

Dear Author,

Here are the final proofs of your article. Please check the proofs carefully.

All communications with regard to the proof should be sent to SpringerOpen_Production@spi-global.com.

Please note that at this stage you should only be checking for errors introduced during the production process. Please pay particular attention to the following when checking the proof:

- Author names. Check that each author name is spelled correctly, and that names appear in the correct order of first name followed by family name. This will ensure that the names will be indexed correctly (for example if the author's name is 'Jane Patel', she will be cited as 'Patel, J.').
- Affiliations. Check that all authors are cited with the correct affiliations, that the author who will receive correspondence has been identified with an asterisk (*), and that all equal contributors have been identified with a dagger sign (†).
- Ensure that the main text is complete.
- Check that figures, tables and their legends are included and in the correct order.
- Look to see that queries that were raised during copy-editing or typesetting have been resolved.
- Confirm that all web links are correct and working.
- Ensure that special characters and equations are displaying correctly.
- Check that additional or supplementary files can be opened and are correct.

Changes in scientific content cannot be made at this stage unless the request has already been approved. This includes changes to title or authorship, new results, or corrected values.

How to return your corrections

Returning your corrections via online submission:

- Please provide details of your corrections in the online correction form. Always indicate the line number to which the correction refers.

Returning your corrections via email:

- Annotate the proof PDF with your corrections.
- Send it as an email attachment to: SpringerOpen_Production@spi-global.com.
- Remember to include the journal title, manuscript number, and your name when sending your response via email.

After you have submitted your corrections, you will receive email notification from our production team that your article has been published in the final version. All changes at this stage are final. We will not be able to make any further changes after publication.

Kind regards,

SpringerOpen Production Team

RESEARCH

Open Access

Spatio-temporal analysis of flood data from South Carolina

Haigang Liu^{1*}, David B. Hitchcock¹ and S. Zahra Samadi²

*Correspondence:

haigang@email.sc.edu

¹Department of Statistics, University of South Carolina, 1523 Greene St 29201 Columbia SC, United States
 Full list of author information is available at the end of the article

Abstract

To investigate the relationship between flood gage height and precipitation in South Carolina from 2012 to 2016, we built a conditional autoregressive (CAR) model using a Bayesian hierarchical framework. This approach allows the modelling of the main spatio-temporal properties of water height dynamics over multiple locations, accounting for the effect of river network, geomorphology, and forcing rainfall. In this respect, a proximity matrix based on watershed information was used to capture the spatial structure of gage height measurements in and around South Carolina. The temporal structure was handled by a first-order autoregressive term in the model. Several covariates, including the elevation of the sites and effects of seasonality, were examined, along with daily rainfall amount. A non-normal error structure was used to account for the heavy-tailed distribution of maximum gage heights. The proposed model captured some key features of the flood process such as seasonality and a stronger association between precipitation and flooding during summer season. The model is able to forecast short term flood gage height which is crucial for informed emergency decision. As a byproduct, we also developed a Python library to retrieve and handle environmental data provided by some main agencies in the United States. This library can be of general usefulness for studies requiring rainfall, flow, and geomorphological information over specific areas of the conterminous US.

Keywords: Flood, Watershed, CAR model, Spatiotemporal analysis

Introduction

During October 2–5, 2015, an extraordinary rainfall event took place in the Carolinas, many parts of which observed 500-year-event levels of precipitation. Accumulation of rainfall amount reached 24.23 inches (61.57 centimeters) near Boone Hall (Mount Pleasant, Charleston County) by 11:00 a.m. Eastern Time on October 4, 2015. The rainfall peaked on October 4 with a 24-hour total of 16.69 inches (42.39 centimeters) of precipitation; and the total 48-hour precipitation during October 3–4 was more than 20 inches (51 centimeters). The likelihood of the rainfall amounts ranged from anywhere between a 1-in-250-year event to a 1-in-1000-year event in the study region with some places such as Columbia and Lexington, SC receiving more than 17 inches (43 centimeters) of rain over a four-day period (Phillips et al. 2018). Columbia, the capital of South Carolina, broke its all-time wettest 1-day, 2-day, and 3-day periods on record (e.g., Bonnin et al. 2006). The

rainfall in Columbia far exceeded the two values of National Oceanic and Atmospheric Administration (NOAA) calculated 1,000-year events of 12.8 inches (32.5 centimeters) and 14.1 inches (35.8 centimeters), respectively (NOAA Atlas 14 volume 2; see Frederick and Miller (1979)). Charleston International Airport observed a record 24-hour rainfall of 11.5 inches (29.2 centimeters) on October 3 (Santorelli, Oct. 4, 2015). Some areas experienced more than 20 inches (51 centimeters) of rainfall over the five-day period.

Flooding from this event resulted in 19 fatalities, according to the South Carolina Emergency Management Department, and South Carolina state officials reported damage losses of \$1.492 billion (National Oceanic and Atmosphere Administration 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 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 four days. A low-pressure system off the U.S. 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 (National Oceanic and Atmosphere Administration 2015).

In this article, we use geostatistical analysis to investigate the stochastic relationship and the dynamics between rainfall and flooding. 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 at 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, (Delfiner and Delhomme 1975) and (Delhomme 1978) applied various geostatistical methods such as variograms and kriging methods in modeling rainfall. The work of Berne et al. (2009), (Ciach and Krajewski 2006; Deidda 2000; Dumitrescu et al. 2016; Ferraris et al. 2003; Georgakakos and Kavvas 1987; Isaaks and Srivastava 1989; Kumar and Foufoula-Georgiou 1994; Ly et al. 2011) Serinaldi and Kilsby (2014), Troutman (1983) Tabios and Salas (1985), (Villarini et al. 2009) further advanced the application of geostatistical methods in rainfall and flood analysis. 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.

This article is arranged as follows: In “Data sources” section, we provide an overview of our use of data munging to obtain the precipitation and gage height data, since the scraping, cleaning, aggregating and transforming of data constitute a major part of our study. “Adjacency matrix and watershed” section discusses the binary adjacency matrix, which is pivotal to the conditional autoregressive model since it accounts for the spatial correlation based on watershed information. In “Model description” section, our model fitting approach and results are detailed, including a remedy for some noted heavy-tailed error behavior. Lastly, we compare our results using the conditional autoregressive model with results using other popular models such as random forest (RF), based on metrics like mean square error.

Q2

Q3

Data sources

In this section, we discuss our data sources and the necessary data munging steps we used in our study. We compiled a dataset for 94 unique locations in South Carolina with precipitation, elevation, gage height, along with basin information, over a span of five years. We primarily cover the collection of variables such as daily rainfall and gage height, since we are interested in exploring the dynamics between them. We mention the watershed information briefly since it is used in defining the proximity matrix. A detailed discussion of this can be found in “[Adjacency matrix and watershed](#)” section.

Precipitation

The National Weather Service (NWS) collects precipitation data at 12 Contiguous United States (CONUS) River Forecast Centers (RFCs). The precipitation is recorded using a multisensor approach. Hourly estimates from weather radars are compared to ground gage reports, and a correction factor is calculated and applied to the radar field (Daly et al. 2000). For areas where radar coverage is not accessible, satellite precipitation estimates can be used to construct the multisensor field (Daly et al. 1994). Note that this method has been applied to South Carolina and most other eastern states, whereas a different method is used to process precipitation data in mountainous areas west of the Continental Divide.

The precipitation data are then mosaicked into a gridded field with a spatial resolution of four by four kilometers. The record is an accumulation of 24-hour periods and 1200 GMT is used as the ending time for a 24-hour total. Spatially, the original dataset extends well beyond the U.S. border, most notably north of Washington and Idaho and west of Texas, in order to model rivers that flow into the United States. However, only the observations within South Carolina and nearby states are retained in our study since the rainfall far outside the state is unlikely to have a major effect on flood gage heights in the short term. Available data dates back to 2004 and still is actively updated by NWS. Rainfall values from 2012 to 2016 (inclusive) were retrieved for our study.

The raw data are archived in <https://water.weather.gov/precip/archive/>. The major challenges of handling this dataset are parsing the raw data (in NetCDF format) and filtering out values from irrelevant regions and dates. “[Miscellaneous code](#)” section is a brief introduction of our proposed approaches to streamline the data preprocessing steps by developing a Python library.

Gage height

Gage height (also known as stage) is the height of the water in the stream above a reference point. Gage height refers to the elevation of the water surface in the specific pool at the streamgaging station, not along the entire stream (USGS, 2011). Gage height also is not exactly the same as the depth of the stream. Since the stage baselines are set in a case-by-case manner across locations, we subtract the station-wise historical median (the median gage height for each location, over a 10-year period) from each gage height measurement to make the measurements comparable (see “[Model description](#)” section for details). This is done as a preliminary centering step before we fit the model.

The U.S. Geological Survey (USGS) provides an archive of approximately 1.9 million observation sites of all kinds in all 50 states, the District of Columbia, Puerto Rico, the Virgin Islands, Guam, American Samoa, under the *Water Data for the Nation* portal on its website. More than 1000 such sites can be found within the border of South Carolina.

127 However, the site count is drastically reduced when we focus on locations measuring
 128 surface water and exclude those that have ceased functioning. Eventually, we have approx-
 129 imately 150 to 200 locations (depending on the timeframe) within South Carolina that
 130 give a valid reading of the gage level on a daily basis. One can either use the interface pro-
 131 vided by USGS or the `data.download_flood()` function from our Python library
 132 (“Miscellaneous code” section) to download the data. The former comes with a graphical
 133 user interface but may be harder to maneuver when multiple sites are needed. The latter,
 134 on the other hand, allows user customization to a greater degree.

135 Notably, the precipitation and the gage height are measured in different locations, since
 136 the former are measured in gridded fields and the latter are located at major rivers and
 137 dams. We implemented a “blurry lookup” approach to combine the two pieces of informa-
 138 tion. For readers familiar with SQL, the algorithm is similar to a left join, where all rows
 139 in the left table (gage height) are retained, and on the right (precipitation) only records
 140 with matching keys are kept. This is different from a typical left join in that although a
 141 latitude and longitude pair serves as the key, typical merging is not feasible due to the
 142 location mismatch. Hence, the merging is done by finding the nearest neighbor. For each
 143 row (location i) in the gage height table, we find a location j in the precipitation table that
 144 is closest to it. We add the rainfall information at location j to location i for each i in the
 145 left table. Admittedly, this is not ideal since the precipitation and gage height are not from
 146 the exact same location, but the high resolution of the precipitation data (4×4 km) makes
 147 this issue less critical.

148 Additionally, since a fair amount of records are missing in the dataset, we first calculate
 149 the missing data ratio, which is the percentage of days with missing records over the total
 150 number of days during the aforementioned time span (2012-2016). We discard the loca-
 151 tion if the missing data ratio is beyond a certain threshold. We strike a balance between a
 152 larger sample size and better data completeness with the help of Fig. 1, which shows how
 153 many locations are retained for different time spans and thresholds. Note that the x-axis
 154 is number of years from 2016 counting backwards. For instance, there are 120 locations
 155 retained in the dataset for 2016 with a 95% complete-data threshold. Based on Fig. 1, we
 156 pick 90% (94 unique stations) as the complete-data threshold for a time span of five years,
 157 since further increasing the threshold leads to a significant decrease in the amount of
 158 available gaging stations.

F1

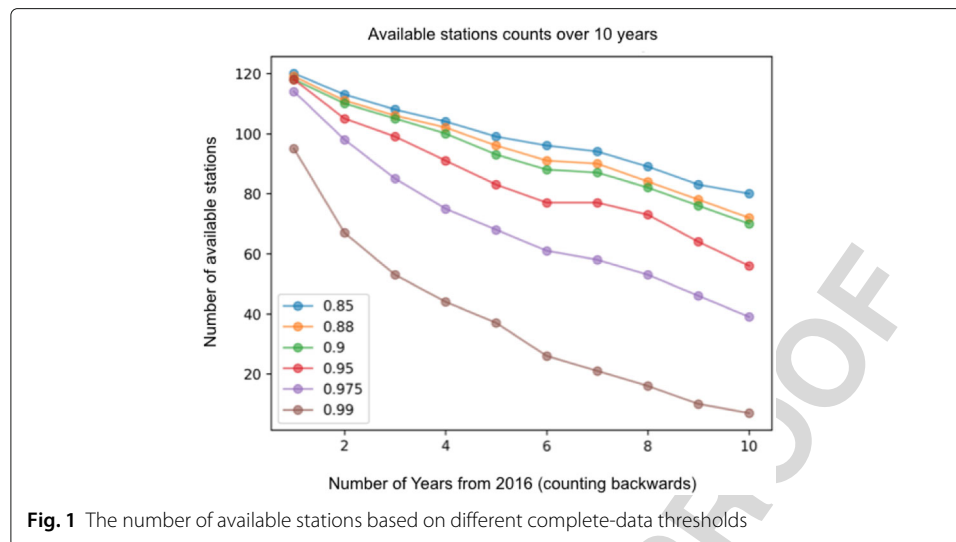
159 Imputation for the remaining locations is based on the temporal adjacency. In other
 160 words, to fill the missing gage height values on certain dates, the weighted average of
 161 values from neighboring dates is used, that is

$$Y_t = \frac{w_{t-2}}{w^*} Y_{t-2} + \frac{w_{t-1}}{w^*} Y_{t-1} + \frac{w_{t+1}}{w^*} Y_{t+1} + \frac{w_{t+2}}{w^*} Y_{t+2},$$

162 where Y_t is the missing value at time t and $w^* = w_{t-2} + w_{t-1} + w_{t+1} + w_{t+2}$. We set
 163 $w_{t-1} = w_{t+1} = 2$ and $w_{t-2} = w_{t+2} = 1$ since observations closer in time to the miss-
 164 ing value should be more informative. Alternatively, one can fill the missing values based
 165 on spatial closeness, but we argue that the gage height measurements in the *same* loca-
 166 tion may change quite steadily and continuously. Filling missing data spatially is less ideal
 167 since doing so would involve pooling together different observation locations, which are
 168 associated with varying gage baseline levels and landscapes.

Q4

Q5



Elevation

The elevation information is obtained based on the Shuttle Radar Topography Mission (SRTM), which is an international research effort that obtained digital elevation models on a near-global scale from 56°S to 60°N (Farr et al. 2007). The 30-meter topographic data products are publicly distributed by the USGS along with the 90-meter data. These data are made available via an Earth Explorer on the US Geological Survey website in a .tiff format. We retrieve the elevation information from the 90-meter data for the aforementioned gage locations by matching the latitude and longitude. Elevation of the nearest neighbor is used if an exact match cannot be found.

Other covariates

Besides the precipitation and elevation, we also include three dummy variables to account for the seasonality in the data. The three dummy variables respectively take values 1 if the data record is from the spring (March through May), summer (June through August), or fall (September through November), and 0 otherwise. More importantly, interaction terms of the season indicator and precipitation are included, so that we can explore whether a difference in rainfall effect on flood levels exists across seasons. Specifically, if the interaction variable between spring and precipitation manifests itself as positive and significant, one can conclude that during March through May, rainfall increases are likely to lead to an even greater average rise in gage heights than in the baseline season (winter).

Basins and watersheds

The watershed information is pivotal to our model in a way that is different from elevation or precipitation. Rather than entering the model as a covariate, the watershed membership is used for the adjacency matrix \mathbf{W} , whose definition can be found in "Adjacency matrix and watershed" section, along with a more detailed account of the watershed system. In this section, we focus on preprocessing such information into a well-structured format.

USGS hosts the watershed information by state on Amazon Web Services (AWS), which is publicly available. It is a repository of contour files with varying sizes. A 4-digit

hydrologic unit code (HUC) is less localized and covers a larger area than a 6-digit HUC (HUC-6), for example. We use the contour information to define the watershed membership. Practically, a categorical variable with the watershed name is added for each available location. We decide to use the 6-digit hydrological unit to categorize all available locations into four regions. We discuss this more in “Adjacency matrix and watershed” section.

Miscellaneous code

A Python library, `climate_data_toolkit`, is developed in parallel with our study, which accomplishes two goals. First, we intend to streamline the process of downloading and preprocessing raw data from different sources. Rather than using varying user interfaces for different databases, one can achieve the same result nearly instantly by function calls like `get_flood()`. Second, we package our models that we use in “Model description” section with a user-friendly interface. Hence, a compilation of a few Python modules, or, a library, is a natural choice for this purpose. In addition, we also have a plotting system, which is a handy tool to visualize spatial data, since it can display spatial elements such as markers and contours on top of an OpenStreet Map in a manner reminiscent of the R package `ggplot` (Wickham 2016). The Python library is hosted on Github, and users can find the source code and help documents at <https://github.com/HaigangLiu/spatial-temporal-py>. Alternatively, the package also supports `pip install`, which is a convenient command line tool for package management.

Adjacency matrix and watershed

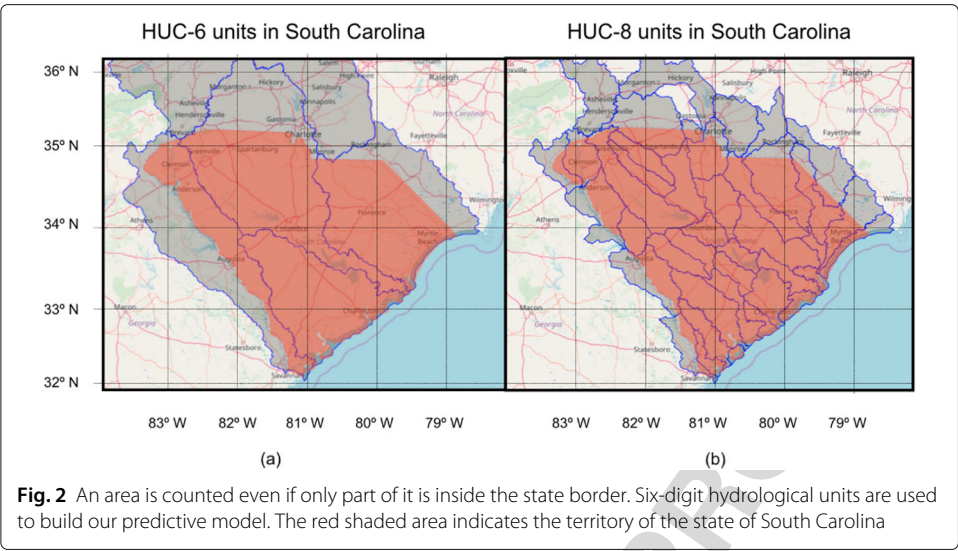
The concept of the adjacency/proximity matrix \mathbf{W} , first introduced by Cressie (1993) in areal data analysis, is pivotal to reflect the dependence among nearby locations. The entries w_{ij} in the adjacency matrix describe the connection between location i and j in some fashion. Typically, one builds the adjacency matrix based on either a distance or a binary status. For instance, one can define $w_{ij} = 1$ if i and j share some common boundary and 0 otherwise. Alternatively, w_{ij} could reflect “distance” between units. Further modifications can be made as well. For instance, we could set $w_{ij} = 1$ for all i and j within a specified distance. Or, for a given i , we could define $w_{ij} = 1$ if j is one of the K nearest (in distance) neighbors of i . In the context of our study, we define the adjacency matrix based on the watershed information since it serves as an indicator of flood activity and its domain. Specifically, if two locations i and j are within the same watershed, then $w_{ij} = 1$ and $w_{ij} = 0$ otherwise. Note that this choice reflects the river network connections.

A watershed is an area of land where rainfall accumulates and drains off into a river, bay or other body of water (Betson 1964). Other terms used interchangeably with watershed are drainage area, catchment basin and water basin. The watersheds have different scales and the hierarchy is reflected by HUC system. For instance, an area indexed by a two-digit code is composed of several smaller four-digit basins. There are six levels in the hierarchy, represented by hydrologic unit codes from two to twelve digits long, called regions, subregions, basins, subbasins, watersheds, and subwatersheds (Seaber et al. 1987). Figure 2 illustrates all the six-digit and eight-digit hydrological units that are located fully or partially in South Carolina.

Notably, basins (areas indexed by a HUC-6 code) appear to be an appropriate granularity when we investigate the watersheds in South Carolina, since these hydrological areas are neither too dense nor sparse in terms of data points. Table 1 summarizes the unique

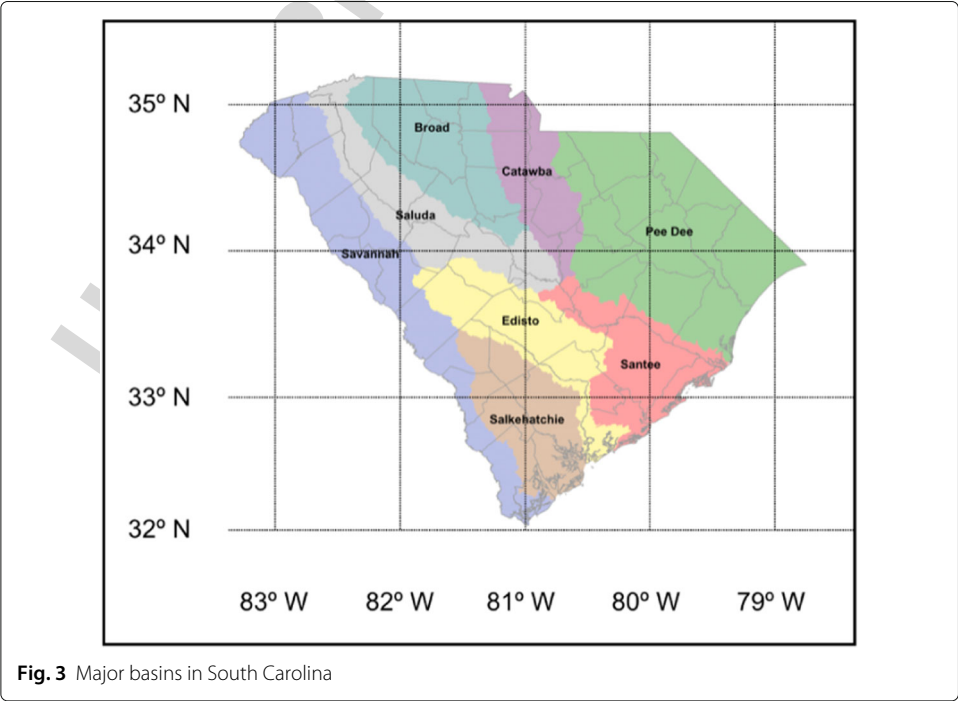
F2

T1



locations in our data set in each HUC region. Since the HUC regions do not exactly match state borders, we retain the regions in which the majority of observations are located in South Carolina, namely, Savannah, Santee, Edisto-South Carolina Coastal and Lower Pee Dee (Fig. 2 left panel, clockwise from left to right) in terms of HUC-6 regions. Note that the Savannah, Lower Pee Dee and Santee watersheds are not exclusively located in South Carolina, as there are initially formed and originated from Georgia and North Carolina. We consider these locations outside South Carolina as part of our data set as well when collecting flood gage height and other variables, since they are integral parts of the watersheds and contribute to the runoff generation, as well.

Q6



t1.1 **Table 1** The counts of available locations in each HUC-6 region

t1.2	Name	Count
t1.3	Santee	108
t1.4	Lower Pee Dee	24
t1.5	Savannah	21
t1.6	Edisto-South Carolina Coastal	19

250 Alternatively, to define the proximity matrix, one can use a *river basin* system as well,
 251 which is a product of the first watershed planning activities in 1970s by the state of South
 252 Carolina. According to the river basin system, eight mutually exclusive areas are defined:
 253 Broad River, Savannah River, Pee Dee River, Santee River, Catawba River, Catawba River,
 254 Saluda River, Edisto River and Salkehatchie River. However, we prefer the watershed
 255 system since it is not constrained by state borders. Furthermore, based on the river
 256 basin segmentation, some river basins, e.g., Salkehatchie, contain as few as two unique
 257 observing stations. Such sparsity might lead to less stable parameter estimates.

258 Model description

259 A neighborhood structure to reflect the spatial structure is pivotal in some spatial and
 260 spatiotemporal analyses. Often, one can define the neighborhood structure based on dis-
 261 tance from certain centroids or similarity of an auxiliary variable (Cressie 1993). In our
 262 study, watershed information is used to construct the neighborhood structure since it
 263 outlines the domain of hydrological water movement activity, and we define measured
 264 stations within the same basin as neighbors.

265 The observed variables in our study include Y_i , the gage height, and p explanatory vari-
 266 ables, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$. In order to compute the shifting patterns in flood records,
 267 this research incorporated exogenous covariates such as precipitation, dummy vari-
 268 ables for spring, summer and fall, as well as the interactions between the seasonality
 269 dummy variables and precipitation into the Conditional Autoregressive (CAR) model.
 270 In addition, the elevation of each location was considered as a covariate but not included
 271 in the final model, as noted in “Result” section. The Conditional Autoregressive model for
 272 the responses, $\mathbf{Y} = (Y_1, \dots, Y_n)'$, is formulated by specifying the set of full conditional
 273 distributions satisfying a form of autoregression given by

$$y_i | \mathbf{Y}_{(i)} \sim N \left(\mathbf{x}_i' \boldsymbol{\beta} + \sum_{j=1, j \neq i}^n c_{ij} (Y_j - \mathbf{x}_j' \boldsymbol{\beta}), \sigma_i^2 \right), \quad i = 1, \dots, n,$$

274 where $\mathbf{Y}_{(i)} = \{Y_j, j \neq i\}$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ are unknown regression parameters. Also,
 275 $\sigma_i^2 > 0$ and c_{ij} are covariance parameters with $c_{ii} = 0$ for all i . It should be noted that
 276 the values of the CAR parameter estimates should reflect reasonable physical mecha-
 277 nisms to guarantee that the patterns observed in the period of record are not just effects
 278 of fluctuations of runoff processes whose dynamics evolve over longer time scales (e.g.,
 279 (Koutsoyiannis 2011; Koutsoyiannis and Montanari 2014)).

280 Banerjee et al. (2014) demonstrate that one can derive the joint distribution based on
 281 full conditional distribution for \mathbf{Y} with Brook's Lemma. The joint distribution of \mathbf{Y} is given

as $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, (\mathbf{I}_n - \mathbf{C})^{-1}\mathbf{M})$, where $\mathbf{M} = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ and the elements of $\mathbf{C} = \{c_{ij}\}$. Note that Brook's Lemma requires $\mathbf{M}^{-1}(\mathbf{I}_n - \mathbf{C})$ to be positive definite and $\mathbf{M}^{-1}\mathbf{C}$ symmetric, which means $c_{ij}\sigma_j^2 = c_{ji}\sigma_i^2$ for $i, j = 1, \dots, n$.

We further simplify this model by assuming $\mathbf{M} = \sigma^2\mathbf{I}_n$, with $\sigma^2 > 0$ and unknown and $\mathbf{C} = \alpha\mathbf{W}$. The parameter α can be interpreted as the unknown “spatial parameter” and $\mathbf{W} = (w_{ij})$ is a known “weight” matrix, which satisfies $w_{ij} = 1$ if and only if sites i and j are neighbors. Oliveira (2010) establishes that setting up the model with these two assumptions automatically satisfies the two aforementioned assumptions (symmetric and positive definite). Hence, the joint distribution of \mathbf{Y} is further reduced to $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{I} - \rho\mathbf{W})^{-1})$. Note that taking advantage of the fact that $\mathbf{I} - \rho\mathbf{W}$ is a sparse matrix can further accelerate the MCMC sampling. See the [Appendix](#) for more on this.

Note that this presentation of the model assumes the random error (or systematic fluctuations in model dynamics; see Clark et al. (2015)) follows a normal distribution, an assumption that should be checked when analyzing a real data set; if this normality assumption is violated, a different error distribution could be specified (e.g., Samadi et al. 2018). In our analysis in “[Model fitting](#)” and “[Model diagnosis](#)” sections, the residuals indicated a heavy-tailed pattern, and we considered a Laplace error specification, but ended up using a t distribution for the error distribution that proved to be proficient for South Carolina's rainfall-runoff processes (see Samadi et al. (2018)). Otherwise, the CAR model outlined in this section was used with our analysis.

Additionally, we define the priors in a relatively non-informative way. Specifically, $\boldsymbol{\beta}_p \sim N_p(\mathbf{0}, 10^6 \cdot \mathbf{I})$, $\rho \sim U(0, 1)$ and $\sigma^2 \sim \text{InvGamma}(0.001, 0.001)$.

Model fitting

Scaling

Scaling is implemented for the gage level measurements since baseline levels vary drastically across locations because they are determined in a fairly arbitrary manner. For instance, Station 02160991, located in the Broad River near Jenkinsville in South Carolina, has an average gage height of more than 200 feet (61 meters), while the Waccamaw River, for example, has a much lower average gage height.

To account for the disparity, we use $y_{ij} - \tilde{y}_i$ as the response variable, where y_{ij} is the original gage height for location i on the j th day, and \tilde{y}_i is the median of location i over 10 years. Figure 4 is a time series plot of the gage heights of five randomly selected locations after scaling.

Autoregressive terms

Autoregressive terms were considered for inclusion in the model with the covariates such as precipitation since it might conceivably take days for precipitation to cause a significant rise in the gage level. The optimal number of terms were determined by inspecting the ACF and PACF of the residuals, along with a comparison of mean squared errors with models with more or fewer autoregressive terms. A first-order autoregressive term was used in the finalized model, and a detailed discussion can be found in “[Model diagnosis](#)” section.

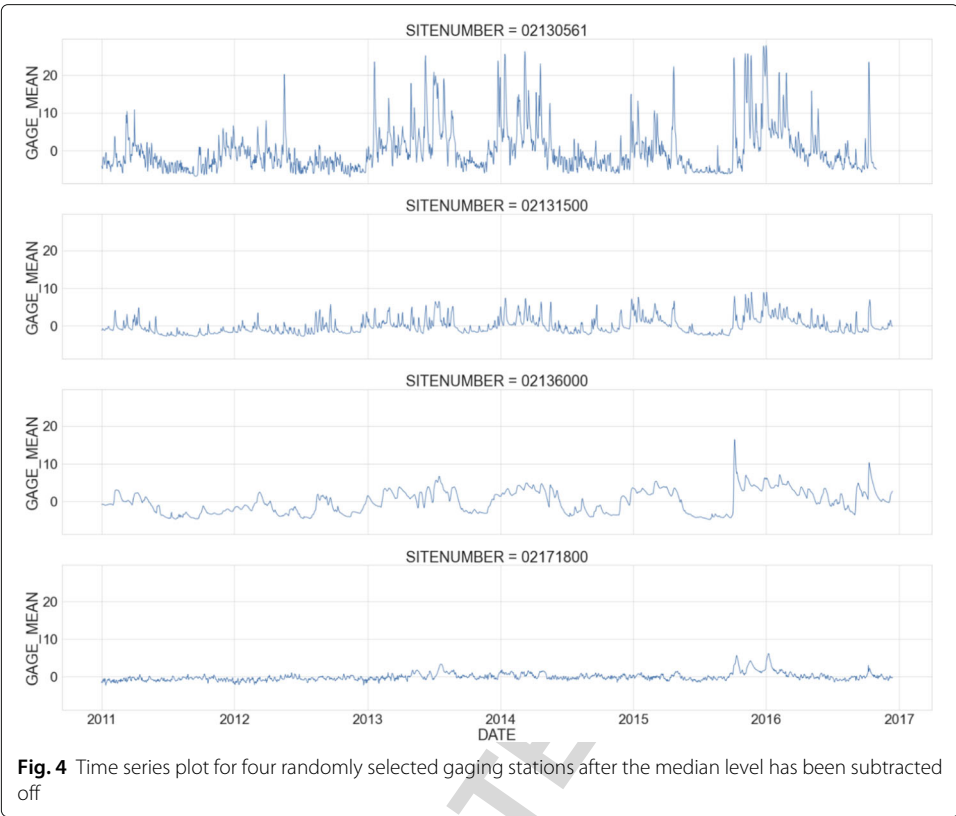


Fig. 4 Time series plot for four randomly selected gaging stations after the median level has been subtracted off

Result

We now present the model fitting results using the CAR model described in “Model description”; recall that we specify a t error distribution to account for the heavy-tailed behavior of the random errors, as explained further in “Model diagnosis” sections. We sample four chains from the posterior distribution of β , τ and ρ , and 95% credible intervals are reported as follows. For each chain we ran 50,000 iterations, and the burn-in period was set to 5,000. We experimented with both the Metropolis sampler and the No-U-Turn sampler and found that both yield similar credible intervals. Winter is used as the baseline season, and a positive estimate for summer, for example, indicates a rise in the predicted gage height compared to winter. Additionally, we found that elevation was not significantly related to the gage level and thus is not included in the final model.

As seen from Table 2, precipitation has a significant effect on the flood level, and a rise of one inch (2.54 centimeter) precipitation leads to an 0.25 inch (0.64 centimeter) increase in the gage measurements on average during the winter season. Among all the seasons, only summer stands out with a statistically significant effect on the gage height. A positive estimate indicates a 0.041 inch (0.10 centimeter) higher predicted gage level for a summer day than a winter day, assuming the days had no precipitation. More importantly, the interaction between the summer season and precipitation has a positive estimate, which indicates that the effect of rainfall on gage levels are different across seasons. Specifically, during summer, rainfall contributes to a larger rise (0.03 inch, or 0.08 centimeter more) in the predicted gage level. In other words, a stronger association between precipitation and flooding can be observed during summer compared to other times of the year. Lastly,

T2

Table 2 Parameter estimates of CAR model

Parameter	Variable	Point Estimate	95% Credible Interval
β_0	Intercept	-0.0376	(-0.5221, 0.5193)
β_1	Precipitation	0.2455	(0.1227, 0.4484)
β_2	Spring	0.005	(-0.0312, 0.0329)
β_3	Summer	0.0413	(0.0018, 0.081)
β_4	Fall	-0.0425	(-0.0793, 0.0018)
β_{12}	Spring * Precipitation	0.0210	(-0.3125, 0.2919)
β_{13}	Summer * Precipitation	0.0445	(0.021, 0.0674)
β_{14}	Fall * Precipitation	-0.0331	(-0.2134, 0.1902)
ρ	Temporal Correlation	0.9007	(0.8788, 0.9826)
α	Spatial Correlation	0.5194	(0.2883, 0.8584)
τ	Spatial Variability	44.5587	(33.7916, 58.3599)

a positive estimate of α suggests that the locations within the same watershed are positively associated, while a positive estimate of ρ indicates that an autoregressive effect is present between different days: For example, a large gage height at a particular location is very likely to be followed by a large gage height measurement the next day at that location. This implies that the relationships between model parameters and covariates can reflect physical mechanisms of runoff generation at a watershed scale. When the model parameters and the covariates have a stochastic pattern/behavior (in time), the model structure reflects more complex nonlinear temporal patterns and relationships between a response variable and the covariates. In this context, spatio-temporal variability of the interface needs to be deduced by meta-data such as effects of water abstraction scheduling, dams' construction and operation, etc. as recently concluded by Serinaldi and Kilsby (2015), and Samadi and Meadows (2017).

Model diagnosis

In this section, we examine the goodness of fit of the CAR model from several perspectives. The autocorrelation function (ACF) and partial autocorrelation (PACF) are employed to examine the residuals from a temporal point of view. Spatially, we display the residuals on the map and check for signs of systematic misprediction in any certain areas.

To examine the residuals from a temporal perspective, the residuals are grouped based on their watershed membership, and averaged over all locations within the watershed. The CAR model was initially fitted without autoregressive terms, and the ACF and PACF of residuals are given in Fig. 5. The slow decay in the ACF plot and the cut-off pattern in the PACF plot suggest an addition of an AR(1) term in the CAR model.

To further evaluate the effectiveness of the autoregressive model, we show a time series plot (Fig. 6, left panel) after averaging out residuals spatially. We also calculate the 2.5% and 97.5% percentiles, and thus the shaded area indicates the range of 95% of all residuals. No apparent autocorrelation pattern is detected, although the last few observations indicate increased volatility in gage height. The absence of an autocorrelation pattern is attributed to the autoregressive term, since a CAR model without the AR(1) term gives the residual time series plot that shows a more obvious autocorrelation pattern and more variability (Fig. 6, right panel).

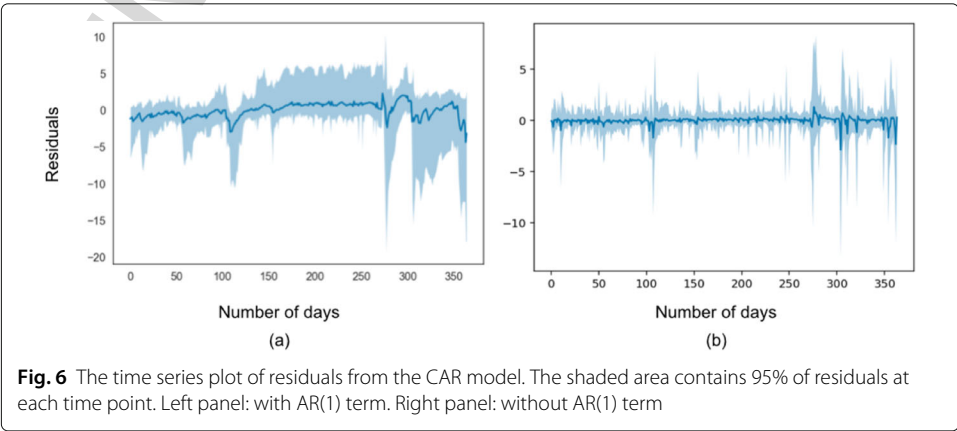
In addition, we examine several time series plots at multiple locations in Fig. 7. We picked four stations of varying degrees of volatility for demonstration purpose, each from

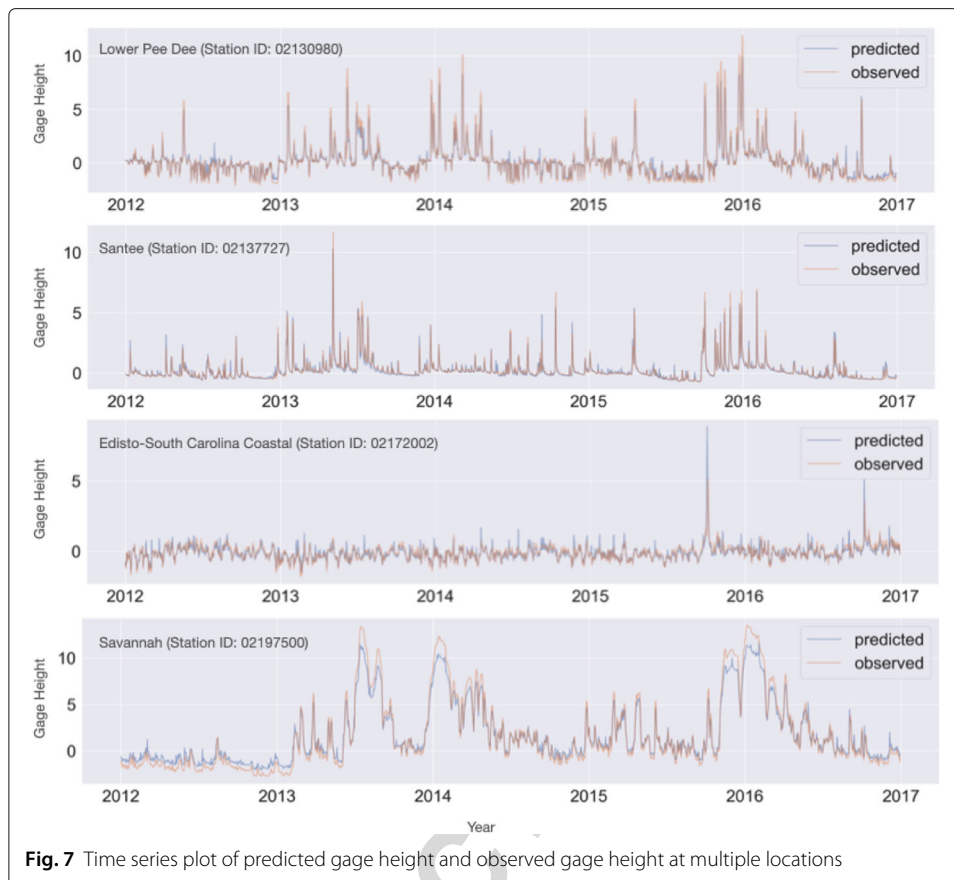


379 a different HUC-6 area. Station 02172002, for example, has rarely seen drastic changes in
380 gage height over the five year between 2012 and 2016. The only exception is during the
381 2015 flood event. In contrast, Station 02197500 demonstrates a different pattern which
382 is more variable and not dominated by the 2015 event. Overall, the autoregressive model
383 consistently captures the trend across different HUC-6 area locations with no systematic
384 overestimation or underestimation. The calculation of mean squared error per site also
385 validates this conclusion.

386 From a spatial perspective, we examine the distribution of residuals by visualizing them
387 on a map with different colors representing negative and positive errors (Fig. 8). The
388 radius of the circle is proportional to the residual. This is a daily snapshot on October 3,
389 2015, from which one can conclude that the residuals are fairly evenly distributed. Sev-
390 eral randomly selected snapshots have been examined during the five-year span and no
391 significant sign of overestimation or underestimation is observed. Additionally, one can

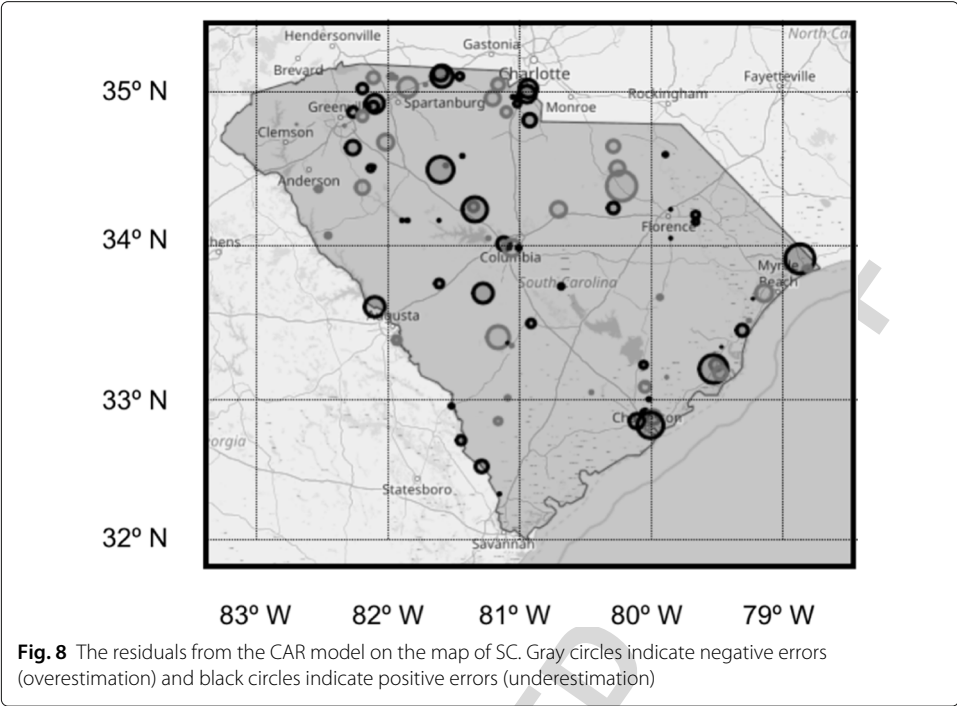
F8





aggregate the residuals over a time period, for instance, a year, and make a yearly residual map for inspection. Such visualization presents a similar picture as Fig. 8 and is thus omitted for the sake of space.

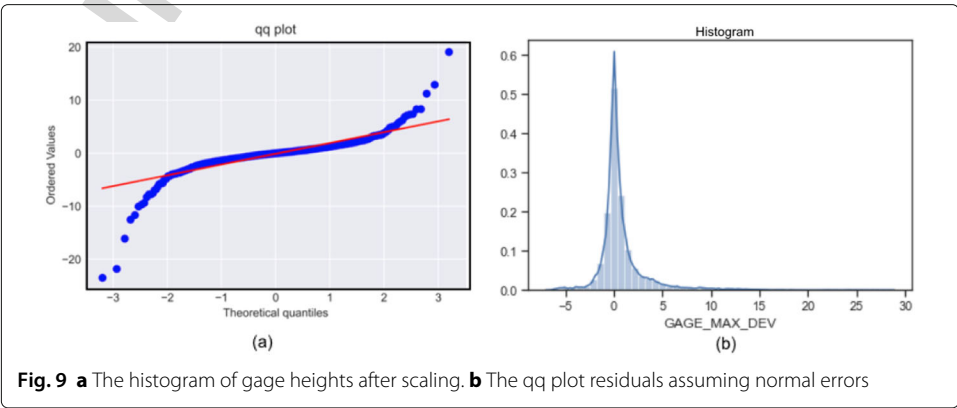
Recall that our fitted model presented in “Model fitting” section used a non-normal error structure; We now explain that choice. If we fit a model with a normal error distribution and examine the qq (quantile-quantile) plot (Fig. 9b), we perceive a pattern that suggests heavy-tailed errors and thus a violation of normality. This is potentially due to the heavy-tailed distribution of maximum gage heights (Fig. 9a). Note that the observations are plotted after the aforementioned scaling operation. The data are also slightly skewed to the right possibly due to occasional thunderstorms, which cause short-term, sharp and severe rises in the gage heights. Flow dynamics in the eastern US are mainly governed by landfall of tropical cyclones and extratropical systems. Tail asymmetry can be partly related to the mixed dynamics forcing floods in different seasons (Smith et al. 2011). Instead of normal errors, using an error structure that follows either a t or Laplace distribution handles extreme rainfall values better. Specifically, we pick a t distribution with 3 degrees of freedom since $\nu = 3$ defines a distribution with reasonably heavy tails and guarantees that both expectation and variance exist. Alternatively, one can also set ν as a hyper-parameter which can be sampled from the posterior distribution. Another possible error distribution could be a skew-t distribution, although for this data set the symmetric heavy-tailed error distributions provide a good fit.



F10 The t distribution where $\nu = 3$ (right panel, Fig. 10) is slightly better in terms of its qq plot than the Laplace (left panel, Fig. 10). Hence, the estimation reported in “Result” section was based on the model assuming that the error term follows a t distribution with 3 degrees of freedom. Note that the parameter estimates would be similar for the two models assuming either of the heavy-tailed distributions (t or Laplace).

Model comparison

It is of interest to evaluate the forecasting capability of the aforementioned CAR model since the gage observations, in and of themselves, are time series data, and forecasting realtime and future flood events might be helpful for early warning systems and emergency management. We compare out-of-sample predictions for the first week of 2017 with the ground truth and calculate the mean squared error and the mean absolute error as metrics since such calculations can be applied to any type of models as long as the



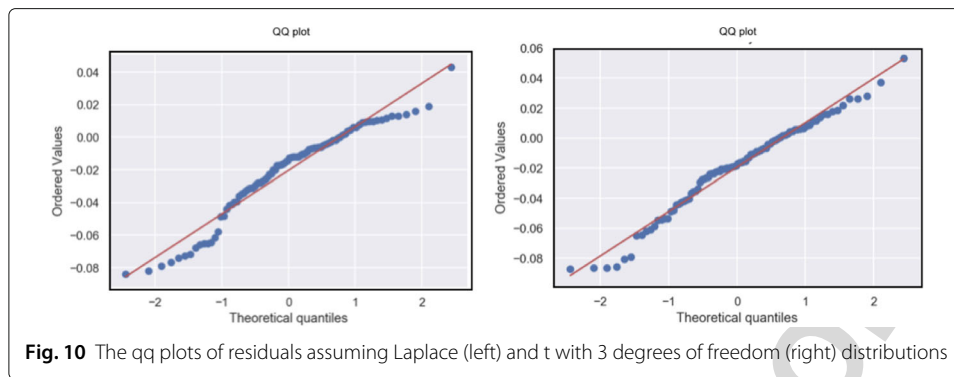


Fig. 10 The qq plots of residuals assuming Laplace (left) and t with 3 degrees of freedom (right) distributions

response variables are continuous. Note that the rainfall and gage height data are measured simultaneously, so that the rainfall at a particular time is used to predict or explain the flooding at the same time. We have found that the time lag between the rainfall cause and the flooding consequence is minimal. However, note that in practice, the use of this model for real-time forecasting of flooding would require some projections of rainfall into the future. The use of the model for retrospective explanatory purposes after the flooding was observed would not require such projections, of course.

In addition, we consider several other models to compare with the CAR model: specifically, the general linear regression model, and two members of a popular family of classification and regression methods: random forest and boosting trees. Comparing the linear model with CAR highlights the necessity of including a proximity matrix since a customized covariance structure is the major difference. Random forest and boosting trees, two popular machine learning algorithms, predict the patterns in data by combining the outputs of individual trees and can give decent benchmarks of model performance. Random forest and boosting trees are both tree-based algorithms and entropy is used as the loss function, but the random forest works in parallel while adaptive boosting works sequentially. Specifically, a random forest obtains results by taking the average of each decision tree prediction, while adaptive boosting builds decision trees iteratively, and the weight of each observation is adjusted until convergence.

The random forest method (Breiman 2001) and boosting trees (Friedman 2001) are chosen as benchmark models since they are routinely used in water sciences. For example, the random forest regression model is demonstrated to be effective in unsaturated hydraulic conductivity and superior to the M5 tree model (Sihag et al. 2019). A random forest approach is also applied in forecasting the spatial distribution of sediment pollution and detecting its predictors (Walsh et al. 2017). Tyrallis et al. (2019) provided a thorough overview of recent papers which involve the study of hydrology and random forest applications, whose themes include rainfall runoff model, streamflow forecasting and drought prediction.

For a fair comparison, all three models include the same seven covariates as the CAR model: precipitation, seasonality variables and all related interaction terms. It is worth noting that the spatial information is handled differently between the CAR model and the other benchmark models. Rather than defining a covariance matrix based on the basin information, we include the water basin indicator as a categorical variable. The mean squared error (MSE) and the mean absolute error (MAE) are reported in Table 3. As seen

t3.1 **Table 3** The comparison of the out-of-sample predictions

t3.2	Model	MAE	MSE
t3.3	CAR model	0.3077	0.2903
t3.4	Linear Model	1.2638	3.0411
t3.5	Random Forest	1.3041	3.4282
t3.6	Boosting Trees	1.5557	4.3659

458 in Table 3, the CAR model outperforms the benchmark models by a considerable margin.
 459 In general, the CAR model is better at exploiting the spatial dependency while the random
 460 forest algorithm makes better use of covariates. Therefore the large difference between
 461 the two models in MAE and MSE can be explained by the small number of covariates.
 462 This notion can be further substantiated by examining the feature importance of the ran-
 463 dom forest and boosting trees (feature importance is measured by the amount of entropy
 464 reduced after a variable is added to the full model), since these two models assign almost
 465 negligible importance to the watershed variables (Table 4). On the other hand, consistent
 466 with the CAR model, the benchmark models such as the random forest recognize pre-
 467 cipitation as a major contributor to the flood height (with a feature importance value of
 468 0.6720 based on random forest, 0.4686 based on boosting trees).

T4

469 To validate the conclusion, we also compute the MSE based on leave-one-out (LOO)
 470 cross-validation based on locations. We use LOO rather than traditional K-fold cross-
 471 validation to avoid the computational burden of partitioning the data (locations) repeat-
 472 edly and fitting the model on every partition. Vehtari et al. (2017) introduced an efficient
 473 computation of LOO from MCMC samples, which uses Pareto-smoothed importance
 474 sampling (PSIS) to provide an estimate of point-wise out-of-sample prediction accuracy.
 475 The CAR model, in this experiment, gives an MSE of 0.32, which outperforms linear
 476 model (1.12), random forest (1.32) and boosting trees (1.66), consistent with outcome of
 477 the one-week forecast task.

478 Discussion

479 We have presented a spatiotemporal model for gage height in South Carolina from 2011
 480 to 2015, a period including one of the most destructive storms in state history. Our model
 481 accounts for the heavy-tailed pattern of the response variable and allows us to determine
 482 several covariates that affect the gage height and to interpret their effects. In particular,
 483 due to the effect of interactions, a stronger association between precipitation and flooding
 484 can be observed during summer compared to other times of the year. If reliable precipi-
 485 tation forecasts are available, our model could be used for forecasting realtime and future
 486 flood events, potentially aiding early warning systems and emergency management.

t4.1 **Table 4** A comparison of feature importances for random forest model and boosting trees

t4.2	Watershed	Random Forest	Boosting Trees
t4.3	Santee	0.0229	0.1751
t4.4	Lower Pee Dee	0.0327	0.0133
t4.5	Savannah	0.0439	0.0245
t4.6	Edisto SC Coastal	0.0042	0.0441

In addition, we developed a Python library to streamline the data preprocessing steps. Data scraping, cleaning, aggregating and transforming steps can be done by simple function calls. We demonstrate several reusable modules we have developed by providing some basic examples in the Github repository of our package. Our hope is that such tools will enable straightforward employment of similar spatio-temporal models for flood data in the future.

Appendix

Sparse matrix

A *sparse* matrix is a matrix where most elements are zero. By contrast, a matrix is considered *dense* if most elements are nonzero. A measure to quantify the sparsity of a matrix is the number of zero-valued elements divided by the total number of elements. As a rule of thumb, a matrix is considered sparse when its sparsity is greater than 0.5. The covariance matrix in the aforementioned CAR model is largely based on our adjacency model, and has a sparsity of 0.86.

Once a sparse matrix is recognized, one can use specialized algorithms and data structures to accelerate computation. This is because memory and computing power are wasted on the zeroes if we employ a standard dense-matrix algorithm. Specifically, a dense matrix is typically stored as a two-dimensional array, and each entry in the array represents an element a_{ij} of the matrix. One can access any element by specifying the row index i and the column index j . In contrast, in a typically sparse matrix representation, only the nonzero entries are stored and thus memory use can be reduced substantially. As a tradeoff, retrieving individual elements becomes more complex in a sparse matrix.

In practice, there are several representations of a sparse matrix. While some types stand out for their efficient modification, such as DOK (Dictionary of Keys) and COO (Coordinate List), others, e.g., Compressed Sparse Row (CSR), support fast matrix operations. CSR suits our needs better since evaluating a multivariate normal distribution involves matrix multiplication, and thus is implemented as part of our model.

The compressed sparse row (CSR) represents a matrix by three one-dimensional arrays: the nonzero values, the row indices, and the column indices. Note that the row indices are not defined in a straightforward manner. An example is given as follows to demonstrate how a CSR representation is implemented.

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

The three vectors to represent the example sparse matrix is
 $A = [5, 8, 3, 6]$ $IA = [0, 0, 2, 3, 4]$ $JA = [0, 1, 2, 1]$.

The array A is the nonzero values, whose column indices are stored in JA . For instance, 3 is in the third column and thus the third element in JA is 2, which stands for the third column since a zero-based index is used. On the other hand, IA contains the row information and is defined recursively, where $IA[0] = 0$ and $IA[i] = IA[i-1] + k$. Note that k is number of nonzero elements on the i th row in the original matrix. According to this definition, the length of IA is $m + 1$ when the matrix has m columns, and the last element in IA is always the number of nonzero values.

The sparse matrix stored in CSR is efficient in matrix-vector multiplication due to the structure of \mathbf{IA} and \mathbf{JA} . For instance, multiplying $[5, 8, 0, 0]$ by another vector, say $[1, 0, 9, 9]$, requires only retrieving nonzero values at location 0 and 1 from the second row. The location information is conveniently stored in \mathbf{JA} , and the length of nonzero values can be found in \mathbf{IA} . Since we only need to compute the dot product of $[5, 8]$ and $[1, 0]$, the computation is reduced by half. In practice, we observe a six-to-ten times boost in sampler performance by switching from a dense matrix implementation, since the adjacency matrix has greater sparsity.

Abbreviations

ACF: Autocorrelation function; AWS: Amazon web services; CAR: Conditional autoregressive (Model); CONUS: Contiguous United States; COO: Coordinate list; CSR: Compressed sparse row; DOK: Dictionary of keys; HUC: Hydrologic unit; MAE: Mean absolute error; MCMC: Markov Chain Monte Carlo; MSE: Mean squared error; NOAA: National oceanic and atmospheric administration; NWS: National weather service; PACF: Partial autocorrelation function; RF: Random forest; SRTM: Shuttle radar topography mission; SST: Sea surface temperature

Acknowledgments

We are grateful for the insightful comments and suggestions of two reviewers, which have greatly improved the article.

Authors' contributions

HL analyzed scraped and prepared the data for this research and coded the model implementation and was a major contributor in writing the manuscript. DH contributed to the methodology for statistical modeling and model validation and was also a contributor in writing. S provided valuable domain knowledge in hydrology modeling, which helped other authors understand the data source, and is also reflected in areas such as variable selection. The author(s) read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author upon request.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Statistics, University of South Carolina, 1523 Greene St 29201 Columbia SC, United States. ²Department of Civil and Environmental Engineering, University of South Carolina, 301 Main St 29208 Columbia SC, United States.

Received: 30 April 2020 Accepted: 12 October 2020

References

- Banerjee, S., Carlin, B. P., Gelfand, A. E.: Hierarchical modeling and analysis for spatial data. CRC Press, Boca Raton (2014)
- Betson, R. P.: What is watershed runoff? *J. Geophys. Res.* **69**(8), 1541–1552 (1964)
- Bonnin, G. M., Martin, D., Lin, B., Parzybok, T., Yekta, M., Riley, D.: Precipitation-frequency atlas of the United States. NOAA Atlas. **14**(2), 1–65 (2006)
- Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
- 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)
- Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., et al.: A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resour. Res.* **51**, 2498–2514 (2015). <https://doi.org/10.1002/2015WR017198>
- Cressie, N.: Statistics for spatial data. John Wiley & Sons, New York (1993)
- Daly, C., Neilson, R. P., Phillips, D. L.: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol.* **33**(2), 140–158 (1994)
- Daly, C., Taylor, G. H., Gibson, W. P., Parzybok, T. W., Johnson, G. L., Pasteris, P. A.: High-quality spatial climate data sets for the United States and beyond. *Transactions of the ASAE*. **43**(6), 1957 (2000)
- Deidda, R.: Rainfall downscaling in a space-time multifractal framework. *Water Resour. Res.* **36**(7), 1779–1794 (2000)
- Delfiner, P., Delhomme, J. P.: Optimum interpolation by kriging. *Ecole Nationale Supérieure des Mines, Paris* (1975)
- Delhomme, J. P.: Kriging in the hydrosociences. *Adv. Water Resour.* **1**(5), 251–266 (1978)
- Diggle, P. J., Tawn, J. A., Moyeed, R. A.: Model-based geostatistics. *J. R. Stat. Soc. Ser. C: Appl. Stat.* **47**(3), 2299–350 (1998)
- 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)
- Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, K., Kobrick, M.: The shuttle radar topography mission. *Rev. Geophys.* **45**(2) (2007)
- Ferraris, L., Gabellani, S., Rebora, N., Provenzale, A.: A comparison of stochastic models for spatial rainfall downscaling. *Water Resour. Res.* **39**(12) (2003)

- Q14 Frederick, R. H., Miller, J. F.: Short duration rainfall frequency relations for California. Third Conference on Hydrometeorology, Bogata, Columbia. Am. Meteorol. Soc., 667–73 (1979) 585
 Friedman, J. H.: Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**(5), 1189–1232 (2001). <https://doi.org/10.1214/aos/1013203451> 586
 Georgakakos, K. P., Kavvas, M. L.: Precipitation analysis, modeling, and prediction in hydrology. *Rev. Geophys.* **25**(2), 163–178 (1987) 587
 Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* **57**(1), 97–109 (1970) 588
 Isaaks, H. E., Srivastava, R. M.: Applied geostatistics. Oxford University Press, New York (1989) 589
 Koutsoyiannis, D.: Hurst-Kolmogorov dynamics and uncertainty. *J. Am. Water Resour. Assoc.* **47**(3), 481–95 (2011). <http://dx.doi.org/10.1111/j.1752-1688.2011.00543.x> 590
 Q15 Koutsoyiannis, D., Montanari, A.: Negligent killing of scientific concepts: the stationarity case. *Hydrol Sci J* (2014). <http://dx.doi.org/10.1080/02626667.2014.959959> 591
 Kumar, P., Foufoula-Georgiou, E.: Characterizing multiscale variability of zero intermittency in spatial rainfall. *J. Appl. Meteorol.* **33**(12), 1516–1525 (1994) 592
 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) 593
 Matheron, G.: Principles of geostatistics. *Econ. Geol.* **58**(8), 1246–1266 (1963) 594
 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) 595
 National Oceanic and Atmosphere Administration, U. S. Department of Commerce: Service assessment: The historic South Carolina floods of October 1–5, 2015 (2015). www.weather.gov/media/publications/assessments/SCFlooding_072216_Signed_Final.pdf. Accessed 4 Dec 2017 596
 Q16 Phillips, R. C., Samadi, S., Meadows, M. E.: How extreme was the October 2015 flood in the Carolinas? An assessment of flood frequency analysis and distribution tails. *J. Hydrol.* **562**, 648–663 (2018). <https://doi.org/10.1016/j.jhydrol.2018.05.035> 597
 Samadi, S. Z., Meadows, M. E.: The transferability of terrestrial water balance components under uncertainty and nonstationarity: A case study of the coastal plain watershed in the southeastern USA. *River Res. Appl.* **33**(5), 796–808 (2017) 598
 Samadi, S., Tufford, D., Carbone, G.: Estimating hydrologic model uncertainty in the presence of complex residual error structures. *Stoch. Env. Res. Risk A.* **32**(5), 1259–1281 (2018). <https://doi.org/10.1007/s00477-017-1489-6> 599
 Q18 Seaber, P. R., Kapinos, F. P., Knapp, G. L.: Hydrologic unit maps (1987). <https://pubs.usgs.gov/wsp/wsp2294/html/pdf.html> 600
 Q19 Serinaldi, F., Kilsby, C. G.: Rainfall extremes: Toward reconciliation after the battle of distributions. *Water Resour. Res.* **50**, 336–352 (2014) 601
 Serinaldi, F., Kilsby, C. G.: Stationarity is undead: Uncertainty dominates the distribution of extremes. *Adv. Water Resour.* **77**, 17–36 (2015). <https://doi.org/10.1016/j.advwatres.2014.12.013> 602
 Sharon, D.: Spatial analysis of rainfall data from dense networks. *Hydrol. Sci. J.* **17**(3), 291–300 (1972) 603
 Q20 Sihag, P., Mohsenzadeh, S., Angelaki, A.: Random forest, MSP and regression analysis to estimate the field unsaturated hydraulic conductivity. *Appl. Water Sci.* **9**(5), 129 (2019). <https://doi.org/10.1007/s13201-019-1007-8> 604
 Smith, J. A., Villarini, G., Baack, M. L.: Mixture Distributions and the Hydroclimatology of Extreme Rainfall and Flooding in the Eastern United States. *J. Hydrometeorol.* **12**, 294–309 (2011). <https://doi.org/10.1175/2010JHM1242.1> 605
 Thiessen, A. H.: Precipitation averages for large areas. *Mon. Weather Rev.* **39**(7), 1082–1084 (1911) 606
 Tyralis, H., Papacharalampous, G., Langousis, A.: A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water.* **11**(5), 910 (2019) 607
 Q21 Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**(5), 1413–1432 (2017) 608
 Villarini, G., Smith, J. A., Serinaldi, F., Bales, J., Bates, P. D., Krajewski, W. F.: Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. *Adv. Water Resour.* **32**, 1255–1266 (2009) 609
 Walsh, E. S., Kreakie, B. J., Cantwell, M. G., Nacci, D.: A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system. *PloS ONE.* **12**(7), e0179473 (2017) 610
 Q22 Wickham, H.: ggplot2: Elegant graphics for data analysis. Springer, New York (2016) 611

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

636

Author Query Form

Journal: Journal of Statistical Distributions and Applications



Article: Spatio-temporal analysis of flood data from South Carolina









Dear Author,












During the copyediting of your paper, the following queries arose. Please respond to these by annotating your proofs with the necessary changes/additions.

- If you intend to annotate your proof electronically, please refer to the E-annotation guidelines.
- If you intend to annotate your proof by means of hard-copy mark-up, please refer to the proof mark-up symbols guidelines. If manually writing corrections on your proof and returning it by fax, do not write too close to the edge of the paper. Please remember that illegible mark-ups may delay publication.

Whether you opt for hard-copy or electronic annotation of your proofs, we recommend that you provide additional clarification of answers to queries by entering your answers on the query sheet, in addition to the text mark-up.

Query No.	Query	Remark
Q1	Author names: Please confirm if the author names are presented accurately, and in the correct sequence (given names/initials, family name). Author 1 Given Name: Haigang Last Name: Liu Author 2 Given Name: David B. Last Name: Hitchcock Author 3 Given Name: S. Zahra Last Name: Samadi	
Q2	Reference: References [Berne et al. (2009), Serinaldi and Kilsby (2014), Troutman (1983), Tabios and Salas (1985), (USGS, 2011), Oliveira (2010)] were mentioned in the manuscript; however, these were not included in the reference list. As a rule, all mentioned references should be present in the reference list. Please provide the reference details to be inserted in the reference list.	

Query No.	Query	Remark
Q3	Please check section citations if cited correctly. Otherwise, kindly advise us on how to proceed.	
Q4	Figures 1–10 contains poor quality and small text inside the artwork. Please do not re-use the file that we have rejected or attempt to increase its resolution and re-save. It is originally poor, therefore, increasing the resolution will not solve the quality problem. We suggest that you provide us the original format. We prefer replacement figures containing vector/editable objects rather than embedded images. Preferred file formats are eps, ai, tiff and pdf.	
Q5	Figure/Table: Please check and confirm if all captions and citations of Figures and Tables have been captured correctly.	
Q6	Figures: Figure “3” e-file was received; however, no citation was provided in the manuscript. Please provide the location of where to insert the citation in the main body of the text. Please note that the order of main figure citations in the text should be sequential to comply with the journal’s standard, e.g., other figures may be cited before Fig. 1, but their main citation should come after Fig. 1 (i.e., Figs. 3, 2, 1, 2, 3, 4, 5, 6, 7).	
Q7	Authors’ contributions: As per standard instruction, the statement “The author(s) read and approved the final manuscript.” is required in the “Authors’ contributions” section. Please note that this was inserted at the end of the paragraph of the said section. Please check if appropriate.	
Q8	Please check all affiliations if captured and presented correctly. Otherwise, please amend if necessary.	
Q9	Affiliations: As per standard instruction, country is required for affiliations; however, this information is missing in affiliations “1 and 2”. Please check if the provided country is correct and amend if necessary.	
Q10	Please check the added authors for Reference [Clark et al. (2015), Villarini et al. (2009)] if captured correctly.	

Query No.	Query	Remark
Q11	Daly et al., 2001 in the manuscript was modified to Daly et al., 2000. Please check if captured correctly. Otherwise, please amend if necessary.	
Q12	Delfmer and Delhomme (1975) in the manuscript was modified to Delfiner and Delhomme (1975). Please check if captured correctly. Otherwise, please amend if necessary.	
Q13	Please provide page range for reference [Farr et al. (2007), Ferraris et al. (2003)].	
Q14	Please provide volume number for Reference [Frederick et al. (1979)]	
Q15	Please provide page range and volume number for Reference [Koutsoyiannis et al. (2014)].	
Q16	Please check Reference [Phillips et al. (2018)] if captured and presented correctly.	
Q17	Philips et al., 2018 in the manuscript was modified to Phillips et al. 2018. Please check if captured correctly. Otherwise, please amend if necessary.	
Q18	Please provide access dates for Reference [Seaber et al. (1987)].	
Q19	Reference: References [Matheron (1963), Serinaldi and Kilsby (2014)] were provided in the reference list; however, these were not mentioned or cited in the manuscript. As a rule, all the references given in the list of references should be cited in the body of a text. Please provide the location of where to insert the reference citation in the main body text.	
Q20	Sihag et al. 2015 in the manuscript was modified to Sihag et al. 2019. Please check if captured correctly. Otherwise, please amend if necessary.	
Q21	Vehtari et al. 2016 in the manuscript was modified to Vehtari et al. 2017. Please check if captured correctly. Otherwise, please amend if necessary.	
Q22	Wickham 2019 in the manuscript was modified to Wickham 2016. Please check if captured correctly. Otherwise, please amend if necessary.	