CAS RESEARCH PAPERS

CLIMATE, SPATIAL DEPENDENCE AND FLOOD RISK: A U.S. CASE STUDY

By Robert J. Erhardt Ph.D., ACAS,
Mathieu Boudreault, Ph.D., FSA, FCIA,
David A. Carozza, Ph.D. and Kejia Yu, MA

CASUALTY ACTUARIAL SOCIETY

Climate, Spatial Dependence, and Flood Risk: A U.S. Case Study

Robert J. Erhardt Ph.D., ACAS,
Mathieu Boudreault, Ph.D., FSA, FCIA, David A. Carozza, Ph.D., and
Kejia Yu, MA

Abstract: Flood represents one of the costliest and most disruptive natural disasters in the United States, and the economic losses from flooding are trending upward. While this trend is known to be driven primarily by an increasing population and wealth exposure, climate change is also affecting flood risk in more subtle ways. We merge data on economic flood losses, historical climate, census population, and geological characteristics to explore drivers of flood losses and climate trends. Our data cover 292 watersheds spanning the continental United States, over the period 1979–2018. We fit a Bayesian spatial mixed-effects model for flood loss frequency and a Bayesian mixed-effects model for flood severity loss per person. Both models control for measured covariates, contain random effects to capture variation from unmeasured covariates, and quantify climate drivers of flood risk. We show empirically that flood losses exhibit spatial dependence that requires spatial statistical models; climate variables are partial drivers of increased frequency and severity; and measures of spatial dependence have been changing over time. And, through a simulation study, we lay a groundwork to disentangle climate and nonclimate drivers of these changing measures of spatial dependence.

Keywords: climate model, global climate change, extreme events, flood risk, spatial dependencies

1 Introduction

Floods represent a major hazard among natural disasters. In 2020, floods resulted in estimated losses of US$76 billion globally (2020 costs). They have dominated the number of presidential disaster declarations, representing almost two-thirds of all declared disasters (Michel-Kerjan, 2010). Floods correctly attract the attention of both private and public insurers as well as risk managers, who seek to model the risk to inform cost-sharing mechanisms, resiliency, and adaptation.

Given the large role floods play in disaster risk management, there is obvious concern that climate change may exacerbate those risks. In 2012 the Intergovernmental Panel on Climate Change (IPCC) published its special report “Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation" (Field at al., 2012). The aims were to aggregate scientific studies specifically linking climate change to environmental extremes and to describe the current state of the scientific literature. Flood risk was one set of extreme events considered. In the Summary for Policymakers, the authors distilled a few messages concerning the risk of future floods and climate change. Somewhat surprisingly, linking climate change to rising flood risk is not as developed or certain as linking climate change to other environmental risks, such as heat waves.

Concerning direct climate impacts on either the frequency or severity of floods, the authors write (Field at al. (2012), page 8):

> There is limited to medium evidence available to assess climate-driven observed changes in the magnitude and frequency of floods at regional scales. . . . Furthermore, there is low agreement in this evidence, and thus overall low confidence at the global scale regarding even the sign of these changes.

In the same paper, they later add (Field at al. (2012), page 13):

> Projected precipitation and temperature changes imply possible changes in floods, although overall there is low confidence in projections of changes in fluvial floods. . . . There is medium confidence (based on physical reasoning) that projected increases in heavy rainfall would contribute to increases in local flooding in some catchments or regions.

In short, merely adjusting catastrophe flood models with increased frequency or severity of physical floods to account for climate change may not be on solid scientific footing at this time.

While it may be unclear if the *physics* of flooding is changing from climate change, there is a clear case that the *economic costs* of flooding are trending up. This is part of a wider pattern of increasing losses from natural disasters, despite a less obvious link between climate and the frequency or severity of the natural hazard itself. A primary driver of this trend is shifting exposures in higher-risk areas, rather than anthropogenic climate change itself (Miller et al., 2008; Field et al., 2012; Bouwer, 2011). However, the shift in exposures is not always an exhaustive explanation. For example Dottori et al. (2018) show a link between anthropogenic warming and increased flood losses globally, though with very large variation by region. Field et al. (2012) write (P. 16):

> Although future flood losses in many locations will increase in the absence of additional protection measures (high agreement, medium evidence), the size of the estimated change is highly variable, depending on location, climate scenarios used, and methods used to assess impacts on river flow and flood occurrence.

Within the United States we see the same pattern holding, with little evidence for climate-driven changes in the physics of flooding (Villarini et al., 2009) but some studies linking climate change to the economic flood losses (Choi and Fisher, 2003). What is clear from the literature is that any studies on the link between climate and economic flood losses must be regional, sensitive to plausible climate scenarios and statistical methodology, and mindful that the drivers of increased losses are more complex than a simple shift in frequency or severity of flood itself.

Some studies have linked together the causal chain from large-scale climate to localized losses through intermediaries such as regional climate, hydrology, and hydraulics (Boudreault et al., 2019). Such a causal chain is logical and serves as a proof of concept that climate inputs can relate to flood risk outputs. However, each link requires a plausible scientific model, necessitating expertise across a range of disciplines. Furthermore, it is

not always obvious how to propagate all sources of uncertainty down the causal chain of events. Focusing solely on the statistical links between climate inputs and economic loss outputs avoids the need to specify each link and propagate uncertainty at each level. Furthermore, if the goal is to inform risk management, avoiding the physical intermediaries of flooding in economic loss studies reduces the known barrier that risk managers and actuaries have limited scientific expertise on floods (Furman et al., 2019). Carozza and Boudreault (2021) have a follow-up study that skips the intermediate steps in favor of a random forest that links climate inputs to flood loss outputs in a statistical framework. From statistical, climate, and risk-management standpoints, there is a benefit to modeling economic losses directly.

When building a statistical model, it is essential to recognize that flood data exhibit spatial dependence. Neighboring locations experience similar large-scale atmospheric processes such as precipitation and temperature, are subject to similar localized weather extremes, are affected by one another's land surface processes, and may show similar geological and economic development profiles. This results in observable spatial dependence in flood risk. The aptly titled "Biases in National and Continental Flood Risk Assessments by Ignoring Spatial Dependence" (Nguyen et al., 2020) captures the spirit of the issue—one cannot hope to estimate flood risk accurately without addressing spatial dependence. Despite this, the authors add "There are very few studies which have discussed the effect of spatial dependence on flood risk estimates" noting exceptions of Lamb et al. (2010), Wyncoll and Gouldby (2015), and Metin et al. (2020).

In Nguyen et al. (2020), time series of annual maximum stream flows at 379 gauges in Europe were modeled and then stitched together using a copula, a method for linking marginal distributions to a joint multivariate distribution by imposing a prespecified dependence structure. The fitted model generated synthetic time series of stream flows, which were then fed into models for inundation area and in turn economic damages. The authors showed a risk of either overestimation or underestimation of total economic damages from varying return periods. Quinn et al. (2019) similarly fit marginal distributions on flow exceedance probabilities at 2,400 gauges in the United States, and then tied those marginal distributions using pairwise regressions to capture spatial dependence. Their statistical model simulated 1,000 events, which were in turn fed into a hydrodynamic model for flood depth, whose output was finally fed into depth-damage curves to estimate flood losses. While these studies incorporated spatial dependence, from a statistical point of view there are three drawbacks worth stating.

First, a suitable joint statistical model has the power to *explain* sources of dependence among variables, and parameter estimates from such a model give scientific insight into the process generating the data. Copulas can be used to model dependence after the fact, but they do not *explain* the source of dependence, and from that perspective they can be less appealing than fitting a suitable dependence model.

Second, both studies relied on fitting a large number of marginal distributions at each study location in a first statistical estimation step, and then tying these marginal distributions together with a dependence model in a second estimation step. This approach is pragmatic and can be necessary in high-dimensional data settings with complicated dependence structures, but is also less efficient from a statistical point of

view. Rather than utilize the full data set to simultaneously estimate all parameters in a single step, two-step procedures use portions of the data set to estimate parameters for each marginal distribution separately, and then assume these estimated marginal parameters are fixed and correct when trying to estimate dependence parameters in a second step. This lowers the estimation efficiency, and may miss important covariances between parameters from the first and second steps. In addition, the number of pairwise copula parameters can grow very rapidly with the number of marginal locations, and this can become computationally infeasible even in a two-step procedure.

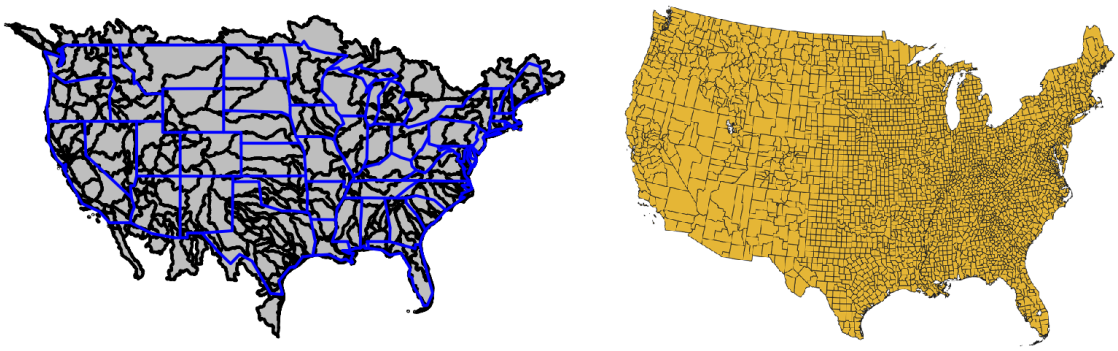
Third, both approaches rely on a causal chain—stream flow outputs feed into an inundation area model, whose outputs are then fed into a damage model. However, each step in the causal chain has associated uncertainties, which are not easy to explicitly capture; nor are they easy to propagate down through the chain. The final economic loss estimates have an implied wide uncertainty bound, which is the net result of parameter uncertainty in the statistical model, stream flow uncertainty from simulations of the model, inundation area uncertainty given stream flow inputs, and then economic loss uncertainty given inundation area. Neither research study makes an explicit attempt to estimate or discuss this uncertainty.

The goal in this study is to estimate the spatial dependence and climate change impacts on flood losses, while incorporating all sources of uncertainty into the final estimate. We define the spatial exposure units based on watersheds drawn from hydrology, and define floods based on economic losses attributable to flooding. We control for a number of climate and geographic features at the watershed level, and train a Bayesian spatial statistical model using 40 years of data in the United States. This watershed-level statistical model explicitly captures spatial dependence, avoids the need to define spatial dependence through copulas, and contains random effects to absorb variability due to unmeasured confounders. Furthermore, our statistical model allows us to track how climate variables are drivers of flood losses and spatial dependence. Simulations from the fitted model under different climate scenarios enable one to study how further shifts in climate may affect flood losses, and since the model is Bayesian all sources of uncertainty are quantified through the posterior and very naturally flow into simulations. Focusing on economic losses rather than a physical definition of flooding keeps attention on the stronger trend in economic losses, avoids the need to fully specify the causal chain of scientific processes spanning from climate to loss, and avoids the need to propagate uncertainty down through this chain of events. We show that the proportion of watersheds experiencing flood risk has increased over the period 1979–2018. We further show that the spatial dependence in flood loss frequency has changed over this same period, and our simulation study shows that climate has been a driver of both.

2 Data

2.1 HydroSHEDS

Figure 1. Left panel: The level 5 watersheds are shown in black, and overlaid in blue are the U.S. state boundaries for reference. The 292 watersheds that intersect any county in the lower 48 states form the spatial exposure unit in this research study. Right panel: The counties of the contiguous lower 48 U.S. states. Flood losses are summarized at the county level in SHELDUS, and attributed to level 5 watersheds by area.



The basic spatial exposure unit in this study is the watershed. We define these using the HydroSHEDS data set, a geometric data set that describes watershed delineation on a global scale. Watersheds are spatial regions divided by ridges or boundaries that separate out the direction of water flows. The major paper describing the data set is Lehner and Grill (2013); the data set can be freely downloaded at https://www.hydrosheds.org/downloads.

The basic principle of the HydroSHEDS data set is that watersheds can be described in a series of nested polygons, spanning from a coarse continental resolution to a very fine local resolution. Level 1 watersheds describe the continental scale, in which there are seven watersheds, one for each continent. At each successive level, each watershed is subdivided into nested regions according to Pfafstetter coding (Verdin and Verdin, 1999). There are 12 total levels in HydroSHEDS. A finer watershed scale allows for more nuanced statistical modeling; on the other hand, such a scale has at least two drawbacks. First, the smaller area makes data collection at the watershed level less reliable. As an example, many climate variables are available at lower spatial resolutions than the smaller watersheds, which makes computation of climate characteristics unreliable at a small scale. Second, smaller-sized watersheds dramatically increase the number of watersheds and therefore computational cost of fitting the spatial model. Since there is a natural trade-off between spatial precision on the one hand and computational cost and data quality on the other, for this analysis we chose level 5 in HydroSHEDS. Figure 1 shows the resulting 292 watersheds that intersect the lower 48 contiguous U.S. states. Those 292 watersheds form the basic spatial exposure units in our study.

2.2 HydroATLAS

HydroATLAS (linke et al., 2019) is a comprehensive data set that describes attributes in each watershed, such as soil type, bedrock type, groundwater depth, and so forth. Attributes divide into two parts: BasinATLAS, which describes the hydrographic features in each polygon, and RiverATLAS, which describes the river network as lines. We chose the attributes corresponding to BasinATLAS because the polygon shapes are the fundamental exposure units. Table 1 shows 19 attributes that are considered potential drivers of flood frequency or severity. Those 19 attributes further subdivide into 41 individual measures according to the spatial extent and summary statistic of each attribute.

These 41 measures form part of the set of covariates in our study. Each attribute listed in Table 1 comes from one of five classes:

- hydrographic (H), referring to the physical features of bodies of water;
- physiographic (P), referring to physical features of the earth's surface;
- land cover (L), which describes the land cover and land usage;
- soils and geology (S), describing the subsurface soil composition; and
- anthropogenic (A), referring to human-induced characteristics.

These measurements are drawn from more than 68 scientific studies and offer consistent baseline data without the need of repetition. It is important to realize that most of the measurements do not vary over time. We made no attempt to second-guess with regard to the inclusion/exclusion of any feature, and relied on this established scientific data source. To acknowledge the limitations of these data and the possibility of further unmeasured confounders, our statistical models include watershed-level random intercepts, which are specifically designed to capture variation from unmeasured confounders at the watershed level. Readers who are curious about how our study results would hold up with an expansion of the covariate set should consider how the random effects capture variability for the unmeasured variable in question.

2.3 SHELDUS

The Spatial Hazard Events and Losses Database for the United States (SHELDUS, Center for Emergency Management & Homeland Security (2020)), or SHELDUS , is a fee-based subscription data set that records hazardous events and losses from natural causes in the United States. Arizona State University's Center for Emergency Management and Homeland Security manages the database. The full SHELDUS database records the property loss, crop loss, and number of injuries and fatalities, each of which is broken out by 18 hazard types (flood, hurricane, wildfire, etc.) and 139 peril types (flood-coastal, flood-riverine, flood-tidal, etc.). For more detail on how and when different losses "count" see Gall et al. (2009). SHELDUS data are available aggregated monthly by U.S. county, from 1960 to present. In this research study, we consider losses only from 1979 to the

present, and we consider only property damages from flood, which includes crop losses from flood.

Table 1. Nineteen HydroATLAS attributes that may contribute to flood loss frequency or severity. All are measured and recorded at the level 5 watershed scale, meaning we have values for every one of the polygons shown in Figure 1. Letters in parentheses after each attribute signify as follows: (H) hydrographic, (P) physiography, (L) landcover, (S) soils and geology, and (A) anthropogenic.

Attribute	Unit	Description
Natural Discharge (H)	m^3/s	Water flow rate
Land Surface Runoff (H)	mm	Water, rain, melted snow, or other source that flows over the land surface
Inundation Extent (H)	%	The scope of the area in the watershed impacted by flood
Limnicity (H)	%	Percentage of lake area in watershed
Lake Volume (H)	$10^6 m^3$	Volume of lake in watershed
Reservoir Volume (H)	$10^6 m^3$	The reservoir's capacity of water in the watershed
Degree of Regulation (H)	%	Shows the influence of dams on the natural flow regime of downstream river reaches
River Area (H)	hm^2	Represents the lands that contain both aquatic and riparian habitats
Groundwater Depth (H)	cm	The depth of water present beneath the earth's surface in soil
Elevation (P)	m	The height to which something is elevated
Terrain Slope (P)	◦	The terrain layer's slope value calculated from elevation data
Stream Gradient (P)	dm/km	The grade measured by the ratio of drop in elevation of a stream per unit of horizontal distance
Land Cover Classes (L)	classes	The classification of the observed physical cover on the earth's surface
Clay Fraction (S)	%	The percentage of clay in soil
Silt Fraction (S)	%	The percentage of silt in soil
Sand Fraction (S)	%	The percentage of sand in soil
Organic Carbon (S)	%	The percentage of organic carbon in soil
Population Density (A)	people/km^2	The number of people per unit of area
Urban Extent (A)	%	The percentage of area in which settlement points are greater than 5,000 persons

Table 2. Example of aggregated SHELDUS data without geometric information (SC data are freely available). FIPS is a unique county identifier. When combined with a county-level shapefile, it can allocate total losses by area to each intersecting watershed.

State Name	County Name	County FIPS	Year	Month	PropertyDmg
South Carolina	Abbeville	45001	1961	2	142.86
South Carolina	Abbeville	45001	1963	8	500.00
South Carolina	Abbeville	45001	1964	3	108.70
South Carolina	Abbeville	45001	1964	8	1086.96
South Carolina	Abbeville	45001	1964	9	108.70
South Carolina	Abbeville	45001	1964	10	10869.57

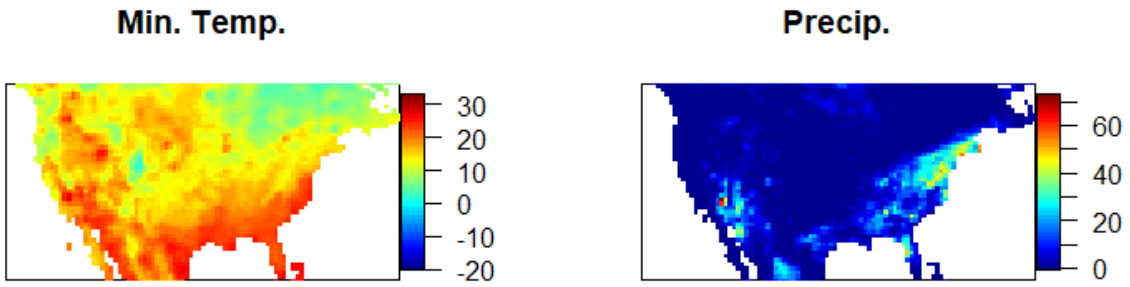
Table 2 shows six observations from SHELDUS. It contains time variables, such as year and month, geometric variables, such as state and county, and the FIPS code, a unique identifier for each county in the United States. From Figure 1, we see that flood losses summarized at the county level are sufficient for our analysis, as there are 3,119 counties in the lower contiguous United States, which ensures that each of the 292 watersheds contains 10+ counties on average. We obtained annual total inflation-adjusted property damage from the flood hazard, by U.S. county from 1979 through 1998 (40 years). These losses include crop losses. This period was selected to match the NOAA CPC time period, described in the next section.

2.4 NOAA CPC

The National Oceanic and Atmospheric Administration Climate Prediction Center (NOAA CPC) constructs and maintains a number of gridded data products that capture historical weather over the recent past. Given the obvious connection between flood losses and weather, we obtained a number of data products for use in our model. Specifically, we obtained CPC Global Temperature and CPC Global Precipitation.

CPC climate data are geometric rasters, with three dimensions of longitude, latitude, and time. The spatial resolution is 0.5 degree, and the temporal resolution is daily, so each annual data set is 720 by 360 by 365. Only cells over land have recorded data. We collected the daily maximum surface temperature and the daily minimum surface temperature (in degrees Celsius) and the accumulated daily precipitation (in millimeters) from 1979 to present. These data are freely available at https://psl.noaa.gov/data/gridded/index.html. We will use daily accumulated precipitation to compute annual maximum daily and annual maximum weekly precipitation to focus on extremes. Figure 2 shows example data from August 11, 1979.

Figure 2. One layer (i.e., one day) of CPC data containing the minimum temperature (left panel) and precipitation (right panel) on August 11, 1979. These CPC data products are raster data, with 0.5-degree-by-0.5-degree grid cells arranged in rows and columns.



2.5 U.S. Census

We obtained estimated U.S. population by county by year, from 1979 to 2018, from the U.S. Census Bureau. Data are available at https://seer.cancer.gov/popdata/download.html. The population data are updated every five years, and we interpolated linearly between update years.

2.6 Data Preprocessing

The fundamental exposure unit is the watershed-year, of which there are 11,680 (292 watersheds times 40 years). These are rows in our final data set. For each watershed-year, we attached

- total property losses from flood, obtained from SHELDUS;
- the HydroATLAS watershed feature data shown in Table 1;
- summarized climate data drawn from CPC; and
- total population within this watershed-year, drawn from the U.S. Census Bureau.

SHELDUS losses and population are at the county level. To attribute these county-level measurements to the level 5 watershed exposure units, we used a shapefile on U.S. counties (https://cemhs.asu.edu/sites/default/files/allcntys2010.zip). We intersected the county shapefile with watersheds, and attributed the variable to each watershed by area—if county A had 70% of its area in watershed 1 and 30% in watershed 2, we attributed 70% and 30% of its total flood property damages in a given year to each watershed, respectively. We did likewise for U.S. Census population.

For the CPC climate data, we intersected the watershed (Figure 1) with the centroids of each CPC grid cell (Figure 2) to identify climate grid cells falling within each watershed. Restricted to only these grid cells, we defined a set of annual variables: the annual maximum daily precipitation value pr; the annual maximum weekly precipitation value pr.week; the annual maximum daily temperature tasmax; and the annual minimum daily temperature tasmin. We deliberately included both short-duration and longer-duration

measures of extreme precipitation, since intense short-duration rainfall is often the driver of floods in small watersheds, while longer-duration precipitation often causes riverine floods in larger watersheds (Kundzeicz et al., 2014).

The result is a database on $292 \cdot 40 = 11{,}680$ watershed-year exposure units, for which we have the total inflation-adjusted property damage caused by flood, climate variables, watershed features, and population. To help the reader visualize this, Table 3 shows a sample of the combined database.

Table 3. Combined raw data at the watershed-year exposure level. Each row corresponds to a unique watershed-year. Columns include the total annual flood loss from SHELDUS, the CPC climate variables, the HydroATLAS watershed features, and U.S. Census Bureau population estimates.

Watershed	yr.in	Loss	Climate Variables				Watershed Features		Population
			pr	pr.wk	tas.max	tas.min	dis_m3_pyr	...	
7050007830	1979	0.00	38.53	67.96	25.65	10.42	87.18	...	41,523.74
7050008340	1979	0.00	33.91	87.96	28.52	12.66	28.71	...	4,442.38
7050008450	1979	3656.19	44.12	55.18	26.16	14.90	12.49	...	145,776.86
...	...	...	...	...	...	...	...	...	...

3 Exploratory Analysis of the Combined Data

3.1 Empirical Changes in Climate

Figure 3 shows the shifts in the four climate variables over the 40-year period from 1979 through 2018. The histograms are computed over the 292 watersheds, and for each watershed we computed the mean of the climate variable over 1999–2018 and subtracted the mean computed over 1979–1998. Solid lines at zero help indicate how many watersheds trended up versus down. The vertical dashed lines show mean change weighting each watershed equally. The same information is shown spatially in Figure 4 and Figure 5.

Figure 3. Differences in the means for each of the four climate variables, computed over 1999–2018 as compared to 1979–1998 for all watersheds. Solid vertical lines at zero show positive versus negative shifts. Dashed lines show the mean change taken over the 292 watersheds.

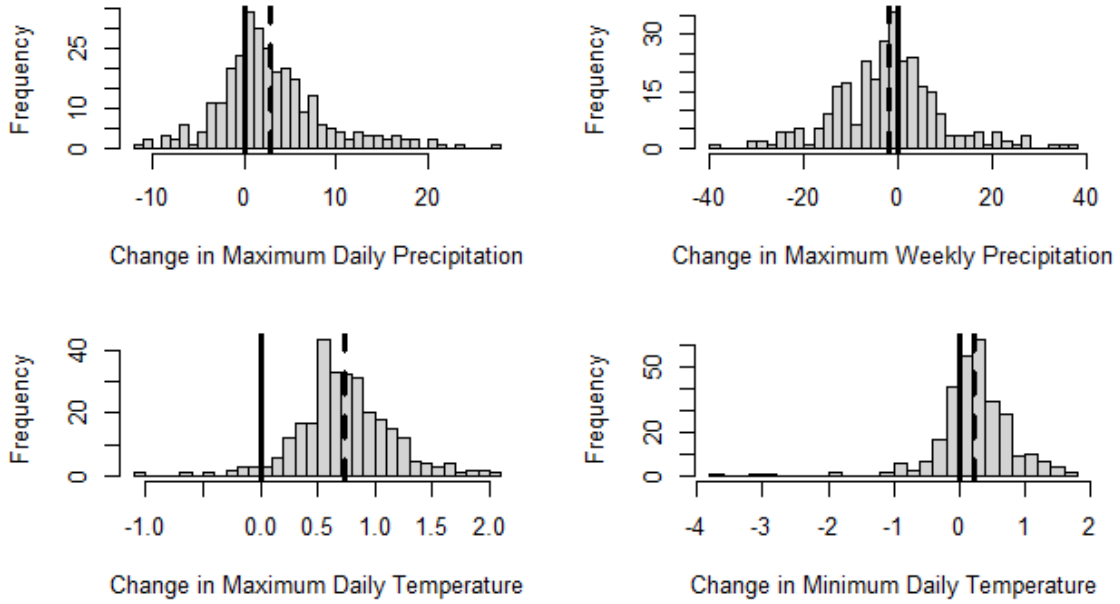


Figure 4. Changes in mean maximum daily and weekly precipitation, from the period 1979-1908 to 1999-2018.

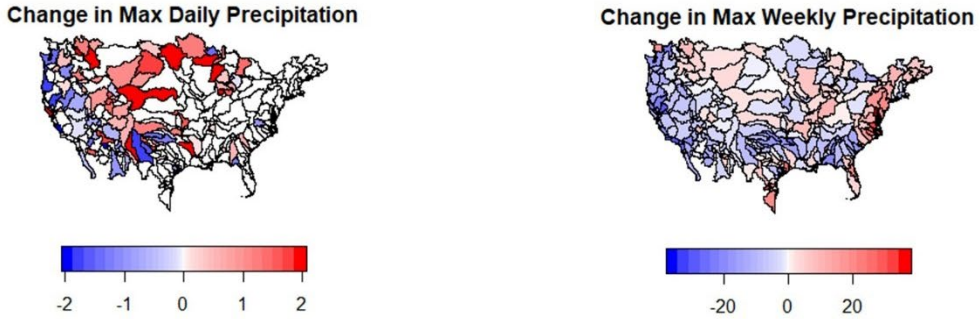
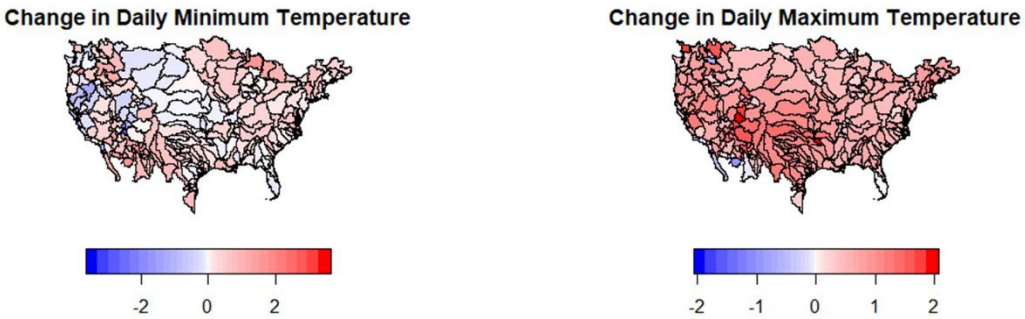


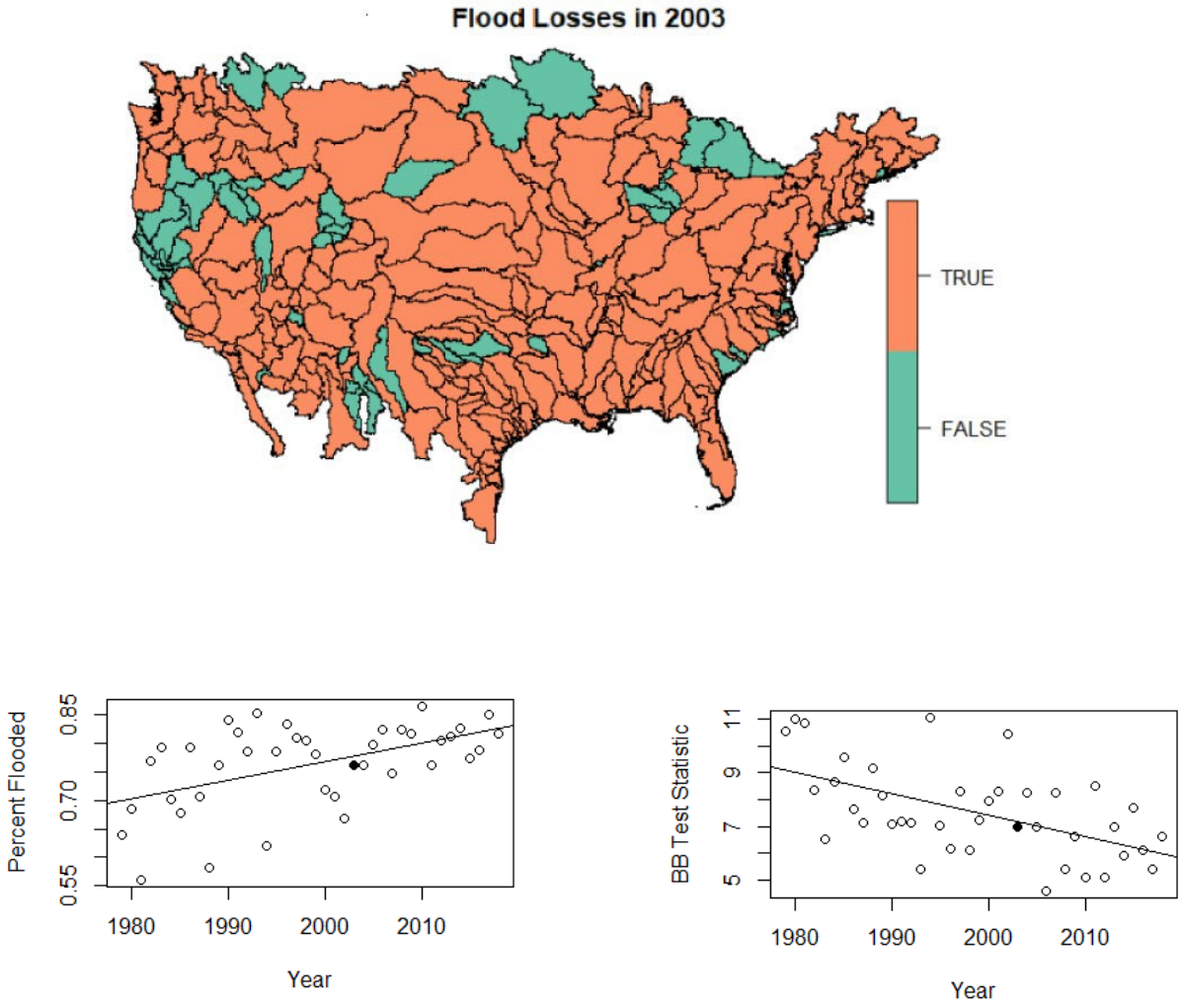
Figure 5. Changes in mean daily maximum and minimum temperatures, from the period 1979–1998 to 1999–2018.



3.2 Exploratory Join Count Analysis

One interest of this study is the spatial pattern of flooding. Specifically, we are interested in whether that pattern exhibits autocorrelation with flooded regions occurring in spatially clustered regions; we are also interested in whether the pattern is shifting over time. Prior to modeling the data, we explore the response variable flood loss occurrence only. A flood loss occurrence is defined as a positive loss-per-person amount in a given watershed and year. Figure 6 shows these data. The bottom left panel of Figure 6 shows the realized flood loss watersheds from the year 2003, when 76% of watersheds experienced a flood loss. Upon visual inspection, it is clear that the data exhibit strong clustering, with flooded and nonflooded watersheds clustering nearby one another. The bottom right panel of Figure 6 shows that the proportion of watersheds experiencing flood losses had increased over the 1979–2018 period, with the solid circle showing data from 2003. We wish to know whether the spatial pattern of flood losses is changing over time after controlling for the observed increase in flood proportion.

Figure 6. Top: Flood losses (red) by watershed in the year 2003. In this year, 76% of watersheds experienced a flood, and the join count BB test statistic is 5.43. Bottom left: the percentage of watersheds flooded over time. Bottom right: the BB statistic for each year computed from Equation 1. In both lower figures, the solid circle refers to 2003 shown at the top.



To answer this, we test for spatial autocorrelation in flood occurrence using the join count test implemented in the R package spdep (Bivand et al., 2008). The join count test is designed for binary spatial data that can take values commonly referred to as black (1) or white (0). Given two neighboring watersheds with a shared boundary, there are three possible cases for the edge (or join)—BB in which both are positive, WW in which neither are, and BW in which one is and one is not. A join count test considers the total number of (possibly weighted) edges, and compares the actual number of BB edges to the expected number of BB edges computed under the null hypothesis of spatial independence.

We defined a first-order Markov neighborhood structure in which two watersheds i and j have $a_{ij} = 1$ if they share a boundary, and $a_{ij} = 0$ otherwise, where the notation a_{ij} refers

to *adjacency*. These values define the symmetric adjacency matrix $\mathbf{A}$, with entries 1 or 0 depending on the adjacency of watersheds i and j for all possible combinations. The total number of edges in the network are

$$\frac{1}{2}\sum_i \sum_j a_{ij},$$

of which the number of BB pairs is

$$BB = \frac{1}{2}\sum_i \sum_j a_{ij}\, y_i y_j,$$

where $y_i = 1$ if exposure i experienced a flood, and $y_i = 0$ otherwise. One can work out the expected number of BB joins under the assumption of spatial independence as

$$E(BB) = \frac{1}{2}\sum_i \sum_j a_{ij} \left(\frac{n_B}{n}\right)^2,$$

where n_B/n is the fraction of black exposures to total exposures. The variance of the number of BB joins σ^2_{BB} can be worked out in closed form, and used to compute the join count test statistic

$$Z(BB) = \frac{BB - E(BB)}{\sqrt{\sigma^2_{BB}}}. \tag{1}$$

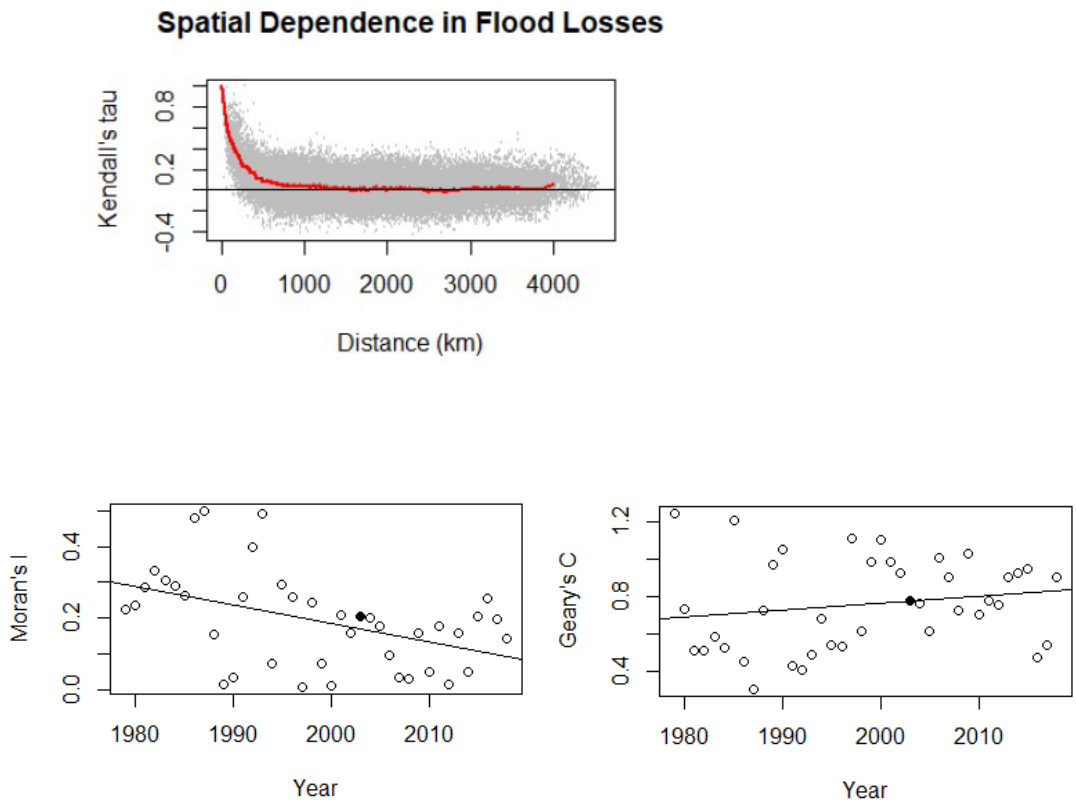
This statistic is asymptotically normal under certain conditions of the spatial network structure (Cliff and Ord, 1981). Positive values indicate spatial clustering, values near zero indicate spatial independence, and negative values indicate a repellent spatial structure that favors dispersion. The bottom-right panel of Figure 6 shows this test implemented for a first-order Markov structure in which two watersheds i and j have $a_{ij} = 1$ if they share a boundary, and $a_{ij} = 0$ otherwise. Z(BB) is shown for each year over the 40-year period. Strong spatial clustering is notable. It has also been trending toward weaker spatial dependence over time.

3.3 Exploratory Spatial Analysis of Losses

Next, we turn our attention to an exploratory analysis of flood losses, with a particular eye toward measures of spatial dependence. Figure 7 shows a plot of Kendall's tau—a nonparametric measure of correlation between two ranked vectors (Kendall, 1938) —for losses per person L_i. We compute this for all pairs of watersheds and plot against the distance between the two watershed centroids. On top of the raw data is a smooth loess curve. As the figure shows, watersheds separated by a short distance display strong positive correlation, meaning both watersheds tend to have large losses or small losses in tandem. This has crucial implications for the ability to diversify a portfolio, since neighboring exposures will not be independent. This positive dependence decays as the distance between watersheds increases. However, even at distances above hundreds of kilometers some evidence of spatial dependence remains (best seen by the red curve

remaining above zero through the 0-to-750-kilometer region). Any statistical model that fails to capture this spatial dependence may understate the risk in a portfolio.

Figure 7. Top: Kendall's tau computed for the 40 years of losses for all 42,486 unique pairs of watersheds. Distances are between the centroids of each watershed, using the haversine method, which assumes a spherical Earth. A smooth loess curve is shown in red. Bottom left: Moran's I for each year's realization, computed from Equation 2. Bottom right: Geary's C for each year's realization, computed from Equation 3. The solid circle shows the statistic computed from 2003.



We further explore two tools used to quantify spatial dependence for a quantitative variable L, which is $ losses per person at each exposure unit. The first measure is Moran's I , defined as

$$I = \frac{n}{\sum_{i,j} a_{ij}} \frac{\sum_i \sum_j a_{ij}(L_i - \bar{L})(L_j - \bar{L})}{\sum_i (L_i - \bar{L})^2}, \tag{2}$$

where L_i is the variable measured at location i, $\bar{L}$ is the mean taken over all spatial locations, a_{ij} is the weight for locations i and j, which is taken to be 1 if neighbors in our setting and 0 otherwise, and n is the total number of locations in the network. Moran's I takes values between -1 and +1, with an expected value

$$E(I) = \frac{-1}{n-1},$$

which is very close to 0 in most settings, including ours. Positive values correspond to positive (clustering) spatial autocorrelation in L, and negative values correspond to negative (repellent) spatial autocorrelation.

A closely related measure is Geary's C (Geary, 1954), defined as

$$C = \frac{(n-1)\sum_i \sum_j a_{ij}(L_i - L_j)^2}{2\left(\sum_{i,j} a_{ij}\right)\sum_i (L_i - \bar{L})^2}, \tag{3}$$

where again a_{ij} is the weight for locations i and j taken to be 1 for neighbors and 0 otherwise. This quantity is positive, with values below 1 indicating positive spatial dependence, and values above 1 indicating negative spatial dependence. Although the two measures appear similar, Geary's C is somewhat more sensitive to localized or regional spatial dependence, whereas Moran's I is more a measure of overall global spatial dependence. Both are implemented in the R package spdep as functions moran.test and geary.test.

We computed Moran's I and Geary's C for losses per person L_i for each of the 40 observed years, with results shown in the bottom two panels of Figure 7. Both measures show positive but declining spatial dependence from 1979 through 2018, matching results from the trends in the BB test statistic over the same period.

4 The Model

Our modeling strategy has two parts. The first piece is a spatially dependent mixed-effects frequency model. It captures the observed spatial dependence in flood losses as shown in the BB statistics in Figure 6. The second piece is a (nonspatial) mixed-effects severity model for losses per person, fit only to nonzero losses. This model assumes severity is conditionally independent across watersheds after controlling for covariates and with a random intercept for watershed to account for unmeasured effects. We fit both models within the Bayesian paradigm, which requires that we specify prior distributions on all unknown parameters and then sample from the posterior using Markov chain Monte Carlo (MCMC). A primary use of both models is to simulate large numbers of realistic flood losses under different climate scenarios and to investigate the differences in those simulations. Accordingly, a distinct advantage of our Bayesian model is that parameter uncertainty is very naturally incorporated into all simulations.

4.1 The Spatial Mixed-Effects Model for Frequency

Let $Y_{i,t}$ be an indicator describing whether the ith exposure did experience flood property damage during time period t ($Y_{i,t} = 1$) or did not ($Y_{i,t} = 0$). As stated earlier, a flood loss is any positive loss-per-person amount in a given year. We define flood loss frequency as a binary variable with probability of occurrence $p_{i,t}$. Although we call this the frequency model, we stress that a watershed either has ($Y_{i,t} = 1$) or does not have ($Y_{i,t} = 0$) a flood loss in a given year, and so this model is binary. Specifically, our model is

$$Y_{i,t} \sim \text{Bernoulli}\left(p_{i,t}\right) \tag{4}$$

$$p_{i,t} = \Phi\left(\boldsymbol{X}'_{i,t}\boldsymbol{\beta} + U_i\right), \tag{5}$$

where $\mathbf{X}_{i,t}$ is a vector of covariates for location i and time t, and $\boldsymbol{\beta}$ is a parameter vector of fixed effects. To control for the effects of unmeasured confounders, we introduce a Gaussian random intercept U_i with mean 0, which allows the model to inflate or deflate flood probabilities for watersheds beyond what the covariates call for. We extend this random effect to incorporate spatial dependence, an example of which was shown in Figure 7. When computing the random effect for location i, this *spatial* random effect allows the process at neighboring locations $j \sim i$ to influence the random effect at location i, and yields a spatially smooth random effect. Specifically, the spatial random effect U_i is defined through conditional distributions

$$U_i \mid \mathbf{U}_{-i} \sim \mathcal{N}\left(\sum_j \frac{a_{ij}}{a_{i+}} U_j, \frac{\sigma_U^2}{a_{i+}}\right), \tag{6}$$

where $U_i \mid \mathbf{U}_{-i}$ refers to the process at location i conditional on the effect all other locations $\mathbf{U}_{-i}$, $a_{i,j} = 1$ if grid cells i and j are neighbors, a_{i+} is the total number of neighbors of grid cell i, and σ_U^2 is the variance.

This spatial random effect is likely to be a new statistical concept for many actuaries, so we unpack it a bit more. This random effect is simply a random intercept indexed by watershed. Observe that Equation 6 says that the distribution of the random effect at location i is normally distributed, with a mean equal to the mean of the random effect at its neighbors. This definition imposes spatial dependence and ensures a spatially smooth random effect, an example of which is seen later in the "Results" section in Figure ***9***. Equation 6 yields the (improper) joint distribution

$$f(\mathbf{U}) \propto \exp\left(-\frac{1}{2\sigma_U^2}\mathbf{U}^T(\mathbf{D} - \mathbf{A})\mathbf{U}\right), \tag{7}$$

where $\mathbf{A}$ is the symmetric adjacency matrix composed of all $a_{i,j}$ and $\mathbf{D} = diag(a_{1+}, \ldots, a_{I+})$. This model is known as the intrinsic conditionally autoregressive model (ICAR), and can be used as a prior model for the spatial process provided a centering constraint $\sum_{i=1}^{I} U_i = 0$ is enforced (Besag and Kooperberg, 1995). Banerjee et al. (2004) and Wall (2004) discuss more on these issues.

We fit this model in the Bayesian paradigm, in which we specify prior distributions on parameters and sample from the posterior using MCMC. For the ICAR model we consider, it is convenient to introduce a latent Gaussian process to facilitate sampling from the posterior. Let $\tilde{Y}_{it}$ be a latent version of the binary flood process Y_{it} such that

$$\tilde{Y}_{it} = \mathbf{X}'_{it}\boldsymbol{\beta} + U_i + \epsilon_{it}, \tag{8}$$

where $\epsilon_{it} \overset{iid}{\sim} \mathcal{N}(0,1)$. Define Y_{it} to be the indicator of $\tilde{Y}_{it} > 0$. Marginally, Equation 8 implies

$$[Y_{it}|\boldsymbol{\beta}, \mathbf{U}] = \text{Bernoulli}\left(\Phi(\mathbf{X}'_{it}\boldsymbol{\beta} + U_i)\right),$$

which brings us back to our desired model shown in Equations 4 and 5.

4.2 Mixed-Effects Model for Severity

Let $L_i, i = 1,\ldots,11{,}680$ represent the loss per person in exposure i of the data set. Such losses are represented by $L_i \geq 0$, but here we focus only on the nonzero losses $L_i \mid L_i > 0$ for the severity model. There are 8,919 nonzero losses in this data set, and they form a strongly right-skewed distribution. We used a Box-Cox transformation with $\lambda = 1/15$ to yield rough normality (see Figure 8), and we call these transformed nonzero losses L_i^*, where now $i = 1,\ldots,8919$. We modeled these with a linear mixed-effects model, with fixed effects drawn from the full set of climate and watershed features (though notably with population removed, as losses have already been normalized per person). The model also contains a (nonspatial) random intercept for watershed to account for unmeasured variables. The reason we opted not to consider a spatial random effect is that a shifting subset of the 292 watersheds experience a nonzero loss each year over the 40 years, whereas we have a full set of observed data for flood loss frequency at all 292 watersheds each year. Specifically, the severity model is

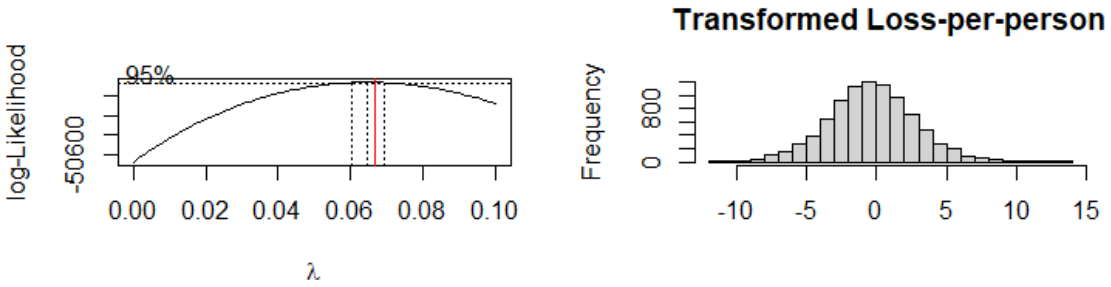
$$L_i^* = \beta_{0,k(i)} + \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i \tag{9}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_{L*}^2) \tag{10}$$

$$\beta_{0,k(i)} \sim \mathcal{N}\left(0, \sigma_{\beta_0}^2\right), \tag{11}$$

where $\beta_{0,k(i)}$ is the random effect for the $k(i)$th watershed, drawn from a normal distribution centered at overall intercept β_0 with variance $\sigma_{\beta_0}^2$, and σ_{L*}^2 is the variance of errors.

Figure 8. Box-Cox transformed losses per person L*=(Lλ-1)/λ, with λ=1/15 closely following a normal distribution.



4.3 Model Fit

We fit the model using the software NIMBLE (de Valpine et al., 2017) on a standard laptop. NIMBLE is a system for programming statistical algorithms for general model structures within . It extends the familiar BUGS (Gilks et al., 1994) and JAGS (Plummer et al., 2003) languages to handle more general or sophisticated hierarchical models. It contains both a number of algorithms for sampling within MCMC and a compiler that

generates (much faster) compiled C++ code, which is run from R and does not require that the user know anything about compilers of C++. These features combine to produce a programming environment that can fit Bayesian spatio-temporal models at a very low computational cost.

For the frequency model, the first question was to determine which subset of the climate and HydroATLAS covariates to include. As a preliminary step we ran an ordinary probit regression (treating all exposures as independent) using all available covariates in the full model, and used stepwise selection with BIC as the selection criteria to select a model. The goal was to remove covariates with minimal statistical connection to flood occurrence. The later inclusion of spatial dependence acknowledges that observations are *not* independent, with therefore less overall information in the data set and (often) larger standard errors on parameter estimates. Accordingly, if a covariate were dropped in a stepwise search assuming independence, we could confidently conclude it would not belong in a spatial model either. All climate variables were included, with log-transformations for both daily and weekly precipitation, and population was also included and log transformed. All parameters for retained covariates received uninformative mean-zero Gaussian priors with variance 1,000. All standard deviations were given uniform (0,100) priors, and the random effect $\mathbf{U}$ had an ICAR prior distribution as shown in Equation 7, with adjacency matrix $\mathbf{A}$ determined from the neighborhood structure of the 292 watersheds. The MCMC had a burn-in of 5,000 iterations, and a run of 5,000 beyond that.

The approach for fitting the severity model was analogous, although the log of population was excluded as a covariate since it was used to define the response of losses per person. We used an initial stepwise selection on a fixed-effects independent model to determine covariates for inclusion, and all other steps and prior distributions as described above and run in NIMBLE. We ran multiple chains at different initial values to assess convergence, and visually inspected trace plots to ensure mixing.

5 Results

5.1 Frequency Model Results

Table 4 shows posterior means, posterior standard deviations, and 95% credible intervals. For display purposes we show the climate variables first, then population, then the HydroATLAS variables, and finally the standard deviation of the random effect.

First, we consider the climate impacts. The model reports $\beta_{log.pr.wk} = 0.7197$, with a 95% credible interval of (0.5623, 0.8756), well away from zero and certainly "significant." If we take a climate change scenario that increases the maximum weekly precipitation by 10% for all watersheds, this would increase the probit of the probability of a flood by $\log(1.1) \cdot 0.7197 = 0.0686$. Similarly, for maximum daily precipitation we see a positive posterior mean of 0.3049, and a 95% credible interval well away from zero (0.1386, 0.4738). A 10% increase in maximum daily precipitation would increase the probit of a flood probability by $\log(1.1) \cdot 0.3049 = 0.029$. Since an overall increase in rainfall would tend to increase both the daily and weekly maximum, these two climate impacts both serve to increase flood probability under a rainier future climate.

For minimum daily precipitation, we see that each increase in minimum temperature by 1 degree serves to decrease the probit of flood probability by 0.0227 (95% credible interval (-0.0591, 0.0162)). This credible interval covers zero, and the magnitude of the impact itself is small compared with the much larger impact from precipitation. Putting these three climate impacts together, a warmer and wetter future climate would overall result in an increased flood probability.

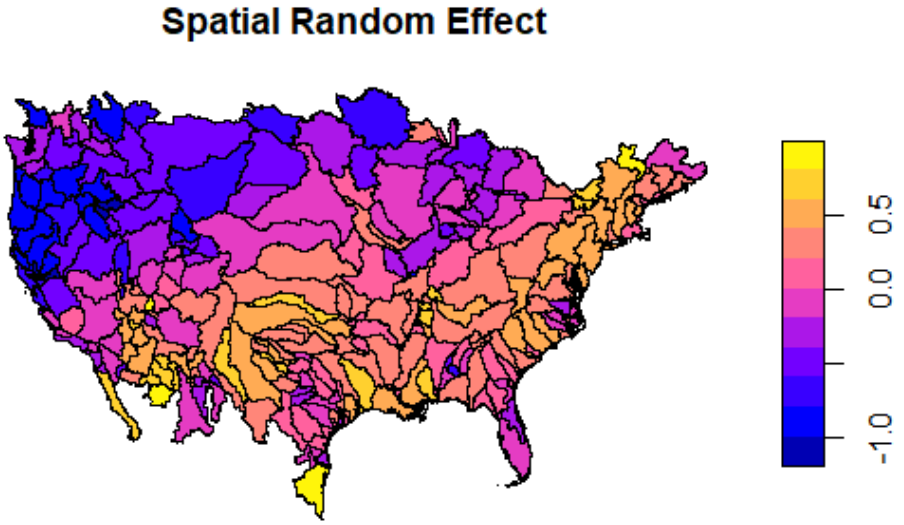
Second, we see that a large number of HydroBASIN features are significant predictors of flood occurrence. More important, the inclusion of these features controls for these watershed characteristics and avoids confounding with climate parameter estimates. Third, observe the strongly significant and positive impact of the log of population, with 95% credible interval (0.1607, 0.2243). This matches previous studies that have found that a primary cause of the increase in flood losses over time is shifting exposures.

Finally, Figure 9 shows the posterior mean of the spatial random effect $\mathbf{U}$. Recall this serves as the random intercept for each distinct watershed, but it also ensures that spatial structure is captured. We see regions (northwestern United States) whose flood occurrence was a bit lower than what the covariates would have suggested. We also see other regions (southern and eastern United States) whose flood risk occurrence was a bit higher than what the covariates alone would have suggested. The spatial random effect accounts for this, regardless of which unmeasured confounders are responsible for it. All model results therefore control for the both spatial dependence as well as the effects of unmeasured confounders.

Table 4. Posterior mean, standard deviation, and 95% credible interval for parameters in the frequency model. The appendix further describes variable names.

	Mean	SD	2.5%	97.5%
log.pr.wk_coef	0.7197	0.0784	0.5623	0.8756
log.pr_coef	0.3049	0.0835	0.1386	0.4738
tasmin_coef	-0.0227	0.0192	-0.0591	0.0162
log.pop_coef	0.1911	0.0156	0.1607	0.2243
class_coef[1]	1.1366	0.1614	0.7980	1.4369
class_coef[2]	0.9489	0.0663	0.8295	1.0796
class_coef[3]	1.0638	0.2048	0.6116	1.4459
class_coef[4]	0.9121	0.0859	0.7436	1.0841
class_coef[5]	1.0572	0.1019	0.8665	1.2606
class_coef[6]	1.3767	0.3213	0.7666	2.0110
class_coef[7]	1.1481	0.1038	0.9464	1.3476
class_coef[8]	0.2944	0.2076	-0.1050	0.7403
class_coef[9]	-1.2800	0.7637	-2.9520	0.1283
cly_pc_sav_coef	0.0209	0.0133	-0.0042	0.0478
dis_m3_pmn_coef	0.0000	0.0001	-0.0001	0.0002
dis_m3_pmx_coef	0.0001	0.0000	0.0000	0.0002
dis_m3_pyr_coef	-0.0002	0.0001	-0.0003	-0.0000
ele_mt_sav_coef	0.0004	0.0001	0.0002	0.0005
gwt_cm_sav_coef	0.0003	0.0002	-0.0002	0.0008
inu_pc_smn_coef	0.0138	0.0060	0.0013	0.0247
inu_pc_smx_coef	-0.0024	0.0044	-0.0110	0.0054
ria_ha_ssu_coef	0.0000	0.0000	0.0000	0.0000
riv_tc_ssu_coef	0.0706	0.0201	0.0333	0.1115
run_mm_syr_coef	0.0003	0.0003	-0.0002	0.0009
slp_dg_uav_coef	-0.0018	0.0013	-0.0043	0.0007
slt_pc_uav_coef	0.0065	0.0095	-0.0111	0.0254
soc_th_sav_coef	-0.0071	0.0033	-0.0134	-0.0007
soc_th_uav_coef	0.0005	0.0011	-0.0019	0.0029
sigma_U	0.6988	0.0502	0.5967	0.7977

Figure 9. Posterior mean of the spatial random effect U from the ICAR model defined in Equations 6 and 7, demonstrating spatial structure and the need to control for unmeasured confounders.



5.2 Severity Model Results

Table 4 shows posterior means, standard deviations, and 95% credible intervals for the severity model. Annual maximum precipitation and annual maximum weekly precipitation have 95% credible intervals of (0.4665, 1.0826) and (2.1561, 2.7618), respectively. These show strong positive associations between precipitation extremes and severity in flood loss per person. The two temperature variables have somewhat offsetting parameter estimates. As with the frequency model, we see that a large number of measured watershed features have significant impacts. Prominent drivers include land cover classification, with several of the classifications differing markedly from one another.

Nevertheless, overall the signal-to-noise ratio is simply much lower for our severity model than for our frequency model, driven by high variability in the losses per person. This results in higher parameter uncertainty, and it will have notable impacts on our simulation study in the next section.

Table 5. Parameter estimates and 95% credible intervals from the severity model. The appendix further describes the variables. We omit individual random effects for each watershed intercept for brevity.

	Mean	SD	2.5%	97.5%
log.pr.wk_coef	2.4555	0.1548	2.1561	2.7618
log.pr_coef	0.7836	0.1567	0.4665	1.0826
tasmax_coef	-0.2585	0.0324	-0.3277	-0.2008
tasmin_coef	0.1428	0.0358	0.0771	0.2148
class_coef[1]	-1.6319	0.4414	-2.4776	-0.7437
class_coef[2]	-0.6987	0.2884	-1.2476	-0.1263
class_coef[3]	0.0568	0.7999	-1.4431	1.6365
class_coef[4]	0.5771	0.4414	-0.2973	1.4258
class_coef[5]	0.0982	0.3425	-0.5437	0.7922
class_coef[6]	1.7137	1.2225	-0.5823	4.1250
class_coef[7]	-0.5197	0.4793	-1.6061	0.4354
class_coef[8]	-2.5406	1.1190	-4.8433	-0.3488
class_coef[9]	-0.2775	3.2000	-6.2157	6.4237
cly_pc_uav_coef	0.1478	0.0858	-0.0109	0.3220
dis_m3_pmn_coef	0.0002	0.0003	-0.0003	0.0007
dis_m3_pmx_coef	-0.0001	0.0001	-0.0002	0.0001
dis_m3_pyr_coef	-0.0001	0.0003	-0.0007	0.0003
dor_pc_pva_coef	-0.0002	0.0001	-0.0004	0.0000
ele_mt_sav_coef	0.0001	0.0008	-0.0015	0.0015
ele_mt_smx_coef	-0.0006	0.0002	-0.0009	-0.0000
ele_mt_uav_coef	0.0002	0.0008	-0.0013	0.0019
gwt_cm_sav_coef	0.0024	0.0009	0.0008	0.0042
inu_pc_smn_coef	0.0400	0.0167	0.0098	0.0736
inu_pc_ult_coef	0.0186	0.0199	-0.0210	0.0559
inu_pc_umx_coef	-0.0224	0.0211	-0.0615	0.0178
ppd_pk_uav_coef	0.0001	0.0003	-0.0005	0.0006
ria_ha_ssu_coef	0.0000	0.0000	-0.0000	0.0001
ria_ha_usu_coef	0.0000	0.0000	-0.0000	0.0000
riv_tc_ssu_coef	0.0260	0.1286	-0.2236	0.2532
riv_tc_usu_coef	-0.0459	0.0958	-0.2143	0.1577
run_mm_syr_coef	-0.0013	0.0008	-0.0029	0.0002
sgr_dk_sav_coef	0.0071	0.0028	0.0019	0.0129
slp_dg_sav_coef	-0.0307	0.0098	-0.0517	-0.0120
slp_dg_uav_coef	0.0219	0.0082	0.0058	0.0359
slt_pc_sav_coef	-0.0351	0.0961	-0.2188	0.1750
snd_pc_sav_coef	0.0298	0.0560	-0.0786	0.1532
snd_pc_uav_coef	0.0276	0.0524	-0.0664	0.1449
soc_th_sav_coef	0.0013	0.0102	-0.0197	0.0208
urb_pc_sse_coef	-0.0907	0.0283	-0.1469	-0.0355
urb_pc_use_coef	0.0149	0.0200	-0.0197	0.0558
sigma_Y	0.1320	0.0020	0.1282	0.1359
sigma_beta0	1.8729	0.1055	1.6773	2.0971

5.3 Simulation Study

Here we describe the use of our two fitted models to explore consequences of different climate scenarios on flood risk. The baseline scenario uses 1979–2018 averages for the four climate variables tasmin, tasmax, log(pr), and log(pr.wk). Scenario A assumes a 4-degree Celsius increase in both minimum and maximum daily temperatures, and a 10% increase in both the maximum daily and maximum weekly precipitation. This scenario is roughly in line with the more aggressive RCP 8.5 projections. The link between warmer temperatures and increased precipitation is assumed because a warming climate system is expected to increase total precipitation overall, with the additional possibility of a shift in extreme precipitation, since the increase in total precipitation is unlikely to be uniform across the full range of precipitation. Given the signs and magnitudes of the estimated parameters for climate variables in Table 4 and Table 5, scenario A should show a modest increase in flood risk.

One computational advantage of a Bayesian model is the ease with which one can simulate from the posterior predictive distributions, $\pi(Y_{new} \mid Y_{observed})$ for occurrence and $\pi(L^*_{new} \mid L^*_{observed})$ for severity. For both the frequency and severity models, we incorporated sampling from the posterior predictive distribution at each iteration of the MCMC to serve as a simulated occurrence and severity, which fully reflected parameter uncertainty from the posterior distribution. This is analogous to sampling a parameter from the posterior and then simulating a new realization of the process given this draw.

The simulated frequency Y_{new} was used to simply zero out simulated losses L^*_{new} for watersheds as needed, which resulted in spatially consistent simulations of flood losses. From our simulated data, we computed the proportion of floods, join count test statistic, Moran's I, and Geary's C under both climate scenarios. We repeated this process 5,000 times to produce distributions of all of the above. Figure 10 and Figure 11 show the distributions. The left panel shows kernel density estimates of both, and the right panel shows a quantile-quantile plot of these two distributions, to aid with seeing the distinction.

Focusing first on frequency, from the top of Figure 10 we see an overall increase in flood probabilities under scenario A, as expected. A warmer, wetter future climate with 10% more precipitation increases the average watershed flood probability. The bottom of Figure 10 shows a *very mild* shift in BB statistics for scenario A as compared to baseline. The difference in means is 0.076, and a t-test confirms it is a statistically significant difference. This implies a shift toward *less* spatial dependence under the climate change scenario A. This simulated shift matches the direction of the observed empirical shift taken over 1979–2018 shown in the bottom-right panel of Figure 6, and demonstrates that climate change may be a mild contributor to a shift in the spatial dependence structure of flood loss occurrence, in addition to any other trends acting on the dependence. However, further study is needed to strengthen this claim.

Moving next to severity, we simulated losses L^* from the severity model for each of the 292 watersheds under the baseline scenario and also scenario A. We simulated $L^*_{new} = \beta_{0,k} + \mathbf{X}\boldsymbol{\beta}$ drawing the parameters from the posterior but ignoring the stochasticity of losses per person through the random error ϵ_i. This was done to allow a greater ability to

detect any shifts in Geary's C or Moran's I across the scenarios, since adding randomness to losses serves only to blur the spatial dependence signal. These simulations were transformed back to the loss-per-person scale L_i, and then multiplied by the simulated flood loss frequency Y_{new} to zero out some losses and retain the others. This was possible by the assumption of independence between frequency and severity in our model. From our simulated realization of losses per person at each watershed, we computed Moran's I and Geary's C under both scenarios, for a total of 5,000 simulations. Figure 11 shows the distributions of Moran's I and Geary's C under both scenarios. Since the severity model has very large unexplained variability, distinctions are not evident in Moran's I or Geary's C. We see evidence of (perhaps) a mild shift in the upper tail for Geary's C; however, this difference is only for simulations with Geary's C above 1.2 and these are exceedingly uncommon. A t-test does not reject the null hypothesis of no differences in means under both scenarios. Our conclusion for severity is that any difference in the spatial dependence of losses per person driven by climate is small in comparison to the random variability of the process, and therefore our study does not detect any climate impacts on changing dependence in flood losses per person.

Figure 10: Results of a simulation frequency study from the fitted model shown in Table 4 and Table 5. Baseline refers to climate variables computed as the means over 1979–2018. Scenario A adds 4 degrees Celsius to minimum daily temperature and 10% to both daily and weekly maximum precipitation. The left panel shows the distributions of over 5,000 simulations. The right panel shows a quantile-quantile plot.

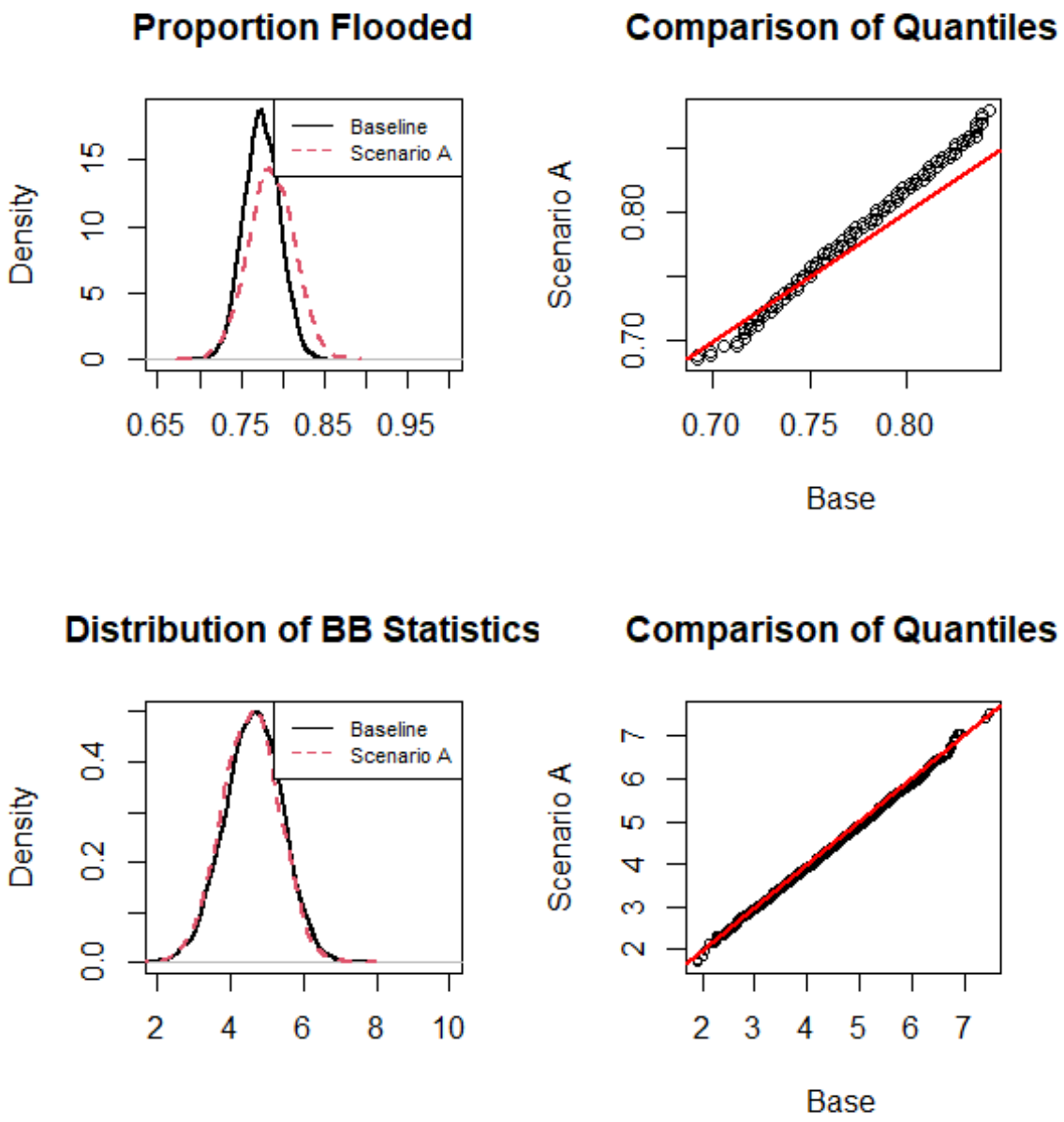
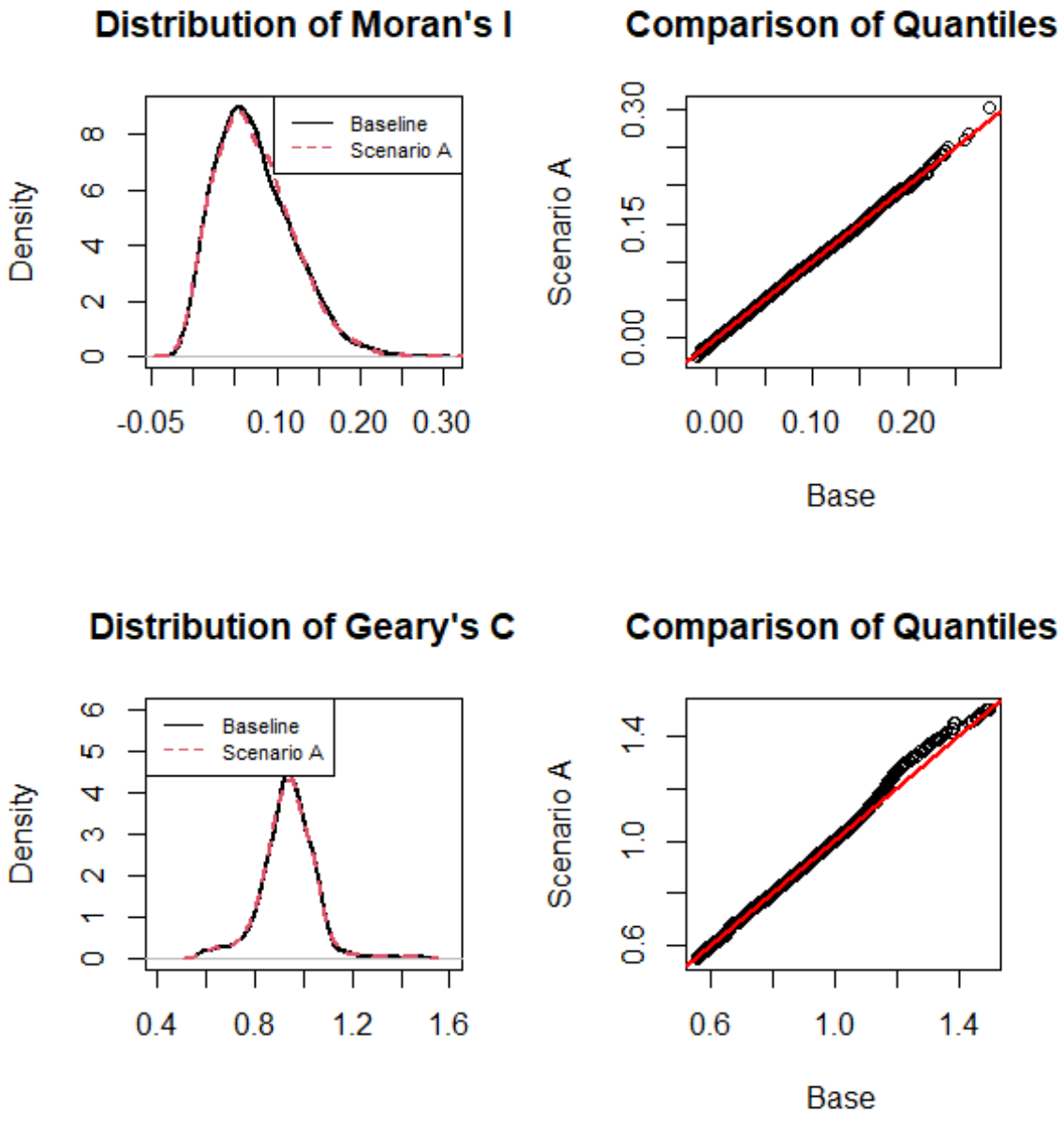


Figure 11: Results of a simulation study from the fitted model shown in Table 4 and Table 5, in which expected losses per person are simulated (without a random error bek added). Baseline refers to climate variables computed as the means over 1979–2018. Scenario A adds 4 degrees Celsius to minimum daily temperature and 10% to both daily and weekly maximum precipitation. The left panel shows the distributions of over 5,000 simulations. The right panel shows a quantile-quantile plot.



6 Discussion

To recap, a review of the literature revealed unresolved questions regarding the links between climate change and the frequency or severity of floods, with measured claims from the IPCC reports on the subject. Strong positive trends in economic losses arising from floods over time are evident, but those trends blend shifting exposures with shifting climate. Studies that normalize total economic losses for exposure shifts yield estimated climate effects that vary by region and study methodology. We also observed a paucity of studies that account for spatial dependence in a statistically optimal manner and propagate uncertainty from initial inputs through simulated outputs in a rigorous manner. To address this, we focused on economic measures of flood loss and built a spatial model across 292 watersheds in the continental United States with the specific goal of quantifying the role that climate plays in the frequency, severity, and spatial dependence of flood loss.

We argued that the watershed is a natural spatial unit to consider flood dependence. We showed a historical shift in flood occurrence probability as well as the spatial dependence of flood loss occurrence. We constructed statistical models that control for a number of known confounders, include a random effect for unmeasured confounders, incorporate spatial dependence in the frequency model, and can be fit in a Bayesian framework that allows simulated floods to fully incorporate all parameter uncertainty. Our models enabled a scientific explanation of the drivers in flood loss occurrence, severity, and spatial dependence, and further enabled a range of simulation studies to explore shifts in spatial dependence arising from changes in climate or exposure. Those simulation studies documented the direction and magnitude of climate change impacts in flood probability. They also suggested that continuation toward a warmer, rainier future climate might continue the trend away from spatial dependence in flood frequency, though other effects dominate the observed trends in spatial dependence and variability in the simulations was high.

We should comment on the granularity of the study, specifically to make clear what extensions to this study are possible. Here we presented a continental U.S. study designed to show the lingering spatial dependence at lengths in the hundreds of kilometers (see Figure 7). We drew upon raw climate variables available as rasters with a 0.5-degree spatial resolution, and specifically sought to ensure reliable measurements of functions of these over a sufficiently large temporal and spatial range. This led us to level 5 watersheds. As mentioned in Section 2.1, HydroSHEDS is available at 12 levels, the last of which provides a very high spatial precision. Depending on the geographic extent and desired environmental and climate covariates, more regional or focused studies using the same methodology at higher spatial resolutions are possible. We hope to see such studies in the future.

As sources of available data continue to grow, actuaries will increasingly face the possibility of constructing ever-larger custom databases to permit studies. The cost of collecting and storing environmental data is falling, new data products have ever-finer spatial and temporal precision, and the possibilities for adding additional variables to risk studies grow daily. This will allow for enhanced actuarial models to be better able to

capture risk and inform risk-management practices. However, these larger high-dimensional data sets will require more sophisticated statistical models and greater computational skills from the actuaries who fit and interpret them. We argue that these trends are inevitable and that actuaries should embrace them. Part of our intention in this paper is to nudge the intrepid actuary toward new computational tools and statistical models needed for environmental risk studies. One of the authors has recently written on this point (Erhardt (2021), https://theactuarymagazine.org). As actuaries extend the computational and statistical reach of the profession, they will simultaneously extend their ability to explain the drivers of risk to stakeholders, regulators, and customers, and give a clearer accounting of the uncertainty in actuarial models.

7 Acknowledgments

The authors thank the Casualty Actuarial Society for its generous financial support of this project. In particular, the authors wish to thank Brian Fannin for his role as project liaison with the CAS and for his deep interest extending the role of statistics and data science for casualty actuarial science.

8 References

Banerjee, S., Carlin, B., & Gelfand, A. (2004). *Hierarchical modeling and analysis for spatial data.* Boca Raton: CRC Press.

Besag, J., & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika, 82*, 733–746.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). *Applied spatial data analysis with R* (Vol. 747248717). Springer.

Bouwer, L. M. (2011). Have disaster losses increased due to anthropogenic climate change? *Bulletin of the American Meteorological Society, 92*, 39–46.

Brunner, M. I., Gilleland, E., Wood, A., Swain, D. L., & Clark, M. (2020). Spatial dependence of floods shaped by spatiotemporal variations in meteorological and land-surface processes. *Geophysical Research Letters, 47*, e2020GL088000.

Carozza, D. A., & Boudreault, M. (2021). A global flood risk modeling framework built with climate models and machine learning. *Journal of Advances in Modeling Earth Systems, 13*.

Center for Emergency Management & Homeland Security, A. S., & Phoenix, A. Z. (2020). Spatial Hazard Events and Losses Database for the United States, Version 19.0. [Online Database].

Choi, O., & Fisher, A. (2003). The impacts of socioeconomic development and climate change on severe weather catastrophe losses: Mid-Atlantic Region (MAR) and the US. *Climatic Change, 58*, 149–170.

Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: models & applications.* Taylor & Francis.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics, 26*, 403–413.

Dottori, F., Szewczyk, W., Ciscar, J.-C., Zhao, F., Alfieri, L., Hirabayashi, Y., . . . others. (2018). Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change, 8*, 781–786.

Erhardt, R. 2021. "How Could Environmental Risk Modeling Help Shape Actuarial Education?" *The Actuary,* June.

Field, C. B., Barros, V., Stocker, T. F., & Dahe, Q. (2012). Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change. Cambridge University Press.

Furman, E., Su, J., Chen, Y., Santoshkumar, S., & Zhang, L. (2019). Modeling, Measuring and Pricing Flood Risk: An Actuarial Perspective. *Society of Actuaries*.

Gall, M., Borden, K. A., & Cutter, S. L. (2009). When do losses count? Six fallacies of natural hazards loss data. *Bulletin of the American Meteorological Society, 90*, 799–810.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician, 5*, 115–146.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *Journal of the Royal Statistical Society: Series D (The Statistician), 43*, 169–177.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*, 81–93.

Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., . . . others. (2014). Flood risk and climate change: global and regional perspectives. *Hydrological Sciences Journal, 59*, 1–28.

Lamb, R., Keef, C., Tawn, J., Laeger, S., Meadowcroft, I., Surendran, S., . . . Batstone, C. (2010). A new method to assess the risk of local and widespread flooding on rivers and coasts. *Journal of Flood Risk Management, 3*, 323–336.

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes, 27*, 2171–2186.

Linke, S., Lehner, B., Dallaire, C. O., Ariwi, J., Grill, G., Anand, M., . . . others. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data, 6*, 1–15.

Metin, A. D., Nguyen, D., Schröter, K., Vorogushyn, S., Guse, B., Kreibich, H., & Merz, B. (2020). The role of spatial dependence for large-scale flood risk estimation. *Natural Hazards and Earth System Sciences (NHESS), 20*, 967–979.

Michel-Kerjan, E. O. (2010). Catastrophe economics: the national flood insurance program. *Journal of economic perspectives, 24*, 165–86.

Miller, S., Muir-Wood, R., & Boissonnade, A. (2008). An exploration of trends in normalized weather-related catastrophe losses. *Climate extremes and society, 12*, 225–247.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*, 17–23.

Nguyen, V. D., Metin, A. D., Alfieri, L., Vorogushyn, S., & Merz, B. (2020). Biases in national and continental flood risk assessments by ignoring spatial dependence. *Scientific reports, 10*, 1–8.

Plummer, M., & others. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*, *124*, pp. 1–10.

Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., . . . Heffernan, J. (2019). The spatial dependence of flood hazard and risk in the United States. *Water Resources Research, 55*, 1890–1911.

Verdin, K. L., & Verdin, J. P. (1999). A topological system for delineation and codification of the Earth's river basins. *Journal of Hydrology, 218*, 1–12.

Villarini, G., Serinaldi, F., Smith, J. A., & Krajewski, W. F. (2009). On the stationarity of annual flood peaks in the continental United States during the 20th century. *Water Resources Research, 45*.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of statistical planning and inference, 121*, 311–324.

Wyncoll, D., & Gouldby, B. (2015). Integrating a multivariate extreme value method within a system flood risk analysis model. *Journal of Flood Risk Management, 8*, 145–160.

Appendix

Table 6. Description of model covariate names.

Variable Name	Explanation
log.pr.wk_coef	Log of the annual maximum weekly precipitation
log.pr_coef	Log of the annual maximum daily precipitation
tasmax_coef	Maximum daily surface temperature
tasmin_coef	Minimum daily surface temperature
class_coef[1]	Land cover classification—tree cover, broad-leaved, deciduous
class_coef[2]	Land cover classification—tree cover, needle-leaved, evergreen
class_coef[3]	Land cover classification—tree cover, regularly flooded
class_coef[4]	Land cover classification—regularly flooded shrub and/or herbaceous cover
class_coef[5]	Land cover classification—regularly flooded cover of mosses
class_coef[6]	Land cover classification—cultivated and managed areas
class_coef[7]	Land cover classification—cropland/tree cover
class_coef[8]	Land cover classification—snow and ice
class_coef[9]	Land cover classification—artificial surfaces and associated areas
cly_pc_sav_coef	Clay fraction in soil sub-basin (average)
cly_pc_uav_coef	Clay fraction in soil upstream of pour point (average)
dis_m3_pmn_coef	Natural discharge at sub-basin pour point (minimum)
dis_m3_pmx_coef	Natural discharge at sub-basin pour point (maximum)
dis_m3_pyr_coef	Natural discharge at sub-basin pour point (annual average)
dor_pc_pva_coef	Degree of regulation at sub-basin pour point
ele_mt_sav_coef	Sub-basin elevation (average)
ele_mt_smx_coef	Sub-basin elevation (maximum)
ele_mt_uav_coef	Elevation upstream of pour point (average)
gwt_cm_sav_coef	Global groundwater map at sub-basin (average)
inu_pc_smn_coef	Inundation extent at sub-basin (minimum)
inu_pc_ult_coef	Inundation extent upstream of pour point (long-term maximum)
inu_pc_smx_coef	Inundation extent in sub-basin (maximum)
inu_pc_umx_coef	Inundation extent upstream of pour point (maximum)
ppd_pk_uav_coef	Population density upstream of pour point (average)
ria_ha_ssu_coef	Total river area in sub-basin
ria_ha_usu_coef	Total river area upstream of pour point
riv_tc_ssu_coef	Total river volume in sub-basin
riv_tc_usu_coef	Total river volume upstream of pour point
run_mm_syr_coef	Annual average land surface runoff in sub-basin
sgr_dk_sav_coef	Average stream gradient in sub-basin
slp_dg_sav_coef	Average terrain slope in sub-basin
slp_dg_uav_coef	Average terrain slope upstream of pour point
slt_pc_sav_coef	Silt fraction in sub-basin soil (average)
slt_pc_uav_coef	Silt fraction upstream of pour point (average)
snd_pc_sav_coef	Sand fraction in sub-basin soil (average)
snd_pc_uav_coef	Sand fraction in soil upstream of pour point
soc_th_sav_coef	Organic carbon content in soil sub-basin (average)
soc_th_uav_coef	Organic carbon content upstream of pour point (average)

Variable Name	Explanation
urb_pc_sse_coef	Spatial urban extent in sub-basin
urb_pc_use_coef	Spatial urban extent upstream of pour point
sigma_Y	Standard deviation
sigma_U	Standard deviation of spatial random effect
sigma_beta0	Standard deviation of random intercept