Spatial and Spatio-Temporal Analysis of Precipitation Data from South Carolina

Haigang Liu, David B. Hitchcock, and S. Zahra Samadi

1 Introduction

Spatial and spatio-temporal data are everywhere: we encounter them on TV, in newspapers, on computer screens, on tablets, and on plain paper maps. As a result, researchers in diverse areas are increasingly faced with the task of modeling geographically referenced and temporally correlated data.

The geostatistical analysis of spatial data involves point-referenced data, where $Y(\mathbf{s})$ is a random vector at a location $\mathbf{s} \in \mathcal{R}^r$, where $\mathbf{s}$ varies continuously over D, a fixed subset of $\mathcal{R}^r$ that contains an r-dimensional rectangle of positive volume (Banerjee et al. 2014). The sample points are measurements of some phenomenon such as precipitation measurements from meteorological stations or elevation heights. The geostatistical analysis models a surface using the values from the measured locations to predict values for each location in the landscape.

Spatial statistics methods have been frequently used in applied statistics as well as water resources engineering. The work of Thiessen (1911) was the first attempt in using interpolation methods in hydrology. Sharon (1972) used an average of the observations from a number of rain gages to obtain estimates of the areal rainfall. Soon after, Benzécri (1973), Delfiner and Delhomme (1975), and Delhomme (1978) applied the various geostatistical methods such as variograms and kriging methods in modeling rainfall. The work of Troutman (1983), Tabios

H. Liu · D. B. Hitchcock (✉)
Department of Statistics, University of South Carolina, Columbia, SC, USA
e-mail: haigang@email.sc.edu; hitchcock@stat.sc.edu

S. Z. Samadi
Department of Civil and Environmental Engineering, University of South Carolina, Columbia, SC, USA
e-mail: samadi@cec.sc.edu

N. Diawara (ed.), *Modern Statistical Methods for Spatial and Multivariate Data*,
STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health,
https://doi.org/10.1007/978-3-030-11431-2_2

and Salas (1985), Georgakakos and Kavvas (1987), Isaaks and Srivastava (1989), Kumar and Foufoula-Georgiou (1994), Deidda (2000), Ferraris et al. (2003), Ciach and Krajewski (2006), Berne et al. (2009), Ly et al. (2011), and Dumitrescu et al. (2016) further advanced the application of geostatistical methods in rainfall prediction. The theoretical basis of the geostatistical approach was strengthened using Bayesian inference via the Markov Chain Monte Carlo (MCMC) algorithm introduced by Metropolis et al. (1953). MCMC was subsequently adapted by Hastings (1970) for statistical problems and further applied by Diggle et al. (1998) in geostatistical studies. Recent developments in MCMC computing now allow fully Bayesian analyses of sophisticated multilevel models for complex geographically referenced data. This approach also offers full inference for non-Gaussian spatial data, multivariate spatial data, spatio-temporal data, and solutions to problems such as geographic and temporal misalignment of spatial data layers (Banerjee et al. 2014).

The data we are studying are monthly rainfall data measured across the state of South Carolina from the start of 2011 to the end of 2015. The precipitation record in 2015 is of particular interest because a storm in October 2015 in North America triggered a high precipitation event, which caused historic flash flooding across North and South Carolina. Rainfall across parts of South Carolina reached 500-year-event levels (NBC News, October 4, 2015). Accumulations reached 24.23 in. near Boone Hall (Mount Pleasant, Charleston County) by 11:00 a.m. Eastern Time on October 4, 2015. Charleston International Airport saw a record 24-h rainfall of 11.5 in. (290 mm) on October 3 (Santorelli, October 4, 2015). Some areas experienced more than 20 in. of rainfall over the 5-day period. Many locations recorded rainfall rates of 2 in. per hour (National Oceanic and Atmospheric Administration (NOAA), U.S. Department of Commerce, 2015).

The extraordinary rainfall event was generated by the movement of very moist air over a stalled frontal boundary near the coast. The clockwise circulation around a stalled upper level low over southern Georgia directed a narrow plume of tropical moisture northward and then westward across the Carolinas over the course of 4 days. A low pressure system off the US southeast coast, as well as tropical moisture related to Hurricane Joaquin (a category 4 hurricane) was the underlying meteorological cause of the record rainfall over South Carolina during October 1–5, 2015 (NOAA, U.S. Department of Commerce 2015).

Flooding from this event resulted in 19 fatalities, according to the South Carolina Emergency Management Department, and South Carolina state officials said damage losses were 1.492 billion dollars (NOAA, U.S. Department of Commerce 2015). The heavy rainfall and floods, combined with aging and inadequate drainage infrastructure, resulted in the failure of many dams and flooding of many roads, bridges, and conveyance facilities, thereby causing extremely dangerous and life-threatening situations.

The chapter is arranged as follows: in Sect. 2, we give an overview of our precipitation data, in conjunction with some other variables, e.g., sea surface temperature, which might help explain the behavior of the precipitation. In Sect. 3, we introduce the kriging method to analyze the precipitation using a pure spatial

analysis. In Sect. 4, some methods in seasonal trend removal are discussed. In Sect. 5, the Gaussian process is introduced to build a spatio-temporal model.

2 Data Description

2.1 Overview

The original data used in this research are the daily precipitation records in South Carolina from National Oceanic and Atmosphere Administration (NOAA) between 2011 and 2015. The original data files include daily precipitation, maximum temperature, and minimum temperature, along with the latitude, longitude, and elevation of each observation's location.

In addition, to investigate the effect of El Niño-Southern Oscillation (ENSO) activity on precipitation, we have calculated an index based on the monthly sea surface temperature (SST). The derivation of our index is given in Sect. 2.3.

2.2 Data Preprocessing

We collected 281 unique meteorological locations in South Carolina with varying completeness of data. For instance, if we look at the most recent 5 years (2011–2015), 31 locations do not have any record of precipitation while 65 locations have a complete record. The other 185 locations contain missing data ranging from 30% to less than 5% of the total data set size.

In Fig. 1, we plot all the meteorological locations with an available precipitation record on October 3, 2015, when the storm struck South Carolina. Note that smoothing is necessary since most of observations are clustered in several regions. See Bivand et al. (2008) for more information about the `sp` package, which provides a comprehensive solution for spatial data visualization.

We aggregate the daily records into monthly variables. The monthly maximum of precipitation is calculated since we are interested in capturing the extreme rainfall behavior which might lead to flooding subsequently. The monthly midrange temperature, which reflects the general warmth of that month, is computed by averaging the highest and the lowest daily temperature for that month.

To incorporate more temperature information, we find the range of daily maxima over a month. We similarly obtain the range of the daily minima. Lastly, for each location, we also find an overall range, the difference of the maximum and minimum temperature of that month.

In the data set, several variables, e.g., precipitation, elevation, and temperature have missing values. We replace each missing observation with the weighted average of its neighbors. The weights are determined by the distance between

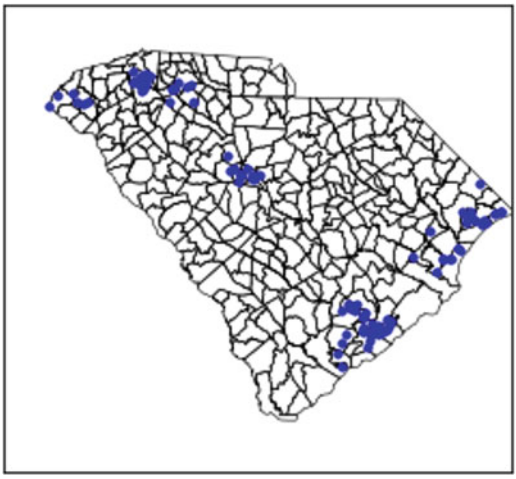

Fig. 1 The meteorological locations with available record of precipitation on October 3, 2015

locations. In other words, if we denote the missing value at $\mathbf{s}^*$ by $Y(\mathbf{s}^*)$, then $\sum_{i=1}^{n} w(\mathbf{s}_i)Y(\mathbf{s}_i)$ can be used as the imputed value, where

$$w(\mathbf{s}_i) = K\left(\frac{||\mathbf{s}^* - \mathbf{s}_i||}{h}\right) / \sum_{i=1}^{n} K\left(\frac{||\mathbf{s}^* - \mathbf{s}_i||}{h}\right). \quad (1)$$

Note that $||\mathbf{s}_i - \mathbf{s}^*||$ refers to the haversine distance rather than the Euclidean distance. We impute missing data based on neighboring observations because doing so takes the spatial correlation into consideration.

2.3 A Sea Surface Temperature (SST)-Related Variable

El Niño-Southern Oscillation (ENSO) is an irregular variation in winds and sea surface temperature (SST) over the tropical eastern Pacific Ocean, affecting much of the tropic and subtropics. Like other climate indices, ENSO occurs irregularly and is associated with changing in physical pattern of temperature and precipitation. Figure 2 gives the plot of sea surface temperature for ocean locations off the coast of South Carolina in June 2015. In this figure, dark colors correspond to cooler sea temperature values. Scientists believe that the ENSO has a significant influence on precipitation and hence controls flood magnitude and frequency. We thus include an SST-based index as a proxy for the ENSO activity. Since our rainfall data are observed for inland locations, we must define our index related to SST for such inland locations, rather than for off-shore locations where sea temperature is actually measured.

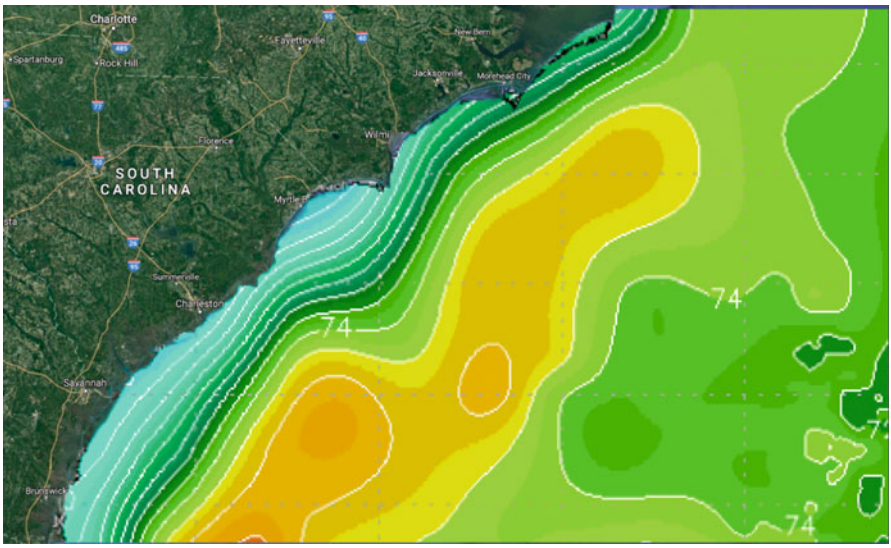

Fig. 2 The sea surface temperature near South Carolina

For any inland location $\mathbf{s}_i$ at a given month, we build an index based on the SST values of the nearest n adjacent ocean observation points $\{\mathbf{z}_j\}$, where $j = 1, \ldots, n$. Denote this SST-based index as $W(\mathbf{s}_i)$ for the ith inland location. It follows that

$$W(\mathbf{s}_i) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{w_j}{\sum_{l=1}^{n} w_l} \right) \mathrm{SST}(\mathbf{z}_j), \tag{2}$$

where the weight w_j can be determined by the kernel function $K(||\mathbf{s}_i - \mathbf{z}_j||)$ for $j = 1, \ldots, n$, which is symmetric around 0. We use the standard normal density as the kernel function. The kernel function includes a bandwidth h, thus making $w_j = \frac{1}{h} K(\frac{||\mathbf{s}_i - \mathbf{z}_j||}{h})$. The bandwidth parameter h is set to 0.25 times the range of all of the distances.

Additionally, we simplify the calculation by considering only locations within a certain threshold. Figure 3 gives a demonstration to calculate the SST-related index for Columbia, South Carolina. We first determine the sea temperature records to be included based on a 300-mile threshold. For the included measurements, we find their weights by calculating their distance to Columbia, and derive the SST-related index based on (2). Note that the closer a location is to the coast, the more sea surface temperature records are used to derive an SST-related index for that location.

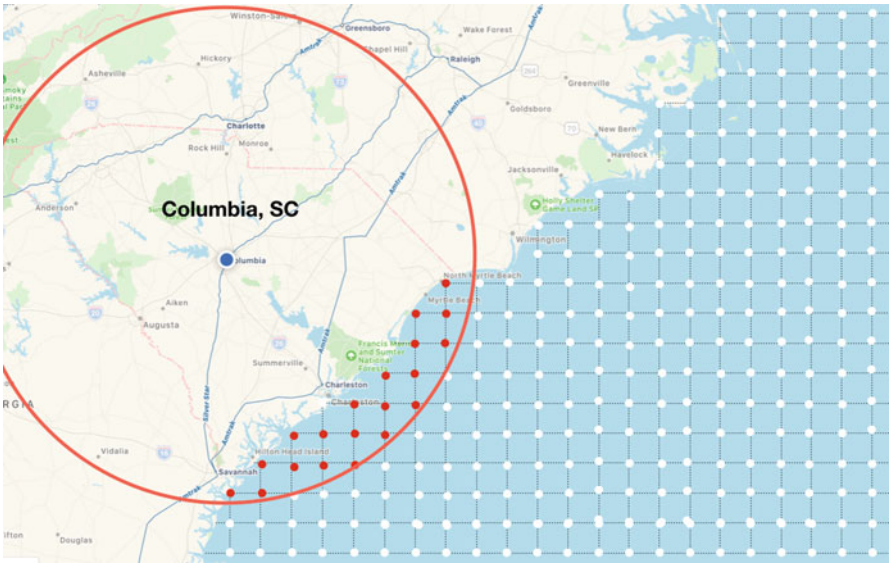

Fig. 3 A demonstration of the calculation of the SST-related variable. The red points are the observations that are included in the calculation

3 Precipitation Modeling: A Spatial Perspective

In this section, we use a spatial model for the rainfall data without considering the temporal aspect. Since geostatistical data feature a strong correlation between adjacent locations, we start by modeling the covariance structure with a variogram, and then we propose two methods of predicting the rainfall for new location.

3.1 Describing the Spatial Structure: Variogram

We assume that our spatial process has a mean, $\mu(\mathbf{s}) = E(Y(\mathbf{s}))$, and that the variance of $Y(\mathbf{s})$ exists for all $\mathbf{s} \in D$. The process $Y(\mathbf{s})$ is said to be Gaussian if, for any $n \geq 1$ and any set of sites $\{\mathbf{s}_1 \ldots, \mathbf{s}_n\}$, $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$ has a multivariate normal distribution. Moreover, the process is *intrinsic stationary* if, for any given $n \geq 1$, any set of n sites $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ and any $h \in \mathcal{R}^r$, we have $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})] = 0$, and $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})]^2 = \text{Var}(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) = 2\gamma(\mathbf{h})$ (Banerjee et al. 2014).

In other words, $E[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})]^2$ only depends on $\mathbf{h}$, and not the particular choice of $\mathbf{s}$. The function $2\gamma(\mathbf{h})$ is then called the *variogram*, and $\gamma(\mathbf{h})$ is called the *semivariogram*. Another important concept is that of an *isotropic* variogram. If the semivariogram function $\gamma(\mathbf{h})$ depends upon the separation vector only through

its length ||**h**|| (distance between observations), then the variogram is isotropic. Otherwise, it is *anisotropic*. Isotropic variograms are popular because of simplicity, interpretability, and, in particular, because a number of relatively simple parametric forms are available as candidates for the semivariogram, e.g., linear, exponential, Gaussian, or Matérn (or K-Bessel).

A variogram model is chosen by plotting the empirical semivariogram, a simple nonparametric estimate of the semivariogram, and then comparing it to the various theoretical parametric forms (Matheron 1963). For demonstration purposes, we choose the precipitation values of October 13 in 2015, shortly after the flood struck South Carolina. Assuming intrinsic stationarity and isotropy, the Matérn model is used due to its better fit to the empirical semivariogram. The correlation function of this model allows control of spatial association and smoothness. See Fig. 4 for a plot of this fit.

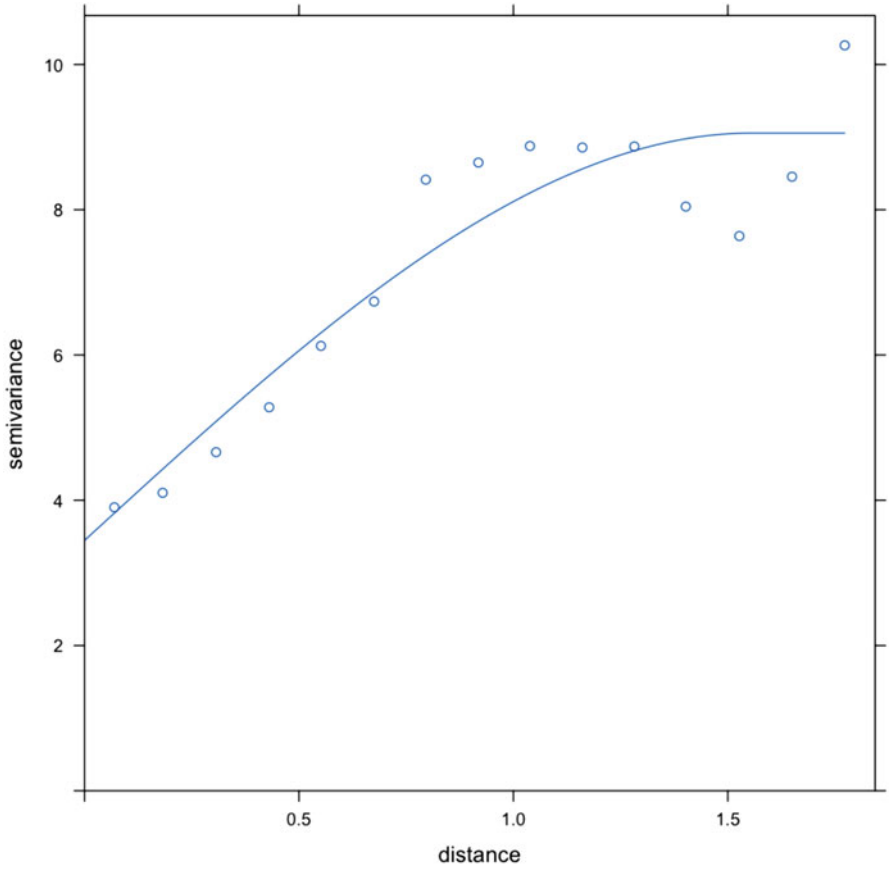

Fig. 4 The empirical and parametric (Matérn) variogram for the precipitation values in October 13, 2015

3.2 Spatial Prediction

Inverse Distance Weighted Interpolation

We use inverse distance weighting (IDW) (Bivand et al. 2008) to compute a spatially continuous rainfall estimate as a weighted average for a given location $\mathbf{s}_0$,

$$\hat{Z}(\mathbf{s}_0) = \frac{\sum w(\mathbf{s}_i) Z(\mathbf{s}_i)}{\sum w(\mathbf{s}_i)}, \quad \text{where } w(\mathbf{s}_i) = ||\mathbf{s}_i - \mathbf{s}_0||^{-p}.$$

In other words, the weight of a given observed location is based on its L_p-distance to the interpolation location. If location $\mathbf{s}_0$ happens to have an observation, then the observation itself will be used to avoid the case of infinite weights. The weight assigned to data points will be more influenced by neighboring points when they are more clustered. The best p found by cross validation for the analysis of our data set is approximately 2.5.

Although this method does not incorporate the covariates, it still possesses some desirable features. For instance, we can make a prediction for the rainfall amount at every single location with a latitude and longitude.

Linear Gaussian Process Model (Kriging)

Since our precipitation data in the study are geostatistical data, we may employ a linear Gaussian process model (Cressie 1993). We start by defining the spatial process at location $\mathbf{s} \in \mathcal{R}^d$ as

$$Z(\mathbf{s}) = X(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}), \tag{3}$$

where $\mathbf{X}(\mathbf{s})$ is a set of p covariates associated with each site $\mathbf{s}$, and $\boldsymbol{\beta}$ is a p-dimensional vector of coefficients. Spatial dependence is imposed via the residual terms, i.e., $w(\mathbf{s})$. Specifically, we model $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{R}^d\}$ as a zero mean Gaussian process. In other words, the vector $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$ follows $\mathbf{w}|\Theta \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}(\Theta))$. We assume $\boldsymbol{\Sigma}$ to be a symmetric and positive definite matrix in order to end up with a sensible distribution. To ensure these conditions, $\boldsymbol{\Sigma}(\Theta)$ can be treated as a function of Θ with certain constraints, which are tantamount to specifying a variogram model.

Among several variogram structures, e.g., spherical, Gaussian, exponential, etc. we choose the exponential covariance with parameters $\Theta = (\psi, \kappa, \phi)$, where $\psi, \kappa, \phi > 0$. The exponential covariance $\boldsymbol{\Sigma}(\Theta)$ has the form

$$\boldsymbol{\Sigma}(\Theta) = \psi \mathbf{I} + \kappa H(\phi), \quad \text{where} H(\phi) = \exp(-||\mathbf{s}_i - \mathbf{s}_j||)/\phi).$$

Note that $||\mathbf{s}_i - \mathbf{s}_j||$ is the Euclidean distance between location i and j. Another type of distance, *Geodesic*, takes the curvature of the earth's surface into consideration.

We use Euclidean distance since most of our distances are between South Carolina counties and the effects of curvature are thus negligible.

The exponential model enjoys a simple interpretation. The "nugget" in a variogram graph is represented by ψ in this model, and this nugget is also the variance of the non-spatial error. Moreover, κ and ϕ dictate the scale and range of the spatial dependence, respectively. Also note that the exponential model assumes the covariance and hence dependence between two locations decreases as distance between locations increases, which is sensible for the study of rainfall behavior.

Letting $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$, we estimate the multivariate normal distribution for $\mathbf{Z}$ after parameter estimation. To find the unknown parameters Θ and $\boldsymbol{\beta}$, we use Bayesian methods implemented by the `spTimer` package in R (Bakar and Sahu 2015), which requires users to provide sensible prior information based on sample variogram graphs. Note that this model fitting process will collapse if we start with initial values far from the true value.

Monte Carlo Simulation for Kriging

Predictions of the process, $\mathbf{Z}^* = (Z(\mathbf{s}_1^*), \ldots, Z(\mathbf{s}_m^*))^T$, where $\mathbf{s}_i^*$ is the ith new location, can be obtained via the posterior predictive distribution

$$\pi(\mathbf{Z}^*|\mathbf{Z}) = \int \pi(\mathbf{Z}^*|\mathbf{Z}, \Theta, \boldsymbol{\beta})\pi(\Theta, \boldsymbol{\beta}|\mathbf{Z})d\Theta d\boldsymbol{\beta},$$

by sampling from the posterior predictive distribution in two steps:

- Step 1: Simulate $\Theta', \boldsymbol{\beta}' \sim \pi(\Theta, \boldsymbol{\beta}|\mathbf{Z})$ by the Metropolis–Hastings algorithm.
- Step 2: Simulate $\mathbf{Z}^*|\Theta', \boldsymbol{\beta}, \mathbf{Z}$ from a multivariate normal density.

For step 1, it suffices to find the posterior distribution $\pi(\Theta, \boldsymbol{\beta}|\mathbf{Z})$ based on (1) and (2). The posterior distribution has low dimension as long as we do not have many covariates. The major challenge is that since covariance parameters might be highly correlated, one must expect autocorrelation issues in the sampler, which can be alleviated by a block updating scheme, a scheme that generates multiple covariance parameters in a single Metropolis–Hastings step.

For step 2, the joint distribution of $\mathbf{Z}$ and $\mathbf{Z}^*$ is given by

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{Z}^* \end{bmatrix} | \Theta, \boldsymbol{\beta} \sim N\left(\begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

based on which one can find the conditional distribution of $\mathbf{Z}^*|\Theta', \boldsymbol{\beta}, \mathbf{Z}$. According to Anderson (2003), it follows that

$$E(\mathbf{Z}^*|\Theta', \boldsymbol{\beta}, \mathbf{Z}) = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_1),$$

$$\text{Var}(\mathbf{Z}^*|\Theta', \boldsymbol{\beta}, \mathbf{Z}) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Hence, one can obtain simulated observations that follow a given covariance structure by iterating between step 1 and step 2. Bivand et al. (2008) suggest the method of sequential simulation: (1) compute the conditional distribution with our given data, (2) draw a value from this conditional distribution, (3) add this value into the data set, and (4) repeat steps (1)–(3).

As $\mathbf{Z}$ becomes a larger matrix as more data are generated, the algorithm becomes more and more expensive. Many strategies are proposed for reducing the considerable computational burden posed by matrix operations, including the use of covariance functions (Hughes and Haran 2013) as well as setting a maximum number of neighbors (Bivand et al. 2008). In our study, we used the maximum number of neighbors with the nearest 40 observations.

We illustrate prediction by modeling rainfall in South Carolina on October 13, 2015 with a kriging model that assumes an exponential spatial covariance structure. Using the Monte Carlo approach described above, we predict by simulating from the posterior predictive distribution. This can be done repeatedly to give a sense of the variability associated with the spatial predictions. Figure 5 demonstrates ten simu-

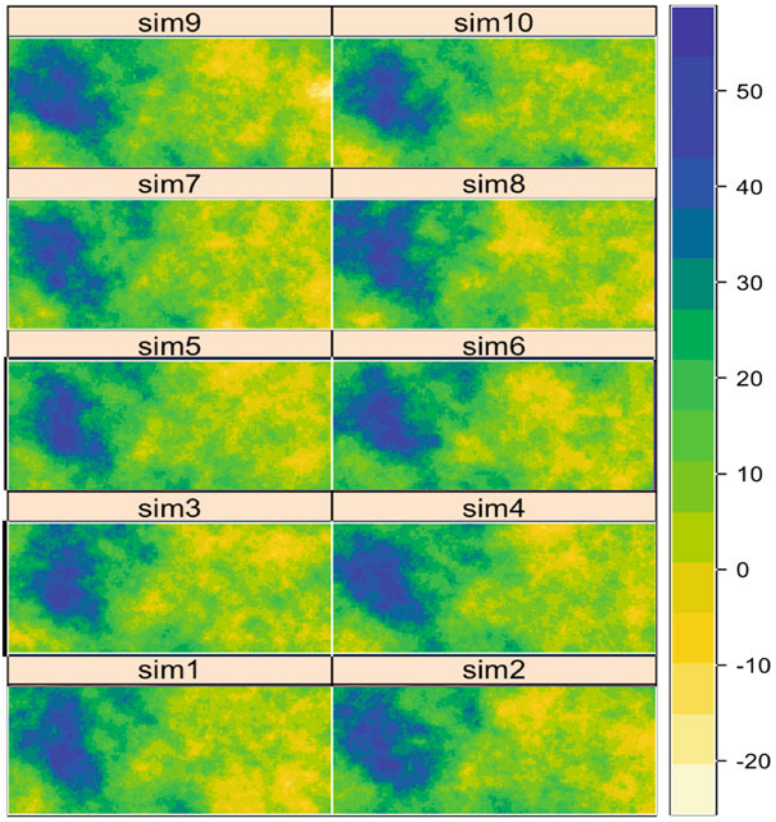

Fig. 5 Ten simulated precipitation heat maps based on kriging. The darker color indicates heavier precipitation and vice versa. A consistent look reveals a robust performance of the kriging model

lated predictions of the spatial distribution of rainfall amounts in a small rectangular spatial area in the northwest corner of South Carolina. The darker color indicates heavier predicted precipitation and the lighter color a small predicted rainfall. The consistent pattern across all ten simulations reveals a robust performance of the kriging model. A pointwise prediction at any spatial location could be obtained by averaging the predicted rainfall values at that location across all ten simulations.

4 Seasonal Trend Removal

We now analyze the geostatistical rainfall data across time. Due to the nature of our rainfall data, the seasonality is of particular interest when we model the temporal trend. We propose two methods to remove the seasonal trend in this section.

4.1 Harmonic Regression

To remove the seasonal trend, one approach is to fit a first-order harmonic regression model with terms $\sin(x)$ and $\cos(x)$. In addition, we set $x = 2\pi t$ if the period is 1. In our case, it is justifiable to set the period as 12 since the monthly rainfall is measured, and thus $x = (\pi/6)t$ is used. Hence, one can regress the precipitation y against dependent variables $\sin((\pi/6)\,t)$ and $\cos((\pi/6)\,t)$. The omnibus F-test to test for the usefulness of the trigonometric terms in this multiple regression model gives a p-value close to 1, which confirms the existence of seasonality.

One can also use a second-order harmonic model to capture more complex behavior, in which two more terms, $\sin[(4\pi/\omega)t]$ and $\cos[(4\pi/\omega)t]$ are included, where ω is the periodic parameter. However, for our rainfall data, it is unnecessary to include these two other terms since we observe no great improvement in model fit by introducing the extra terms (see Fig. 6).

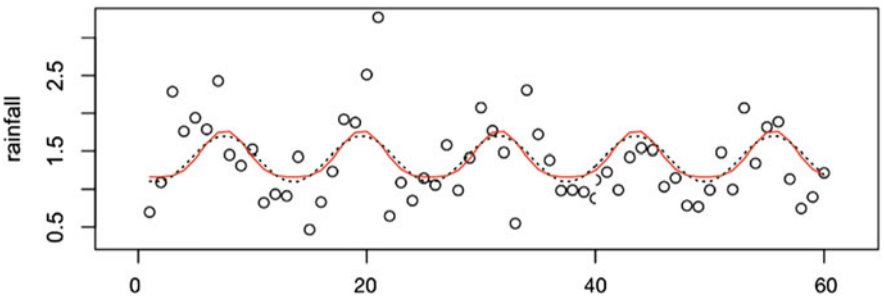

Fig. 6 The fitted model based on the first- and second-order harmonic models. The dotted line corresponds to the second-order model, and the solid red line corresponds to the first-order model

4.2 Seasonality Indicator

Another approach to model seasonality in the spatio-temporal model is the seasonal means model. Specifically, one indicator variable will be 1 if the record is collected from January to March, and will be 0 otherwise. Similarly, another dummy variable indicates the month April to June while a third dummy variable indicates July to September. Lastly, if all three variables are 0, then the observation is from the last 3 months of the year. Note that one could also include dummy variables for months in a similar way if necessary, but we have found that it is sufficient to model the means of the four seasons.

5 Precipitation Modeling: A Spatio-Temporal Perspective

In this section, we discuss how to model spatio-temporal data with two different methods, the Gaussian process (GP) model and autoregressive (AR) model. The latter model is an extension of the Gaussian process model obtained by introducing an autoregressive term.

5.1 Gaussian Process (GP) Model

The independent Gaussian process (GP) model (Cressie and Wikle 2015; Gelfand et al. 2010) is specified hierarchically in two stages,

$$\mathbf{Z}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t \tag{4}$$

$$\boldsymbol{\mu}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\eta}_t, \tag{5}$$

in which $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_n, t))^T$, which defines the response variable for all n locations at time t. It is known that $\mathbf{s}_1, \ldots, \mathbf{s}_n$ can be indexed by latitude and longitude. In the first layer, $\mathbf{Z}_t$ is defined by a simple mean model plus a pure white noise term, $\boldsymbol{\epsilon}_t$. We therefore assume that

$$\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_n), \tag{6}$$

where the σ_ϵ^2 is the pure error variance and $\mathbf{I}_n$ is the identity matrix.

The second level models $\boldsymbol{\mu}_t$ as the sum of fixed covariates and random effects at time t. The fixed term, $\mathbf{X}_t\boldsymbol{\beta}$, comes from the covariates, and $\boldsymbol{\eta}_t$ is the spatio-temporal random effects, $\boldsymbol{\eta}_t = (\eta(\mathbf{s}_1, t), \ldots, \eta(\mathbf{s}_n, t))^T$. Similar to $\boldsymbol{\epsilon}_t$, $\boldsymbol{\eta}_t$ also follows a multivariate normal distribution whose mean vector is $\mathbf{0}$. However, $\boldsymbol{\eta}_t$ has a more complicated covariance matrix than does $\boldsymbol{\epsilon}_t$.

We use the exponential function to specify the correlation matrix of the random effects. The correlation strength is solely based on the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, which is given by

$$\mathbf{\Sigma}_\eta = \sigma_\eta^2 H(\phi) + \tau^2 \mathbf{I}_n,$$

where $H(\phi) = \exp(-||\mathbf{s}_i - \mathbf{s}_j||)/\phi)$, and $||\mathbf{s}_i - \mathbf{s}_j||$ indicates the spatial distance between location i and j. This function is used to determine each element in the matrix $\mathbf{S}_\eta$, where $\mathbf{\Sigma}_\eta = \sigma_\eta^2 \mathbf{S}_\eta$. This parameterization allows σ_η^2 to capture the invariant spatial variance, and $\mathbf{S}_\eta$ is used to capture the spatial correlation.

The posterior distribution involves three layers, i.e., the prior distribution for parameters, the mean model, and the random effects model. We will set aside the prior for later discussion and use $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta}, \nu, \phi, \sigma_\eta^2, \sigma_\epsilon^2)$ to refer to the prior in general. Thus the posterior is given by

$$g(\boldsymbol{\theta}, \boldsymbol{\mu}|\mathbf{Z}) = \pi(\boldsymbol{\theta}) \times \prod_{t=1}^{N} f_n(\mathbf{Z}_t|\boldsymbol{\mu}_t, \sigma_\epsilon^2) g_n(\boldsymbol{\mu}_t|\boldsymbol{\beta}, \upsilon, \phi, \sigma_\eta^2). \tag{7}$$

To be specific, we use $f_n(\cdot)$ and $g_n(\cdot)$ to indicate an n-dimensional distribution function. In this case, each of them is a multivariate normal distribution, and n is the number of locations in the data set and N is the number of time points. $\boldsymbol{\mu}_t$ is the vector of random effects for time t and we use $\boldsymbol{\mu}$ on the left-hand side to refer to the collection of all random effects.

Since both $\mathbf{Z}_t$ and $\boldsymbol{\mu}_t$ follow a multivariate normal distribution, their density functions are given as follows:

$$f_n(\mathbf{Z}_t|\boldsymbol{\mu}_t, \sigma_\epsilon^2) = \frac{1}{\sqrt{(2\pi)^n |\sigma_\epsilon^2 \mathbf{I}_n|}} \exp\left(-\frac{1}{2\sigma_\epsilon^2}(\mathbf{Z}_t - \boldsymbol{\mu}_t)^T (\mathbf{Z}_t - \boldsymbol{\mu}_t)\right), \tag{8}$$

$$g_n(\boldsymbol{\mu}_t|\boldsymbol{S}_\eta, \sigma_\eta^2, \boldsymbol{\beta}) = \frac{1}{\sqrt{(2\pi)^n |\sigma_\eta^2 \boldsymbol{S}_\eta|}} \exp\left(-\frac{1}{2\sigma_\eta^2}(\boldsymbol{\mu}_t - \mathbf{X}_t\boldsymbol{\beta})^T \boldsymbol{S}_\eta^{-1} (\boldsymbol{\mu}_t - \mathbf{X}_t\boldsymbol{\beta})\right), \tag{9}$$

Thus the posterior distribution is given by plugging (8) and (9) into (7). The logarithm of the joint posterior distribution of the parameters for this Gaussian process model is given by

$$\log\pi(\sigma_\epsilon^2, \sigma_\eta^2, \boldsymbol{\mu}, \boldsymbol{\beta}, \upsilon, \phi|\mathbf{Z}) \propto \frac{N}{2} \log \sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2} \sum_{t=1}^{N} (\mathbf{Z}_t - \boldsymbol{\mu}_t)^T (\mathbf{Z}_t - \boldsymbol{\mu}_t)$$

$$- \frac{N}{2} \log |\sigma_\eta^2 \boldsymbol{S}_\eta| - \frac{1}{2\sigma_\eta^2} \sum_{i=1}^{N} \left(-\frac{1}{2\sigma_\eta^2}(\boldsymbol{\mu}_t - \mathbf{X}_t\boldsymbol{\beta})^T \boldsymbol{S}_\eta^{-1} (\boldsymbol{\mu}_t - \mathbf{X}_t\boldsymbol{\beta})\right) + \log \pi(\boldsymbol{\theta}).$$

We specify the prior $\pi(\boldsymbol{\theta})$ to reflect the assumption that $\boldsymbol{\beta}$, ν, ϕ, σ_η^2, and σ_ϵ^2 are mutually independent, so the joint prior is the product of the marginal prior densities, which are given as follows: All the parameters describing the mean, e.g., $\boldsymbol{\beta}$ and ρ (see Sect. 5.2) are given independent normal prior distributions, with the prior on ρ truncated to have support on $(-1, 1)$. We assume ϕ and ν both follow uniform distributions, while the prior for the precision (inverse of variance) parameter is a gamma distribution. We choose the hyperparameters to make these prior distributions very diffuse.

5.2 *Autoregressive (AR) Model*

In this section, we introduce the autoregressive model (Sahu and Bakar 2012). The hierarchical AR(1) model is given as follows:

$$\begin{aligned}\mathbf{Z}_t &= \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{\mu}_t &= \rho\boldsymbol{\mu}_{t-1} + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\eta}_t,\end{aligned}$$

where ρ denotes the unknown temporal correlation parameter assumed to be in the interval $(-1, 1)$. Obviously, for $\rho = 0$, these models reduce to the GP model described in Sect. 5.1.

The autoregressive model requires specification of the initial term, the first random effect, which has mean $\boldsymbol{\beta}_0$ and covariance matrix $\sigma_0^2\boldsymbol{S}_0$. The correlation matrix $\boldsymbol{S}_0$ is obtained using the exponential correlation function. The derivation of the posterior distribution is similar to that in GP model with $\rho = 0$. The logarithm of the posterior distribution of the parameters is now given by

$$\begin{aligned}\log\pi(\sigma_\epsilon^2, \sigma_\eta^2, \boldsymbol{\mu}, \boldsymbol{\beta}, \upsilon, \phi|\mathbf{Z}) \propto\ & \frac{N}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{N}(\mathbf{Z}_t - \boldsymbol{\mu}_t)^T(\mathbf{Z}_t - \boldsymbol{\mu}_t) \\ & - \frac{N}{2}\log|\sigma_\eta^2\boldsymbol{S}_\eta| \\ & - \frac{1}{2\sigma_\eta^2}\sum_{i=1}^{N}\left(-\frac{1}{2\sigma_\eta^2}(\boldsymbol{\mu}_t - \rho\boldsymbol{\mu}_{t-1} - \mathbf{X}_t\boldsymbol{\beta})^T\boldsymbol{S}_\eta^{-1}(\boldsymbol{\mu}_t - \rho\boldsymbol{\mu}_{t-1} - \mathbf{X}_t\boldsymbol{\beta})\right) \\ & - \frac{1}{2}\log|\sigma_0^2\boldsymbol{S}_0| - \frac{1}{2\sigma_0^2}(\boldsymbol{\mu}_0 - \boldsymbol{\beta}_0)^T\boldsymbol{S}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\beta}_0) + \log\pi(\boldsymbol{\theta})\end{aligned}$$

Note that $\boldsymbol{\beta}_0$ is only a mean vector for the initial random effect term, which is different from $\boldsymbol{\beta}$, which refers to regression coefficients corresponding to covariates $\mathbf{X}$. In other words, the terms in the last line (except $\log\pi(\boldsymbol{\theta})$) derive from the initial random effect term.

5.3 Model Fitting

In this section, we fit the AR(1) model with monthly precipitation data from the beginning of year 2011 to the end of year 2015. A natural log transformation was initially applied to the precipitation to improve the model fit and ensure positive predicted rainfall values once we back-transform by exponentiating the predicted log-rainfall values. We include temperature range, sea surface temperature, and elevation as monthly covariates.

We initially found that ordinary temperature measurements such as the monthly average temperature were not apparently related to precipitation after accounting for the season and thus we did not include these in the model. However, measurements of variability in temperature over each month, e.g., the range of daily maxima and the range of daily minima over a month, were believed to have an effect on precipitation and thus we include these to determine whether their effects are significant.

We also include a flood-year indicator as a dummy variable, where data from 2015 is labeled as 1 and otherwise 0, to account for the unusual October precipitation amounts in this year. Interaction terms involving the dummy variable were also tested, none of which were statistically significant and were thus removed from the final model. The acceptance rate from Metropolis step for all parameters is 42.97% and a brief summary of model fitting details is given as follows:

```
-----------------------------------------------------
Model: AR
Call: LOG ~ RANGE_OVERALL + RANGE_LOW  + RANGE_HIGH
+ SST + ELEVATION + SST * RANG E_LOW + Year2015

Iterations: 5000
nBurn: 1000
Acceptance rate: 29.76
-----------------------------------------------------
Parameters
                  Mean  Median     SD Low2.5p Up97.5p
(Intercept)     0.3635  0.3689 0.1363  0.0894  0.6265
RANGE_OVERALL -0.0006 -0.0006 0.0017 -0.0039  0.0027
RANGE_LOW       0.0017  0.0017 0.0030 -0.0040  0.0078
RANGE_HIGH      0.0006  0.0007 0.0011 -0.0016  0.0028
SST            -0.0057 -0.0058 0.0045 -0.0142  0.0033
ELEVATION       0.0001  0.0001 0.0001  0.0000  0.0002
Year2015        0.0808  0.0810 0.0180  0.0450  0.1154
RANGE_LOW:SST -0.0001 -0.0001 0.0001 -0.0003  0.0001
rho             0.0756  0.0757 0.0151  0.0466  0.1054
sig2eps         0.0054  0.0054 0.0002  0.0051  0.0057
sig2eta         0.0764  0.0739 0.0121  0.0617  0.1073
phi             0.0501  0.0502 0.0090  0.0322  0.0659
-----------------------------------------------------
```

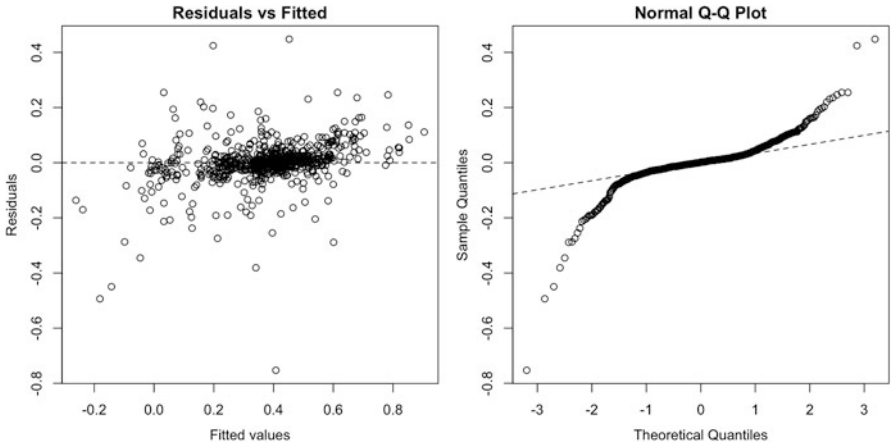

Fig. 7 The residual plot and QQ plot from AR(1) prediction

The dummy variable for year 2015 is significant. After back-transforming, we can say the predicted monthly rainfall for 2015 is $\exp(0.0808) = 1.084$ times greater than the predicted monthly rainfall in other years, holding other predictors fixed. This is consistent with the flooding event in the fall of 2015. Another finding is that elevation might be an explanatory factor to the rainfall since higher elevation relates to higher volumes of precipitation. In addition, a statistically significant and positive ρ indicates that a rainy month might tend to precede another one. On the other hand, the SST has a marginally negative effect on the rainfall prediction but is not significant based on the 95% credible interval.

We also obtain the residuals and the QQ plot in Fig. 7. There is no obvious pattern in the residual plots. However, the residuals show deviations in the tails to some extent from normality based on the QQ plot on the right panel, which indicates a heavy-tailed error distribution and lack of symmetrical pattern (e.g., Samadi et al. 2017).

6 Model Comparison: State-Space Model vs. Gaussian Process

Another framework for spatio-temporal data analysis is the dynamic state-space model. A formulation of the spatio-temporal framework (Stroud et al. 2001) is specified as follows:

$$
\begin{aligned}
y_t(\mathbf{s}) &= \mathbf{x}_t(\mathbf{s})^T \boldsymbol{\beta}_t + u_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad \epsilon_t(\mathbf{s}) \sim N(0, \tau_t^2) \\
\boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N(0, \boldsymbol{\Sigma}_\eta) \\
\mu_t(\mathbf{s}) &= \mu_{t-1}(\mathbf{s}) + w_t(\mathbf{s}), \quad w_t(\mathbf{s}) \sim GP(\mathbf{0}, C_t(\cdot, \boldsymbol{\theta}_t)).
\end{aligned}
$$

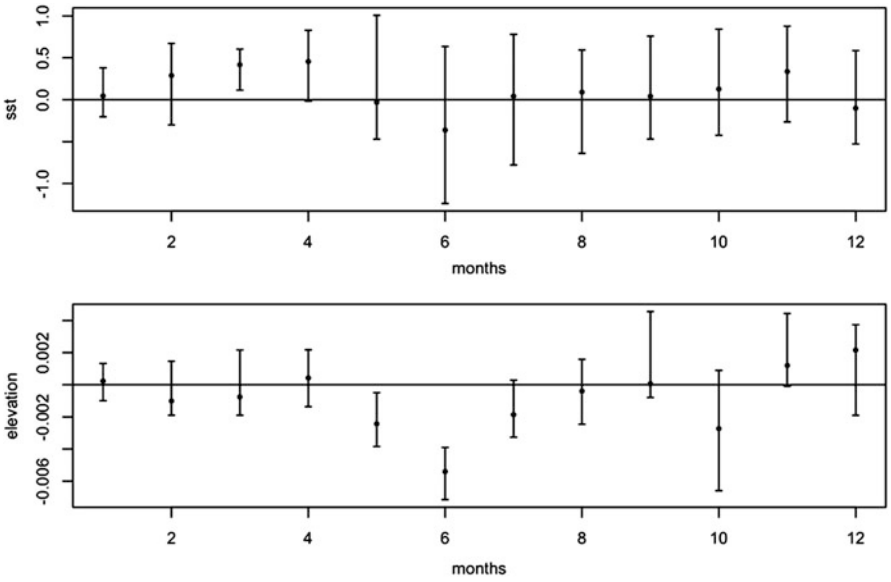

Fig. 8 The 95% confidence interval for β_1 (the SST-related variable) and β_2 (elevation) over 12 months in 2015

Here $\mathbf{x}_t(\mathbf{s})$ is a $p \times 1$ vector of predictors and $\boldsymbol{\beta}_t$ is a $p \times 1$ vector of coefficients. The $GP(\mathbf{0}, C_t(\cdot, \boldsymbol{\theta}_t))$ denotes a spatial Gaussian process with covariance function $C_t(\cdot, \boldsymbol{\theta}_t)$. We further specify $C_t(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}_t)) = \sigma_t^2 \rho(\mathbf{s}_1, \mathbf{s}_2; \phi_t)$, where $\boldsymbol{\theta}_t = \{\sigma_t^2, \phi_t\}$ and $\rho(\cdot; \phi)$ is a correlation function with ϕ controlling the correlation decay.

The same response variable and covariates with AR(1) model are used when fitting the state-space model. The R package `spBayes` (Finley et al. 2007) provides a framework to sample from parameters and posterior. The 95% credible interval for sea surface temperature and elevation are plotted for all 12 months in 2015.

The state-space model allows for a more detailed monthly look of the effect of covariates. For instance, one can conclude that, based on Fig. 8, the SST-based variable effects the rainfall amount in a more significant manner during the first few months of the year. These results strengthen the previous findings of Häkkinen (2000), Mehta et al. (2000), Wang et al. (2006), and Dima and Lohmann (2010), and further support the hypothesis that the variability of North Atlantic SST is coherent with the fluctuations of the rainfall pattern and occurrence. In other words, intense ocean–atmosphere coupling exists in the North Atlantic, particularly during winter. In contrast, elevation is more related to the precipitation in June and October, when heavier rainfall data are observed. This covariate specifies a convective mode that is widely recognized as an important contributor to the probability and type of severe convective rainfall during summer and early fall in the southeast region. The residual plot and the QQ plot for the state-space are shown in Fig. 9. We see the heavy-tailed error pattern is still apparent in this model, based on the QQ plot.

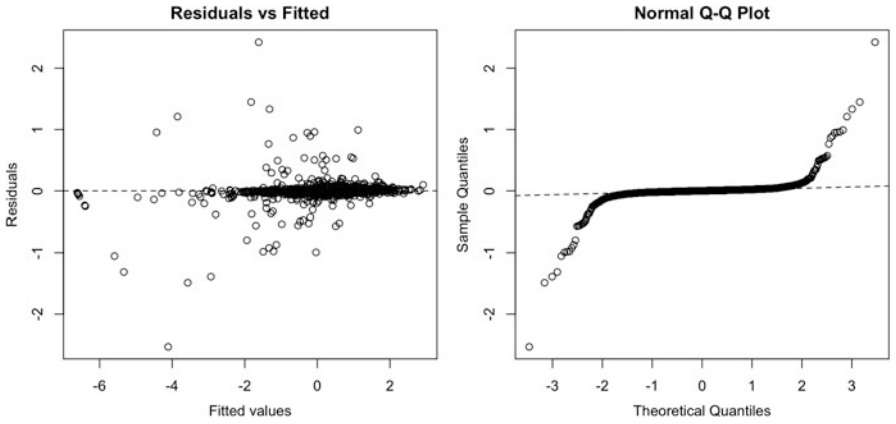

Fig. 9 The residual plot and QQ plot from the state-space model

7 Discussion

We have presented both spatial and spatio-temporal models for rainfall in South Carolina during a period including one of the most destructive storms in state history. Our models have allowed us to determine several covariates that affect the rainfall and to interpret their effects. In particular, the flood year of 2015 was an important indicator of rainfall and elevation also had a positive significant effect on precipitation. There was a significant positive correlation in rainfall measurements over time. Finally, our novel SST index provided some evidence that cooler nearby sea temperatures corresponded to higher rainfall at in land sites although this SST effect was not significant at the 0.05 level based on a 95% credible interval for its effect.

A spatial prediction at a new location and a temporal prediction at a future time point can be obtained based on the posterior predictive distribution for $Z(\mathbf{s}_0, t')$, where $\mathbf{s}_0$ denotes a new location and t' is a future time point. Further details regarding these predictions are provided in Cressie and Wikle (2015) for the GP models, and Sahu and Bakar (2012) for the AR models.

A limitation of the study, and a direction for future research, is that the model does not account for the apparent heavy-tailed nature of the errors. Methods involving generalized extreme value distribution (Rodríguez et al. 2016) could possibly be adapted to this model to help handle this heavy-tailed error structure, but such research is still relatively new in the spatio-temporal modeling literature.

References

Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. Wiley, Hoboken (2003)
Bakar, K.S., Sahu, S.K.: spTimer: spatio-temporal Bayesian modeling using R. J. Stat. Softw. **63**(15), 1–32 (2015)

Banerjee, S., Carlin, B.P., Gelfand, A.E.: Hierarchical Modeling and Analysis for Spatial Data. CRC Press, Boca Raton (2014)
Benzécri, J.P.: L'Analyse des Données. Dunod, Paris (1973)
Berne, A., Delrieu, G., Boudevillain, B.: Variability of the spatial structure of intense Mediterranean precipitation. Adv. Water Resour. **32**(7), 1031–1042 (2009)
Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., Pebesma, E.J.: Applied Spatial Data Analysis with R. Springer, New York (2008)
Ciach, G.J., Krajewski, W.F.: Analysis and modeling of spatial correlation structure in small-scale rainfall in central Oklahoma. Adv. Water Resour. **29**(10), 1450–1463 (2006)
Cressie, N.: Statistics for Spatial Data. Wiley, New York (1993)
Cressie, N., Wikle, C.K.: Statistics for Spatio-Temporal Data. Wiley, New York (2015)
Deidda, R.: Rainfall downscaling in a space-time multifractal framework. Water Resour. Res. **36**(7), 1779–1794 (2000)
Delhomme, J.P.: Kriging in the hydrosciences. Adv. Water Resour. **1**(5), 251–266 (1978)
Delfiner, P., Delhomme, J.P.: Optimum Interpolation by Kriging. Ecole Nationale Supérieure des Mines, Paris (1975)
Diggle, P.J., Tawn, J.A., Moyeed, R.A.: Model-based geostatistics. J. R. Stat. Soc.: Ser. C: Appl. Stat. **47**(3), 299–350 (1998)
Dima, M., Lohmann, G.: Evidence for two distinct modes of large-scale ocean circulation changes over the last century. J. Clim. **23**(1), 5–16 (2010)
Dumitrescu, A., Birsan, M.V., Manea, A.: Spatio-temporal interpolation of sub-daily (6 h) precipitation over Romania for the period 1975–2010. Int. J. Climatol. **36**(3), 1331–1343 (2016)
Ferraris, L., Gabellani, S., Rebora, N., Provenzale, A.: A comparison of stochastic models for spatial rainfall downscaling. Water Resour. Res. **39**(12), 1368 (2003). https://doi.org/10.1029/2003WR002504
Finley, A.O., Banerjee, S., Carlin, B.P.: `spBayes`: an R package for univariate and multivariate hierarchical point-referenced spatial models. J. Stat. Softw. **19**(4), 1 (2007)
Gelfand, A.E., Diggle, P., Guttorp, P., Fuentes, M.: Handbook of Spatial Statistics. CRC Press, Boca Raton (2010)
Georgakakos, K.P., Kavvas, M.L.: Precipitation analysis, modeling, and prediction in hydrology. Rev. Geophys. **25**(2), 163–178 (1987)
Häkkinen, S.: Decadal air-sea interaction in the North Atlantic based on observations and modeling results. J. Clim. **13**(6), 1195–1219 (2000)
Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970)
Hughes, J., Haran, M.: Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. J. R. Stat. Soc. Ser. B Stat Methodol. **75**(1), 139–159 (2013)
Isaaks, H.E., Srivastava, R.M.: Applied Geostatistics. Oxford University Press, New York (1989)
Kumar, P., Foufoula-Georgiou, E.: Characterizing multiscale variability of zero intermittency in spatial rainfall. J. Appl. Meteorol. **33**(12), 1516–1525 (1994)
Ly, S., Charles, C., Degre, A.: Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. Hydrol. Earth Syst. Sci. **15**(7), 2259–2274 (2011)
Matheron, G.: Principles of geostatistics. Econ. Geol. **58**(8), 1246–1266 (1963)
Mehta, V., Suarez, M., Manganello, J.V., Delworth, T.D.: Oceanic influence on the North Atlantic oscillation and associated northern hemisphere climate variations: 1959–1993. Geophys. Res. Lett. **27**(1), 121–124 (2000)
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)
National Oceanic and Atmosphere Administration, U.S. Department of Commerce: Service assessment: the historic South Carolina floods of October 1–5, 2015. www.weather.gov/media/publications/assessments/SCFlooding_072216_Signed_Final.pdf (2015). Accessed 4 Dec 2017

Rodríguez, S., Huerta, G., Reyes, H.: A study of trends for Mexico city ozone extremes: 2001–2014. Atmósfera **29**(2), 107–120 (2016)

Sahu, S.K., Bakar, K.S.: Hierarchical Bayesian autoregressive models for large space–time data with applications to ozone concentration modeling. Appl. Stoch. Model. Bus. Ind. **28**(5), 395–415 (2012)

Samadi, S., Tufford, D., Carbone, G.: Estimating hydrologic model uncertainty in the presence of complex residual error structures. Stoch. Environ. Res. Risk Assess. **32**(5), 1259–1281 (2018)

Sharon, D.: Spatial analysis of rainfall data from dense networks. Hydrol. Sci. J. **17**(3), 291–300 (1972)

Stroud, J.R., Müller, P., Sansó, B.: Dynamic models for spatio-temporal data. J. R. Stat. Soc. Ser. B Stat. Methodol. **63**(4), 673–689 (2001)

Tabios III, Q.G., Salas, J.D.: A comparative analysis of techniques for spatial interpolation of precipitation. Water Resour. Bull. **21**(3), 365–380 (1985)

Thiessen, A.H.: Precipitation averages for large areas. Mon. Weather Rev. **39**(7), 1082–1084 (1911)

Troutman, B.M.: Runoff prediction errors and bias in parameter estimation induced by spatial variability of precipitation. Water Resour. Res. **19**(3), 791–810 (1983)

Wang, C., Enfield, D.B., Lee, S.K., Landsea, C.W.: Influences of the Atlantic warm pool on western hemisphere summer rainfall and Atlantic hurricanes. J. Clim. **19**(12), 3011–3028 (2006)