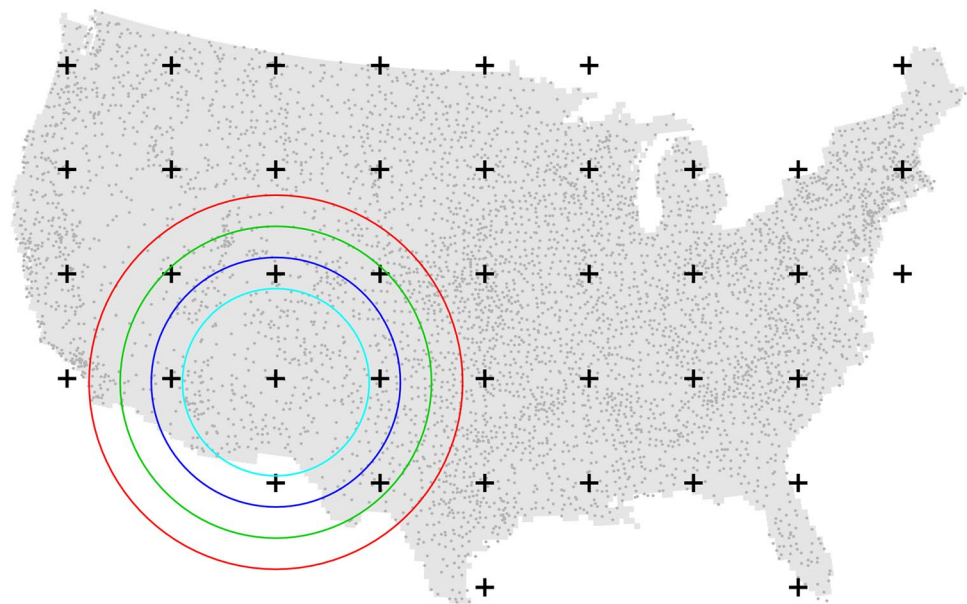
Furthermore, specific to this example, we must also select

4. The local correlation model
5. An elevation-based covariate for the mean structure

We must make these choices separately for each climatological coefficient  $\{\mu_0, \mu_1, \log \sigma, \xi\}$  in each season {DJF, MAM, JJA, SON}. *A priori*, we consider the mixture component locations fixed, using the  $K = 41$  evenly spaced locations over CONUS shown with the “+” symbol in Fig. 11, and we also fix the bandwidth parameter to  $h = 3$  units, so that the weights at the mixture components themselves, i.e.,  $w_k(\mathbf{b}_k)$ , are approximately 0.95 (on average) and the weights for neighboring mixture component locations are approximately 0.0125 (on average). We would like the data to indicate “best” choices for the other selections, as we do not have strong *a priori* expectations about the most appropriate fit radius, local correlation model, and mean covariate. As such, we use fivefold (out-of-sample) cross-validation and the continuous rank probability score (CRPS) to select the best model, considering all possible combinations of the following:

- Fit radius: intuitively, a very large fit radius implies that the spatial mean and covariance parameters are very

**Fig. 11** The  $K = 41$  mixture components used for the non-stationary covariance function (black “+” symbols), and the four fit radii considered (red = Radius 1, green = Radius 2, blue = Radius 3, and light blue = Radius 4). The GHCN stations used in the analysis are shown with the dark gray dots in the background



smooth across the spatial domain, while a very small fit radius implies heterogeneity in the parameters. As such, we consider four different fit radii  $r_1 = 9$ ,  $r_2 = 7.5$ ,  $r_3 = 6$ , and  $r_4 = 4.5$ , as these cover a reasonable range of values for defining the local models. Intuitively, these fit radii indicate varying degrees of second-order nonstationarity, with  $r_4$  yielding the strongest nonstationarity and  $r_1$  yielding the weakest nonstationarity. To compare the fit of a stationary Gaussian process, we also fit a globally stationary model, where all mean and covariance parameters are constant over CONUS (we denote this model as “radius 0” or  $r_0$ ).

- Local correlation: a popular choice in spatial statistics for modeling environmental processes is the Matérn correlation function (see, e.g., Stein 1999), which includes a smoothness parameter  $\kappa$  that specifically controls the roughness of the fitted surface. We consider two values of  $\kappa$ : 0.5, which corresponds to the exponential correlation and non-differentiable surfaces, and 2.5, which gives surfaces that are twice-differentiable.
- Mean covariate: given the well-established relationships between extreme precipitation and orography, we consider using both elevation and the log of elevation in the Gaussian process mean.

Several other items should be noted: first, we are interested in characterizing large-scale features in the location parameter time trend  $\mu_1(\mathbf{s})$  and shape parameter  $\xi(\mathbf{s})$ , as we expect these features to vary less over space than the location intercept and scale parameters. As a result, for  $\mu_1(\mathbf{s})$  and  $\xi(\mathbf{s})$  we estimate a locally isotropic correlation (i.e., circular spatial correlation contours), while for  $\mu_0(\mathbf{s})$  and  $\log \sigma(\mathbf{s})$  we estimate a locally anisotropic correlation (i.e., elliptical spatial

**Table 1** Best spatial models for each season, in terms of the continuous rank probability score for out-of-sample cross validation

Season	Coefficient	Radius	Mean covariate	Smoothness
DJF	$\mu_0$	4	Elevation	0.5
DJF	$\mu_1$	3	Elevation	2.5 (fixed)
DJF	$\log \sigma$	4	Elevation	0.5
DJF	$\xi$	1	Elevation	0.5
MAM	$\mu_0$	4	Elevation	0.5
MAM	$\mu_1$	4	Elevation	2.5 (fixed)
MAM	$\log \sigma$	3	Elevation	0.5
MAM	$\xi$	3	Log elevation	0.5
JJA	$\mu_0$	3	Elevation	0.5
JJA	$\mu_1$	2	Elevation	2.5 (fixed)
JJA	$\log \sigma$	4	Elevation	0.5
JJA	$\xi$	3	Elevation	0.5
SON	$\mu_0$	4	Elevation	0.5
SON	$\mu_1$	3	Log elevation	2.5 (fixed)
SON	$\log \sigma$	2	Elevation	0.5
SON	$\xi$	1	Log elevation	0.5

correlation contours). Second, given that the true time trend  $\mu_1(\mathbf{s})$  is likely to have only large scale features, we only consider the  $\kappa = 2.5$  model for  $\mu_1(\mathbf{s})$ .

The best model (in terms of CRPS) for each coefficient and each season is shown in Table 1. While we wish to allow the coefficient in each season to have a unique degree of nonstationarity (i.e., fit radius), we opt to use a fixed mean covariate (elevation) and smoothness ( $\kappa = 0.5$ ) across all seasons and coefficients, as these two models are most often chosen as “best” (except for the smoothness in  $\mu_1(\mathbf{s})$ , which is fixed at

2.5). Note that across all seasons and parameters, the nonstationary Gaussian process models (Radii 1–4) are always preferred to the stationary Gaussian process model (Radius 0).

### Appendix 3: Computational details

The computational resources required for the analysis in this paper are extensive in light of the bootstrap fitting, even given the computational efficiency of the local likelihood method for estimating a nonstationary covariance function. The analysis was feasible in our case thanks to having access to a large supercomputer through the National Energy Research Scientific Computing Center (NERSC). Specifically, we utilized NERSC's newest supercomputer Cori, a Cray XC40 machine comprised of 2388 Intel Xeon "Haswell" processor nodes (there are two other node partitions in Cori that we did not use). Each Haswell node consists of 32 cores, with a total of 128 GB DDR4 2133 MHz memory per node.

There are three primary tasks that are time consuming for the analysis. The approximate number of required hours (both wall clock and compute time) and node configuration for each task are given in the following subsections.

#### Appendix 3.1: Station-specific extreme value analysis

The four-parameter GEV analysis for each station using `climextRemes` (Paciorek 2016)—i.e., estimating  $\mu_0$ ,  $\mu_1$ ,  $\sigma$ , and  $\xi$ —is actually quite fast; however, recall that there are 5202 stations and we must conduct the analysis at each station for four seasons and  $B = 250$  bootstrap samples (plus one more for the full data fits). On the Cori cores, estimating the GEV coefficients for all 5202 stations takes approximately 15 min. Overall, the configuration for this step is as follows:

- One "job" is defined as estimating the GEV coefficients for all 5202 stations
- Configuration: the 4 seasons  $\times$  (250 bootstrap + 1 full data) = 1004 jobs are spread out over 16 nodes (512 cores)
- Compute hours: 1004 jobs  $\times$  15 min each  $\approx$  250 h
- Wall clock (elapsed) hours: approximately 30 min total

#### Appendix 3.2: Cross-validation for spatial model selection

Recall from Appendix 2 that we must fit a set of spatial models to each of (4 seasons)  $\times$  (4 GEV coefficients)  $\times$  (2 mean covariates)  $\times$  (2 smoothness values)  $\times$  (5 holdout sets)

= 320 models. The benefit of the nonstationary models (radii 1–4) is that we can split each of these fits into  $K = 41$  local fits, one for each mixture component. The stationary model (radius 0), unfortunately, cannot be divided into any smaller tasks.

Given that the stationary model involves fitting a spatial model to over 5000 stations and the nonstationary model involves fitting local models to sample sizes ranging between 50 and 2000, for computation we split these tasks into two computing jobs:

##### *Stationary model fitting*

- One "job" is defined as fitting a stationary model to all 5202 stations
- Configuration: the 320 jobs are spread out over 5 nodes (160 cores)
- Compute hours: 160 cores  $\times$  10 hours  $\approx$  1600 hours total
- Wall clock (elapsed) hours: approximately 10 h.

##### *Nonstationary model fitting*

- One "job" is defined as fitting a locally stationary model
- Configuration: the 320 combinations  $\times$  4 radii  $\times$  41 local fits = 52480 jobs spread out over 32 nodes (1024 cores)
- Compute hours: 1024 cores  $\times$  5 h  $\approx$  5120 h total
- Wall clock (elapsed) hours: approximately 5 h.

Once the GP parameters have been estimated, it remains to conduct prediction for the held-out data, but this is a relatively low-cost operation computationally, and was conducted on a personal laptop.

#### Appendix 3.3: Spatial model fitting for the full and bootstrap data sets

The cross-validation is conducted to select a best model for each coefficient/season, shown in Table 1. The next task is to fit the best model to all 5202 stations, for both the full data estimates and the bootstrapped estimates. Given that the best model across all coefficients/seasons is one of the nonstationary models, we can split things up as follows:

- One "job" is defined as fitting a locally stationary model
- Configuration: (250 bootstrap + 1 full data)  $\times$  4 coefficients  $\times$  4 seasons  $\times$  41 local fits = 164,656 jobs spread out over 64 nodes (2048 cores)
- Compute hours: 2048 cores  $\times$  6 h  $\approx$  12288 h total  $\approx$  1.4 years
- Wall clock (elapsed) hours: approximately 6 h



### Appendix 3.4: Kriging for the full and bootstrap data sets

After estimating the local GP parameters, we need to obtain best estimates for the coefficients (across all seasons and bootstrap data sets) at both the GHCN station locations (5202 locations) and the  $0.25^\circ$  grid over CONUS (13,073 locations). This step is non-trivial, as it involves computing a  $5202 \times 5202$  covariance matrix and a  $13073 \times 5202$  cross-covariance matrix, as well as computing the Cholesky decomposition of a  $5202 \times 5202$  matrix and the subsequent large matrix operations (see Eq. 11 in the main text). Given the large memory requirements of these calculations (recall that for a data set of size  $n$ , a single evaluation of the Gaussian likelihood requires  $\mathcal{O}(n^2)$  storage and  $\mathcal{O}(n^3)$  calculations), the following calculations were done using an alternative computing cluster (2x Eight Core Xeon E5-2680 2.7 GHz machine with 384 GB total memory):

- One “job” is defined as calculating the kriging predictor for 5202 + 13,073 locations
- Configuration: (250 bootstrap + 1 full data)  $\times$  4 coefficients  $\times$  4 seasons = 4004 jobs spread out over 16 cores
- Compute hours: 4004 jobs  $\times$  6.2 min each  $\approx$  410 h
- Wall clock (elapsed) hours: approximately 17 h.

### Appendix 3.5: Total time

Aggregating all of the above:

- Compute hours:  $250 + 1600 + 6144 + 12288 + 410 = 20692$  hours
- Wall clock (elapsed) hours:  $0.5 + 10 + 6 + 6 + 17 = 39.5$  hours.

### References

- Barnard J, McCulloch R, Meng XL (2000) Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat Sin* 10:1281–1311
- Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London, United Kingdom
- Cooley D, Nychka D, Naveau P (2007) Bayesian spatial modeling of extreme precipitation return levels. *J Am Stat Assoc* 102(479):824–840. <https://doi.org/10.1198/016214506000000780>
- Craigmile P, Guttorp P (2013) Can a regional climate model reproduce observed extreme temperatures? *Statistica* 73(1):103–122. <https://doi.org/10.6092/issn.1973-2201/3988>
- Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int J Climatol* 28(15):2031–2064
- Daniels MJ, Kass RE (2001) Shrinkage estimators for covariance matrices. *Biometrics* 57(4):1173–1184
- Davison AC, Padoan SA, Ribatet M (2012) Statistical modeling of spatial extremes. *Stat Sci* 27(2):161–186. <https://doi.org/10.1214/11-STS376>
- Demarta S, McNeil AJ (2005) The  $t$  copula and related copulas. *Int Stat Rev* 73:111–129
- Diaconescu EP, Gachon P, Laprise R (2015) On the remapping procedure of daily precipitation statistics and indices used in regional climate model evaluation. *J Hydrometeorol* 16(6):2301–2310
- Easterling DR, Arnold J, Knutson T, Kunkel K, LeGrande A, Leung LR, Vose R, Waliser D, Wehner M (2017) Precipitation change in the United States. In: Wuebbles DJ, Fahey DW, Hibbard KA, Dokken DJ, Stewart BC, Maycock TK (eds) Climate science special report: fourth national climate assessment, vol I. U.S. Global Change Research Program, Washington, DC, USA, pp 207–230. <https://doi.org/10.7930/JOH993CC>
- Fuentes M, Henry J, Reich B (2013) Nonparametric spatial models for extremes: application to extreme temperature data. *Extremes* 16(1):75–101. <https://doi.org/10.1007/s10687-012-0154-1>
- Gelfand AE, Kottas A (2002) A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *J Comput Graph Stat* 11(2):289–305. <https://doi.org/10.1198/106186002760180518>
- Gervais M, Tremblay LB, Gyakum JR, Atallah E (2014) Representing extremes in a daily gridded precipitation analysis over the united states: impacts of station density, resolution, and gridding methods. *J Clim* 27(14):5201–5218
- Haan LD (1984) A spectral representation for max-stable processes. *Ann Probab* 12(4):1194–1204. <https://doi.org/10.1214/aop/1176993148>
- Hannart A, Naveau P (2014) Estimating high dimensional covariance matrices: a new look at the Gaussian conjugate framework. *J Multivar Anal* 131:149–162
- Holland DM, De OV, Cox LH, Smith RL (2000) Estimation of regional trends in sulfur dioxide over the eastern United States. *Environmetrics* 11(4):373–393
- Hüsler J, Reiss RD (1989) Maxima of normal random vectors: between independence and complete dependence. *Stat Probab Lett* 7:283–286
- Kabluchko Z, Schlather M, de Haan L (2009) Stationary max-stable fields associated to negative definite functions. *Ann Probab* 37(5):2042–2065. <https://doi.org/10.1214/09-AOP455>
- King AD, Alexander LV, Donat MG (2013) The efficacy of using gridded data to examine extreme rainfall characteristics: a case study for Australia. *Int J Climatol* 33(10):2376–2387
- Kottas A, Wang Z, Rodríguez A (2012) Spatial modeling for risk assessment of extreme values from environmental time series: a Bayesian nonparametric approach. *Environmetrics* 23(8):649–662. <https://doi.org/10.1002/env.2177>
- Krupskii P, Huser R, Genton MG (2018) Factor copula models for replicated spatial data. *J Am Stat Assoc* 113(521):467–479
- Kunkel KE, Karl TR, Brooks H, Kossin J, Lawrimore JH, Arndt D, Bosart L, Changnon D, Cutter SL, Doesken N (2013) Monitoring and understanding trends in extreme storms: state of knowledge. *Bull Am Meteorol Soc* 94(4):499–514
- Livneh B, Rosenberg EA, Lin C, Nijssen B, Mishra V, Andreadis KM, Maurer EP, Lettenmaier DP (2014) A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: update and extensions. *J Clim* 27(1):478
- Lovejoy S, Schertzer D, Allaire V (2008) The remarkable wide range spatial scaling of TRMM precipitation. *Atmos Res* 90(1):10–32. <https://doi.org/10.1016/j.atmosres.2008.02.016>. <http://linkinghub.elsevier.com/retrieve/pii/S0169809508000562>
- Mannshardt E, Craigmile PF, Tingley MP (2013) Statistical modeling of extreme value behavior in North American tree-ring density series. *Clim Change* 117(4):843–858. <https://doi.org/10.1007/s10584-012-0575-5>

- Maskey ML, Puente CE, Sivakumar B, Cortis A (2016) Encoding daily rainfall records via adaptations of the fractal-multifractal method. *Stoch Environ Res Risk Assess* 30(7):1917–1931. <https://doi.org/10.1007/s00477-015-1201-7>
- Menne MJ, Durre I, Korzeniewski B, McNeal S, Thomas K, Yin X, Anthony S, Ray R, Vose RS, Egleason B, Houston TG (2012a) Global Historical Climatology Network—Daily (GHCN-Daily), version 3, 10.7289/V5D21VHZ, nOAA National Climatic Data Center (accessed 19 March 2018)
- Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG (2012b) An overview of the global historical climatology network-daily database. *J Atmos Ocean Technol* 29(7):897–910
- Morris SA, Reich BJ, Thibaud E, Cooley D (2017) A space-time skew- $t$  model for threshold exceedances. *Biometrics* 73(3):749–758
- Norbiato D, Borga M, Sangati M, Zanon F (2007) Regional frequency analysis of extreme precipitation in the eastern Italian Alps and the August 29, 2003 flash flood. *J Hydrol* 345(3–4):149–166
- Paciorek C (2016) *climextRemes*: Tools for analyzing climate extremes. R package version 0.1.2. <https://CRAN.R-project.org/package=climextRemes>. Accessed 15 Oct 2016
- Paciorek CJ, Schervish MJ (2006) Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* 17(5):483–506
- Reich BJ, Shaby BA (2012) A hierarchical max-stable spatial model for extreme precipitation. *Ann Appl Stat* 6(4):1430–1451. <https://doi.org/10.1214/12-AOAS591>
- Risser MD (2016) Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. arXiv preprint [arXiv:161002447](https://arxiv.org/abs/161002447)
- Risser MD, Calder CA (2015) Regression-based covariance functions for nonstationary spatial modeling. *Environmetrics* 26(4):284–297
- Risser M, Calder C (2017) Local likelihood estimation for covariance functions with spatially-varying parameters: the convoSPAT package for R. *J Stat Softw* 81(14):1–32. <https://doi.org/10.18637/jss.v081.i14>. <https://www.jstatsoft.org/v081/i14>
- Risser MD, Wehner MF (2017) Attributable human-induced changes in the likelihood and magnitude of the observed extreme precipitation during Hurricane Harvey. *Geophys Res Lett*. <https://doi.org/10.1002/2017GL075888>
- Risser MD, Paciorek CJ, Wehner MF, O'Brien TA, Collins WD (2019) A probabilistic gridded product for daily precipitation extremes over the United States. <https://doi.org/10.7910/DVN/LULNUQ>
- Russell BT, Cooley DS, Porter WC, Heald CL (2016) Modeling the spatial behavior of the meteorological drivers' effects on extreme ozone. *Environmetrics* 27(6):334–344
- Sang H, Gelfand AE (2010) Continuous spatial process models for spatial extreme values. *J Agric Biol Environ Stat* 15(1):49–65. <https://doi.org/10.1007/s13253-009-0010-1>
- Saunders K, Stephenson AG, Taylor PG, Karoly D (2017) The spatial distribution of rainfall extremes and the influence of El Niño Southern Oscillation. *Weather Clim Extremes* 18:17–28
- Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol*. <https://doi.org/10.2202/1544-6115.1175>
- Schlather M (2002) Models for stationary max-stable random fields. *Extremes* 5(1):33–44. <https://doi.org/10.1023/A:1020977924878>
- Serinaldi F, Kilsby CG (2014) Simulating daily rainfall fields over large areas for collective risk estimation. *J Hydrol* 512:285–302
- Shaby BA, Reich BJ (2012) Bayesian spatial extreme value analysis to assess the changing risk of concurrent high temperatures across large portions of European cropland. *Environmetrics* 23(8):638–648. <https://doi.org/10.1002/env.2178>
- Smith RL (1990) Max-stable processes and spatial extremes. Unpublished manuscript
- Stein ML (1999) *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, New York
- Sylla M, Giorgi F, Coppola E, Mariotti L (2013) Uncertainties in daily rainfall over Africa: assessment of gridded observation products and evaluation of a regional climate model simulation. *Int J Climatol* 33(7):1805–1817
- Timmermans B, Wehner M, Cooley D, O'Brien T, Krishnan H (2019) An evaluation of the consistency of extremes in gridded precipitation data sets. *Clim Dyn*. <https://doi.org/10.1007/s00382-018-4537-0>
- Tye MR, Cooley D (2015) A spatial model to examine rainfall extremes in Colorado's Front Range. *J Hydrol* 530:15–23
- Volosciuk C, Maraun D, Semenov VA, Park W (2015) Extreme precipitation in an atmosphere general circulation model: impact of horizontal and vertical model resolutions. *J Clim* 28(3):1184–1205. <https://doi.org/10.1175/JCLI-D-14-00337.1>
- Wallis J, Schaefer M, Barker B, Taylor G (2007) Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington State. *Hydrol Earth Syst Sci Discuss* 11(1):415–442
- Wang Z, Rodríguez A, Kottas A (2011) A nonparametric mixture modeling framework for extreme value analysis. Technical Report UCSC-SOE-11-26, University of California, Santa Cruz. <http://www.soe.ucsc.edu/research/technical-reports/ucsc-soe-11-26>. Accessed 1 Nov 2016
- Wehner MF (2013) Very extreme seasonal precipitation in the NARC-CAP ensemble: model performance and projections. *Clim Dyn* 40(1–2):59–80
- Wehner MF, Reed KA, Li F, Bacmeister J, Chen CT, Paciorek C, Gleckler PJ, Sperber KR, Collins WD, Gettelman A (2014) The effect of horizontal resolution on simulation quality in the Community Atmospheric Model, CAM5.1. *J Adv Model Earth Syst* 6(4):980–997
- Westra S, Alexander LV, Zwiers FW (2013) Global increasing trends in annual maximum daily precipitation. *J Clim* 26(11):3904–3918
- Yang T, Shao Q, Hao ZC, Chen X, Zhang Z, Xu CY, Sun L (2010) Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. *J Hydrol* 380(3–4):386–405

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.