Astin Bulletin (2025), pp. 1-21
doi:10.1017/asb.2025.7

RESEARCH ARTICLE

Risk modeling of property insurance claims from weather events

Lisa Gao[1] and Peng Shi[2]

[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[2]Department of Risk and Insurance, Wisconsin School of Business, University of Wisconsin–Madison, Madison, WI 53706, USA
Corresponding author: Lisa Gao; Email: lisa.gao@uwaterloo.ca

Received: 5 June 2024; **Revised:** 21 December 2024; **Accepted:** 3 February 2025

Keywords: Replicated spatial data; factor copula; claims management; weather property risk; insurance analytics

Abstract

The localized nature of severe weather events leads to a concentration of correlated risks that can substantially amplify aggregate event-level losses. We propose a copula-based regression model for replicated spatial data to characterize the dependence between property damage claims arising from a common storm when analyzing its financial impact. The factor copula captures the location-based spatial dependence between properties, as well as the aspatial dependence induced by the common shock of experiencing the same storm. The framework allows insurers to flexibly incorporate the observed heterogeneity in marginal models of skewed, heavy-tailed, and zero-inflated insurance losses, while retaining the model interpretation in decomposing latent sources of dependence. We present a likelihood-based estimation to address the computational challenges from the discreteness and high dimensionality in the outcome of interest. Using hail damage insurance claims data from a US insurer, we demonstrate the effect of dependence on claims management decisions.

1. Introduction

The (re)insurance industry has traditionally placed greater emphasis on modeling natural catastrophe perils with the highest loss potentials, such as earthquakes and cyclones, compared to the more frequent, small-to-mid-sized events sometimes termed "secondary perils," such as thunderstorms, hail, and flash floods. This delineation has blurred in recent years, with secondary perils being the predominant driver of catastrophe insurance losses over the last decade (Aon, 2020) and causing over 70% of the \$81 billion (USD) insured catastrophe losses in 2020 worldwide (Swiss Re Institute, 2021). As shifting weather and socioeconomic patterns continue generating larger property exposure (Changnon, 2009), increased modeling and monitoring for secondary perils offer opportunities for insurers to incorporate detailed weather and claims information to manage weather risk more effectively.

The purpose of our work is to provide a statistical method for insurers to predict the financial impact of spatially dependent property insurance losses arising out of convective storms. Since the 1970s, severe convective storms have been responsible for more insurance damage than any other secondary peril in the US, with significantly increasing insured losses throughout this period even after normalization (Barthel and Neumayer, 2012). Despite being short-lived and geographically smaller-scaled, convective storms can exhibit extreme intensity that can accumulate very large losses at the aggregate event level (Insurance Information Institute, 2020b). Their highly localized nature results in a setting where concentrated groups of properties are simultaneously affected by storms of varying intensity, inducing complicated dependence structures between claims.

The dependence between insured property losses from a storm can arise from two sources: the inherent spatial dependence between properties based on their physical proximity to each other and the dependence induced by the common shock of experiencing the same storm. To accommodate the distinctive features of insurance losses, we propose a copula-based regression model for replicated spatial data. The proposed method demonstrates two key advantages, aligning well with the goal of this study: first, it establish a predictive model for non-normal insurance claims data, mixed outcomes in our context, while incorporating heterogeneity from weather, property, and policy characteristics; second, it extends the model to the multivariate context in the spatial dimension, accommodating both spatial and aspatial dependence among insurance claims. It is notable that predictive modeling for replicated spatial data, particularly for non-normal data, has received limited attention in the literature. Our work builds on and further extends the copula-based regression literature. The central piece of our model is the spatial factor copula proposed by Krupskii *et al*. (2018), where the dependence structure is formulated as the copula extracted from a spatial factor process. The underlying factor process consists of a random field specifying location-based dependence and a latent factor jointly affecting all observations in a storm regardless of location. The resulting copula-based regression framework allows us to flexibly accommodate the unique features of insurance losses such as skew, excess zeros, and heavy tails, under well-studied univariate models incorporating rich covariate information, while retaining the interpretation of latent sources of dependence among the joint losses.

The particular type of convective storm event in our application focus is hail, the leading sub-peril comprising 50–80% of convective storm losses in any given year (Insurance Information Institute, 2020b). Hail and wind damage have consistently accounted for the most substantial portion of annual US homeowners insurance losses over the past two decades (Insurance Information Institute, 2020a). In 2021, the US experienced 20 weather and climate disasters where the damages exceeded $1 billion, 7 of which were related to hail damage (NOAA NCEI, 2022). We analyze a replicated spatial dataset of hail property claims that leverages hail radar maps to identify claims arising out of a given hailstorm. Incorporating rich weather, property, and policy characteristics, we demonstrate how insurers can harness the growing availability of detailed weather and claims information to enhance their claims management for hail and other storm perils.

There are limited studies in the literature assessing hail risk using property insurance claims, primarily due to the lack of data (Changnon, 1999). Recent studies are Brown *et al*. (2015), Shi *et al*. (2022), and Gao and Shi (2022). From a loss control perspective, Brown *et al*. (2015) evaluate the resilience of various roofing materials to hail by analyzing claims associated with a single large storm in the Dallas-Fort Worth area from multiple local insurers. Shi *et al*. (2022) study the arrival times of insurance claims, as well as their relationship with the subsequent loss amounts using a marked counting process. To support proactive insurer claims management, Gao and Shi (2022) predict the geographical distribution of insurance claim reports immediately after a hailstorm occurs by incorporating high-resolution weather information in a spatial point process framework. In contrast, our work emphasizes the dependence among joint insured losses from a storm that are particularly prone to concentration as a hazard and examines claim severity given the reporting of claims using methods for spatial modeling of geostatistical data. Our model allows insurers to decompose sources of dependence in hail property claims to make claims management decisions based on predictive distributions of the insured losses. For example, predicting the joint losses stemming from a hailstorm can inform insurers' weather risk retention and transfer decisions, which we demonstrate with an application in the reinsurance context. We also supplement the literature on insurance claims modeling for other secondary perils and convective storm events. Recent settings include flooding (Lyubchich and Gel, 2017), water damage (Haug *et al*., 2011; Scheel *et al*., 2013; Spekkers et al., 2013), and thunderstorm winds in Hua *et al*. (2017).

More broadly, our work extends the actuarial literature on the role of statistical methods and predictive analytics in improving insurance operations. A comprehensive review on analytics in key operational areas of non-life insurance is given in Frees (2015), which emphasizes the value of adopting a micro-oriented view of individual insurance contract developments. Micro-level models using granular insurance data have been actively developing in the key functional areas of ratemaking (for example, Shi *et al*., 2016; Yang and Shi, 2019; Henckaerts *et al*., 2021, and Shi and Zhao, 2020) and loss reserving (for example, Pigeon *et al*., 2013; Antonio and Plat, 2014; Wüthrich, 2018, and Badescu *et al*., 2016; Badescu *et al*., 2019). We contribute to the burgeoning literature on micro-level models supporting insurer claims management, particularly for insurance claims due to weather-related perils (for example, Shi *et al*., 2022 and Gao and Shi, 2022). As losses from secondary perils continue to increase, insurers are reevaluating their reinsurance protection, and aggregate excess-of-loss reinsurance is an effective way for insurers to manage the volatility in event severity (Aon, 2021). Our model allows insurers to capture the dependence in property claims stemming from a storm event and obtain predictive distributions of storm-level losses to better understand excess layers for retention and transfer decisions.

Finally, on the computational front, our contribution involves a two-stage estimation procedure to estimate the parameters in the proposed model. Existing literature on spatial modeling of insurance claims is sparse, with a few examples including Shi and Shi (2017) and Tufvesson *et al*. (2019) in automobile insurance, Zhao *et al*. (2021) in property insurance, and Huang *et al*. (2024) in crop insurance. To the best of our knowledge, our work is among the first to address the unique features posed by replicated spatially dependent insurance claims data. A number of characteristics unique to our context make existing estimation approaches, such as maximum likelihood estimation (Huang *et al*., 2024) and Bayesian estimation (Shi and Shi, 2017; Tufvesson *et al*., 2019) prohibitively expensive. First, unlike the repeated measurements taken at a low number of fixed locations, our replicated spatial data are unbalanced, where the number of claims in a storm may range from one to thousands at varying locations, leading to dimensionality concerns. Second, insurance losses often exhibit a probability mass at zero from coverage denials or modifications such as the deductible. The discreteness in the outcome greatly complicates the evaluation of the copula density in the likelihood function by introducing a numerical integration problem. At the intersection of these two challenges, a large number of zero-payment claims within a storm require computing high-dimensional multiple integrals with dimensions that may be in the hundreds or thousands. The proposed two-stage estimation, inspired by the multilevel composite likelihood method in Zhao *et al*. (2021), extends its scope to the context of replicated spatial data.

The rest of the article proceeds as follows: Section 2 describes the property insurance hail claims data, Section 3 introduces our spatial factor copula regression model, Section 4 discusses the estimation procedure of our proposed model, Section 5 analyzes the hail damage claims using our model, Section 6 explores a claims management application, and Section 7 concludes.

2. Data

The severity of hail damage property insurance claims that arise from multiple hailstorms constitutes replicated spatial data, where the insurer observes the geographical locations of the reported claims associated with each hailstorm. Similar to the approach in Shi *et al*. (2022) and Gao and Shi (2022) examining the occurrence of claims, we construct the dataset of hail claims by overlaying hail radar maps on top of the insurer's exposure map of in-force policies to study the dependence in the severity of the reported claims from each storm.

Our hail property insurance claims data consist of policyholder claims from a major U.S. insurer operating in the personal homeowners line of business in Kansas over 2011–2015. The radar maps of hail experience were obtained by the insurer from a third-party vendor. We define hailstorms based on hail event day within a county, which assumes that all hail experienced in a county on a given day belong to a distinct storm. We acknowledge that the definition of a hailstorm is a data limitation in this setting since the hail maps do not distinguish between different hail swaths on the same day, and only identify regions that experienced hail on a particular day. In total, we observe 839 hailstorms in Kansas producing 23,882 hail property claims over the sampling period.

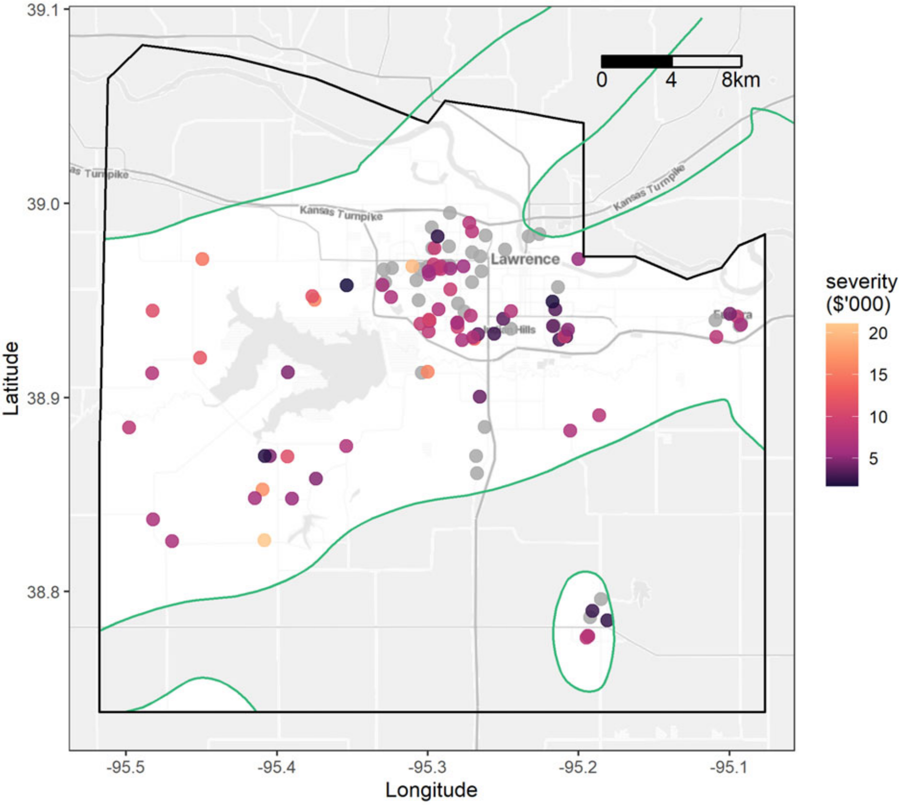

Figure 1. *Locations and severity of hail claims from storm on April 3, 2011, in Douglas County (black) with hail radar outline (green). Claims with zero payment are in gray.*

As an example, Figure 1 displays the locations and severity of the claims arising out of a hail event on April 3, 2011, as identified by the hail radar map (green) in Douglas County (black), with zero payment claims in gray. As foreshadowed by the figure, not all filed claims result in positive payments. There are filed claims with "denied coverage," a general term we use to encompass any reason for zero payments, including, but not limited to, the case where the loss amount falls below the policy deductible. The hail claim severity outcomes are thus semicontinuous, with a discrete probability mass at zero and a positive continuous component. Figure 2 exhibits the distribution of the insured losses on the original scale in the left panel and on the log scale in the right panel. Notably, the distribution features inflated zero outcomes and heavy tails.

The observed heterogeneity at the claim locations for each hailstorm consists of traditional policyholder-level rating variables and claim characteristics, as well as event-level features that can be spatially invariant or spatially varying. At the policyholder level, property characteristics consist of the building age in years, construction type (frame, masonry, or other), roof type (asphalt, slate, metal, tile, wood, flat, or other), and property type (outbuilding, single family residence, or multi-family residence). Policy characteristics consist of the deductible and coverage amount, and claim characteristics consist of the reporting lag in days between the hailstorm occurrence and claim filing dates.

Event-level features are predictors that are uniquely measured for each hailstorm. As a result, these features are storm specific and may either be spatially varying or identical for all properties exposed to the same storm. In our analysis, event-level features include the season (winter, spring, summer, or fall), density of the insured properties affected by the storm (number of exposures per unit area of the radar-defined storm observation window), as well as spatially varying weather information on hail size, wind

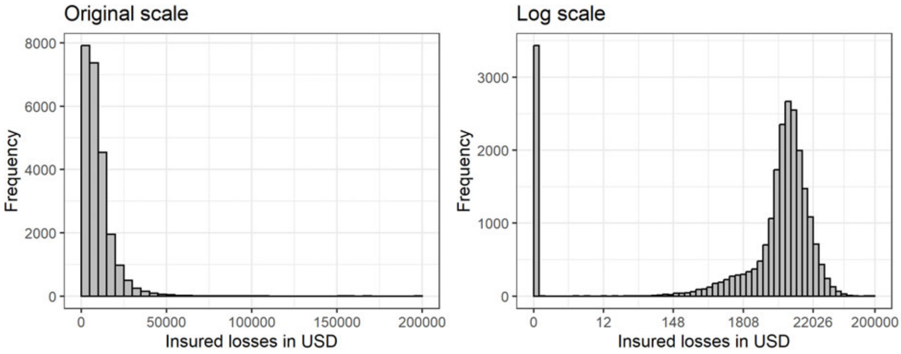

Figure 2. *Distribution of claim severity following a hailstorm on the original scale (left) and log scale (right).*

speed, and wind direction. Due to the data limitation of only observing the date of a hailstorm rather than the exact timestamp, corresponding weather information is also at a daily resolution. Hail sizes for individual hailstorms (measured as hailstone diameter in inches) are given in the hail radar maps from the insurer's third-party vendor. We obtain wind information across each storm from daily U.S. weather data collected from land surface stations by the National Oceanic Atmospheric Administration (NOAA; see NOAA *et al*., 1998 for data collection details) and archived in the Global Historical Climatology Network (GHCN)-Daily database (Menne *et al*., 2012). We incorporate wind information from the 262 Kansas climate-monitoring stations on the days with hailstorm occurrences over 2011–2015. To illustrate our proposed method, we use ordinary kriging to interpolate wind measurements at the claim locations following Gao and Shi (2022), but note that spatial interpolation is a large topic in atmospheric science and statistical science literature. Wind speed measures the rate at which air is moving horizontally past a given point in meters per second and records the fastest consecutive 2-min average speed over the 24-h period ending at local midnight. Wind direction describes the prevailing direction from which the fastest 2-min wind is blowing, in tens of degrees from 10° clockwise through 360°, where 360° represents a north wind (i.e., blowing from the north to the south). As a circular predictor where 10° is closer to 360° than to 30°, we decompose the wind direction into the sine and cosine (vertical and horizontal components respectively) prior to interpolation (Al-Daffaie and Khan, 2017).

The event-level, policy/claim, and property characteristics at the claim locations from each hailstorm are summarized separately across claims that resulted in zero loss payment and claims that resulted in a positive loss payment in Table 1. On average, larger hail sizes and higher wind speed are observed in claims that resulted in positive loss payments compared to claims with no payment. In addition, there is a longer average reporting lag between the hailstorm occurrence and claim filing for claims with no payment, which may indicate that minor property damage takes longer to detect and is less likely to exceed the deductible. On average, relatively more claims from properties with asphalt roofing or frame construction resulted in positive payments. The observed heterogeneity from the policy-level and storm-level weather covariates, along with the spatial and aspatial dependence between claims from a given hailstorm, are characteristics that motivate our factor copula model.

3. Methodology

We use a factor copula regression framework to jointly model the insured losses of property insurance claims that are filed following the occurrence of a hailstorm. Assume a hailstorm generates n property damage claims at locations $\{s_1, \ldots, s_n\} \in \mathbb{R}^2$. Let $\{Y(s_1), \ldots, Y(s_n)\}$ be the collection of corresponding claim payments and denote $Y(s_j) = Y_j$ for $j = 1, \ldots, n$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Further, denote

Table 1. *Descriptive statistics for covariates by whether the claim paid was positive.*

	Zero loss paid				Positive loss paid			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Weather characteristics								
Hail size	1.460	0.623	0.750	4.000	1.752	0.638	0.750	4.000
Wind speed	14.093	2.859	7.572	18.712	14.368	2.988	6.213	18.848
Wind direction (sin)	−0.079	0.265	−0.665	0.921	−0.112	0.284	−0.666	0.921
Wind direction (cos)	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
Other event-level characteristics								
Exposure density	0.095	0.110	0.001	1.750	0.093	0.110	0.001	1.750
Season								
Summer	0.158	0.364			0.142	0.349		
Fall	0.028	0.165			0.025	0.157		
Winter	0.019	0.137			0.022	0.145		
Spring	0.795	0.404			0.811	0.391		
Policy/claim characteristics								
Deductible ($’000)	0.805	0.817	0.050	10.000	0.719	0.594	0.050	10.000
Coverage amount ($’000)	202.297	120.431	5.200	1552.700	198.137	115.278	1.000	2019.100
Reporting lag	38.784	65.193	0.555	363.834	31.193	60.404	0.474	364.785
Property characteristics								
Property type								
Outbuilding	0.001	0.030			0.001	0.030		
Single family	0.990	0.100			0.992	0.090		
Multi-family	0.009	0.096			0.007	0.084		
Roof type								
Slate	0.001	0.024			0.001	0.030		
Metal	0.006	0.074			0.008	0.087		
Tile	0.008	0.088			0.005	0.073		
Asphalt	0.888	0.316			0.925	0.263		
Wood	0.094	0.292			0.057	0.232		
Flat	0.003	0.051			0.002	0.044		
Other	0.001	0.034			0.002	0.042		
Construction type								
Masonry	0.052	0.222			0.045	0.206		
Frame	0.946	0.226			0.953	0.212		
Other	0.002	0.045			0.003	0.053		

the set of predictors for the property located at s by $\boldsymbol{X}(s) = (X_1(s), \ldots, X_p(s))$, which may consist of a combination of weather information, property characteristics, and contract features. We similarly use the notation $\boldsymbol{X}(s_j) = \boldsymbol{X}_j$ for $j = 1, \ldots, n$, and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. By Sklar’s theorem for conditional distributions (Patton, 2006), the joint distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ can be represented using a multivariate copula such as

$$\begin{aligned} F_Y(y_1, \ldots, y_n|\boldsymbol{X}) &= \Pr(Y_1 \le y_1, ..., Y_n \le y_n|\boldsymbol{X}) \\ &= C(F_1(y_1|\boldsymbol{X}), ..., F_n(y_n|\boldsymbol{X})|\boldsymbol{X}), \end{aligned} \tag{3.1}$$

where $F_j(\cdot|\boldsymbol{X})$ is the conditional distribution function for Y_j given $\boldsymbol{X}$ and $C(\cdot|\boldsymbol{X})$ is the conditional copula. We make two additional assumptions: (i) $F_j(\cdot|\boldsymbol{X}) = F_j(\cdot|\boldsymbol{X}_j)$ and (ii) $C(\cdot|\boldsymbol{X}) = C(\cdot|\boldsymbol{D})$, where $\boldsymbol{D}$ is a nonstochastic symmetric matrix of pairwise spatial distances, that is, $\boldsymbol{D} = \left[r_{ij}\right]_{i,j=1}^{n}$ with $r_{ij} = r_{ji}$. The first assumption permits different sets of covariates to be used for each marginal distribution. In our application, it is reasonable to assume that, given the predictors measured at the location of the property, predictors at other locations carry no additional information on the insured losses. The second assumption is standard in spatial statistics literature; we assume second-order spatial stationarity, where conditional on the observed heterogeneity, the association between two observations is a function of their spatial distance.

In our context, a policyholder may be denied payment due to contract features such as the deductible. As a result, the distribution of Y_j features a probability mass at zero. For vector $\boldsymbol{Y}$, let $I_+ = \{i_1, \ldots, i_k\}$ be the indices for the positive payments. The joint density of (3.1) can be expressed as

$$f_{\boldsymbol{Y}}(y_1, \ldots, y_n|\boldsymbol{X}) = \begin{cases} C(F_1(y_1|\boldsymbol{X}_1), \ldots, F_n(y_n|\boldsymbol{X}_n)|\boldsymbol{D}), & k = 0 \\ C^{(k)}(F_1(y_1|\boldsymbol{X}_1), \ldots, F_n(y_n|\boldsymbol{X}_n)|\boldsymbol{D}) \prod_{j\in I_+} f_j(y_j|\boldsymbol{X}_j), & 0 < k < n \\ c(F_1(y_1|\boldsymbol{X}_1), \ldots, F_n(y_n|\boldsymbol{X}_n)|\boldsymbol{D}) \prod_{j=1}^{n} f_j(y_j|\boldsymbol{X}_j), & k = n, \end{cases} \tag{3.2}$$

where $C^{(k)}(u_1, ..., u_n|\boldsymbol{D}) = \dfrac{\partial^k C}{\partial u_{i_1} \cdots \partial u_{i_k}}$ and $c(\cdot|\boldsymbol{D})$ is the density of the conditional copula.

3.1. Marginal model for insured losses

We consider two strategies to accommodate the probability mass at zero in claim payments. The first strategy assumes that zero payment outcomes are a result of the policy deductible, that is the policyholder receives no indemnification if the ground-up loss is below the per-occurrence deductible. Let Y_j^* be the ground-up loss for the claim at location s_j, and d_j be the deductible that applies to the claim. The insured loss Y_j and ground-up loss Y_j^* satisfy the relation:

$$Y_j = (Y_j^* - d_j)_+ = \max\{Y_j^* - d_j, 0\} = \begin{cases} 0, & Y_j^* \le d_j \\ Y_j^* - d_j, & Y_j^* > d_j. \end{cases}$$

From the above, the distribution and density of claim payment Y_j are

$$F_j(y|\boldsymbol{X}_j) = F_j^*(y + d_j|\boldsymbol{X}_j) \;\; y \ge 0 \tag{3.3}$$

$$f_j(y|\boldsymbol{X}_j) = \begin{cases} F_j^*(d_j|\boldsymbol{X}_j) & y = 0 \\ f_j^*(y + d_j|\boldsymbol{X}_j) & y > 0, \end{cases} \tag{3.4}$$

where $F_j^*(\cdot|\boldsymbol{X}_j)$ and$f_j^*(\cdot|\boldsymbol{X}_j)$ are the distribution and density functions of the ground-up loss, respectively.

The second strategy models the claim payment directly. We consider a two-part mixture regression:

$$F_j(y|\boldsymbol{X}_j) = p(\boldsymbol{X}_j) + (1 - p(\boldsymbol{X}_j))\; G(y|\boldsymbol{X}_j) \;\; y \ge 0 \tag{3.5}$$

$$f_j(y|\boldsymbol{X}_j) = \begin{cases} p(\boldsymbol{X}_j) & y = 0 \\ (1 - p(\boldsymbol{X}_j))\; g(y|\boldsymbol{X}_j) & y > 0 \end{cases} \tag{3.6}$$

where $p(\boldsymbol{X}_j) = \Pr(Y_j = 0|\boldsymbol{X}_j)$ and $G(\cdot)$ and $g(\cdot)$ are the distribution and density functions for a random variable defined on the positive real line. The approach of the two-part model is data driven and ignores the mechanism behind the zero-inflated outcomes; thus, it can be readily applied in settings other than the presence of a policy deductible. In addition, this strategy is consistent with the approximation method for pricing the deductible contract discussed in Lee (2017), where the deductible is treated as a predictor in the regression model.

3.2. Factor copula model for spatial dependence

The association among insurance claims from a storm could arise from two sources: the spatial dependence and the unobserved storm-specific effect. The former reflects a situation where outcomes observed at different locations are correlated with one another, with the correlation between two observations diminishing as their distance increases. The latter points to a special type of aspatial dependence where the unobserved storm-specific effect is a common shock inducing an exchangeable dependence structure among all observations affected by a particular storm.

To account for both the spatial and aspatial dependence, we employ the factor copula introduced by Krupskii *et al*. (2018). The factor copula is derived from a random spatial process $Q(s)$ that varies by location $s \in \mathbb{R}^2$:

$$Q(s) = Z(s) + V_0, \tag{3.7}$$

where $Z(s)$ is a Gaussian random field and V_0 is an independent common factor that does not depend on location s. The Gaussian random field $Z(s)$ in (3.7) captures the spatial dependence among observations in different locations in the correlation function. For instance, the Matérn class of spatial correlation functions characterizes the common empirical observation that the correlation between outcomes at two locations decreases as the distance r increases:

$$\rho(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r}{\xi}\right)^{\nu} K_\nu \left(\frac{r}{\xi}\right),$$

where $K_\nu(\cdot)$ is the modified Kessel function of order ν, $\nu > 0$ is a shape parameter determining the smoothness of the process, and $\xi > 0$ is a scale parameter for the range of the spatial dependence (see Matérn, 1986 and Guttorp and Gneiting, 2006 for more details). In addition, the factor V_0 accommodates the aspatial dependence among all observations within a storm due to the common shocks.

For observations at n unique locations $s_1, \dots, s_n$, the finite-dimension representation of the spatial process (3.7) is

$$Q_j = Z_j + V_0, \tag{3.8}$$

for $j = 1, \dots, n$, where Q_j and Z_j denote $Q(s_j)$ and $Z(s_j)$, respectively. Then $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ follows a zero-mean, unit-variance multivariate normal distribution $\Phi_{\boldsymbol{\Sigma}_Z}(\cdot|\boldsymbol{\Sigma}_Z)$ with correlation matrix $\boldsymbol{\Sigma}_Z$ and let V_0 follow a distribution $F_{V_0}(\cdot|\boldsymbol{\gamma}_0)$ with parameters $\boldsymbol{\gamma}_0$, independent of $\boldsymbol{Z}$. By Sklar's theorem (Sklar, 1959), we extract the dependence structure of the spatial process $Q(s)$ via the copula implied by the joint distribution of $\boldsymbol{Q} = (Q_1, \dots, Q_n)$ at locations $s_1, \dots, s_n$. Thus, we obtain the factor copula C as

$$C(u_1, \dots, u_n|\boldsymbol{\Sigma}_Z, \boldsymbol{\gamma}_0) = F_{\boldsymbol{Q}}\left((F_{Q,1})^{-1}(u_1|\boldsymbol{\gamma}_0), \dots, (F_{Q,n})^{-1}(u_n|\boldsymbol{\gamma}_0)|\boldsymbol{\Sigma}_Z, \boldsymbol{\gamma}_0\right), \tag{3.9}$$

where $F_{\boldsymbol{Q}}(q_1, \dots, q_n|\boldsymbol{\Sigma}_Z, \boldsymbol{\gamma}_0) = \displaystyle\int_{-\infty}^{\infty} \Phi_{\boldsymbol{\Sigma}_Z}(q_1 - v_0, \dots, q_n - v_0|\boldsymbol{\Sigma}_Z) dF_{V_0}(v_0|\boldsymbol{\gamma}_0)$ is the multivariate distribution function of $\boldsymbol{Q}$. In addition, $F_{Q,j}(q|\boldsymbol{\gamma}_0) = \int_{-\infty}^{\infty} \Phi(q - v_0)\, dF_{V_0}(v_0|\boldsymbol{\gamma}_0)$ is the marginal distribution function of Q_j, where $\Phi(\cdot)$ is the univariate standard normal distribution function.

Factor copulas arise from factor models, where latent random variables are employed to induce dependence among observed variables. These models provide a highly flexible framework for capturing complex dependence structures and are particularly well-suited to high-dimensional settings, where traditional methods may struggle with scalability or interpretability. The development of various factor copulas in the literature has been driven by the need to address unique structural characteristics of different data types across a range of disciplines, including time series data (e.g., Krupskii and Joe, 2013; Oh and Patton, 2017), spatiotemporal data (e.g., Krupskii and Genton, 2017; Krupskii *et al*., 2018), mortality data (e.g., Chen *et al*., 2015), and ordinal data (e.g., Nikoloulopoulos and Joe, 2015), among others.

Lastly, it is worth noting that the joint distribution $F_{\boldsymbol{Q}}$ is only used to extract the dependence structure of the factor spatial process (3.7) and construct the factor copula C, whereupon C joins the marginal models discussed in Section 3.2. The spatial factor copula thus allows us to flexibly accommodate various features of insurance claim payments such as skewness, heavy tails, and excess zeros, as well as incorporate the rich covariate information in the marginal models, while retaining the interpretation of the association arising from spatial and aspatial sources in the dependence model.

3.3 Model specification

In this section, we discuss the detailed specification of the marginal models and the factor copula for our hail property damage claims application. The marginal claim payments are modeled based on the generalized beta of the second kind (GB2) distribution, a four-parameter family that flexibly accommodates the skew and heavy tails commonly exhibited by insurance claims data. The GB2 distribution is well studied in the context of insurance claims modeling (see Shi, 2014 for a review). For a GB2 distributed variable Y with location parameter μ, scale parameter σ, and shape parameters α_1 and α_2, the density is defined on $y > 0$:

$$g(y|\boldsymbol{X}) = \frac{\exp(z)^{\alpha_1}}{y|\sigma|B(\alpha_1, \alpha_2)(1+\exp(z))^{\alpha_1+\alpha_2}},$$

where $z = (\log(y) - \mu(\boldsymbol{X}))/\sigma$ and $B(\cdot,\cdot)$ is the beta function.

In Section 3.1, we presented two strategies for accommodating the probability mass at zero in claim payments. For the first strategy, we assume the ground-up loss Y_j^* follows a GB2 distribution and is left-censored at the per-occurrence policy deductible d_j. The covariates $\boldsymbol{X}_j$ are linearly incorporated in the GB2 location parameter μ, so that $Y_j^* \sim GB2(\mu(\boldsymbol{X}_j), \sigma, \alpha_1, \alpha_2)$, where $\mu(\boldsymbol{X}_j) = \boldsymbol{X}_j\,\boldsymbol{\beta}$ for coefficients $\boldsymbol{\beta}$. The marginal model parameters for the first strategy are $\boldsymbol{\theta}_{1,\text{cens}} = (\boldsymbol{\beta}, \sigma, \alpha_1, \alpha_2)$. For the second strategy of directly modeling the claim payment, we specify $p(\boldsymbol{X}_j)$ using a generalized linear model with a logit link so that

$$p(\boldsymbol{X}_j) = \frac{1}{1 + e^{\boldsymbol{X}_j \boldsymbol{\beta}_1}}$$

for covariates $\boldsymbol{X}_j$, where $\boldsymbol{\beta}_1$ are the coefficients associated with whether a claim has a positive payment. Then $G(\cdot)$ and $g(\cdot)$ defined on the positive real line are the distribution and density functions of the $GB2(\mu(\boldsymbol{X}_j), \sigma, \alpha_1, \alpha_2)$ describing the positive payments, where $\mu(\boldsymbol{X}_j) = \boldsymbol{X}_j\,\boldsymbol{\beta}_2$ for coefficients $\boldsymbol{\beta}_2$. The marginal model parameters for the second strategy are $\boldsymbol{\theta}_{1,\text{2pt}} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma, \alpha_1, \alpha_2)$. Note that the covariate set in the two-part model contains the policy deductible, while that of the censored regression does not.

In our factor copula specification, we assume the correlation matrix of $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_Z)$ corresponding to the Gaussian random field $Z(s)$ takes the form

$$\boldsymbol{\Sigma}_Z = \left[\sigma_{jj'}\right]_{n\times n} = \begin{cases} 1, & j = j' \\ \kappa\tau_{jj'} & j \neq j', \end{cases} \tag{3.10}$$

for $j, j' \in \{1, \ldots, n\}$, where $\kappa \in (0, 1)$ captures the nugget effect for microscale variation and measurement error. Let $r_{jj'} = ||s_j - s_{j'}||$ denote the distance between locations s_j and $s_{j'}$. In addition, $\tau_{jj'} = \exp\{-3r_{jj'}/\psi\}$ is the exponential spatial correlation function, a special case of the Matérn family where the correlation between Y_j and $Y_{j'}$ decreases as the distance $r_{jj'}$ between their locations increases. The parameter ψ controls how quickly the spatial correlation decays with distance and is parameterized to represent the practical range, the distance at which the correlation is 0.05 (Diggle and Ribeiro, 2007). We further assume $V_0 \sim N(0, \sigma_0^2)$ for simplicity, which results in a tractable special case of a structured Gaussian factor copula. Note that $\boldsymbol{Q}$ follows a multivariate normal distribution with correlation matrix

$$\boldsymbol{\Sigma}_Q = (1-\rho)\boldsymbol{\Sigma}_Z + \rho\boldsymbol{J}_n, \tag{3.11}$$

where $\rho = \sigma_0^2/(1+\sigma_0^2) \in (0, 1)$ can be interpreted as the correlation introduced due to the unobserved storm-specific effect and $\boldsymbol{J}_n$ is an n-dimensional square matrix of ones. The resulting factor copula is thus the Gaussian copula

$$C(u_1, \ldots, u_n | \boldsymbol{D}) = \Phi_{\boldsymbol{\Sigma}_Q}\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n) | \boldsymbol{D}\right), \tag{3.12}$$

where $\Phi_{\boldsymbol{\Sigma}_Q}$ is the zero-mean, unit-variance multivariate normal distribution function with correlation matrix $\boldsymbol{\Sigma}_Q$. Denote the association parameters by $\boldsymbol{\theta}_2 = (\kappa, \psi, \rho)$. If the distance $r_{jj'} \to 0$, then the spatial correlation equals κ, reflecting the nugget effect, while the overall correlation is $(1-\rho)\kappa + \rho$. If the distance $r_{jj'} \to \infty$, then the spatial correlation approaches 0 and the overall dependence approaches $\rho > 0$, reflecting the positive contribution of the aspatial dependence in the severity of claims from the same storm regardless of their locations.

The Gaussian factor copula is also particularly convenient for our replicated spatial data and prediction application. Due to the heterogeneity in claim frequency and geographical distribution, the factor copula dimensionality n can vary substantially across storms. The marginal consistency of the Gaussian copula allows us to easily adapt the correlation matrix under changing dimensions while maintaining the same underlying factor covariance structure.

4. Statistical estimation

The parameters in the spatial factor copula regression model are estimated using a fully parametric likelihood-based approach. Assume that for m total hailstorms, the ith hailstorm $(i = 1, \ldots, m)$ is observed to have $n_i > 0$ claims at locations $\{s_{i1}, \ldots, s_{in_i}\}$. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ be the observed claim payments from storm i at the corresponding claim locations for our point-referenced data. Let $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$ be the corresponding covariates associated with the claims from hailstorm i, where $\boldsymbol{x}_{ij} = (x_{i,1}(s_{ij}), \ldots, x_{i,p}(s_{ij}))$ is the covariate vector for claim j in storm i for $j = 1, \ldots, n_i$. The data are summarized by $\{\boldsymbol{y}_i, \boldsymbol{x}_i : i = 1, \ldots, m\}$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the vector of model parameters containing the unknown parameters in the marginal distributions and the factor copula dependence. The log-likelihood of the data can be expressed as

$$\begin{aligned} l(\boldsymbol{\theta}) = &\sum_{\{i:k_i=0\}}^{m} \log C(F_{i1}(y_{i1}|\boldsymbol{x}_{i1}), \ldots, F_{in_i}(y_{in_i}|\boldsymbol{x}_{in_i})) + \\ &\sum_{\{i:0<k_i<n_i\}}^{m} \left\{ \log C^{(k_i)}(F_{i1}(y_{i1}|\boldsymbol{x}_{i1}), \ldots, F_{in_i}(y_{in_i}|\boldsymbol{x}_{in_i})) + \sum_{j \in I_{i+}} \log f_{ij}(y_{ij}|\boldsymbol{x}_{ij}) \right\} + \\ &\sum_{\{i:k_i=n_i\}}^{m} \left\{ \log c(F_{i1}(y_{i1}|\boldsymbol{x}_{i1}), \ldots, F_{in_i}(y_{in_i}|\boldsymbol{x}_{in_i})) + \sum_{j=1}^{n_i} \log f_{ij}(y_{ij}|\boldsymbol{x}_{ij}) \right\}, \end{aligned} \tag{4.1}$$

based on the joint density defined in (3.2). The discreteness in the claim payment outcome and the high dimensionality of the Gaussian factor copula in our insurance claims context offer a more general setting that greatly complicates the traditional maximum likelihood estimation presented in Krupskii *et al.* (2018). In particular, directly evaluating the data likelihood based on the joint density (3.2) involves an $(n_i - k_i)$-dimensional integration where $k_i \leq n_i$ is the number of claims with positive payments as defined in (3.2). As discussed in Section 2, the number of zero components in a hailstorm can be several hundred, which is not computationally viable. Motivated by the computational challenges associated with the high dimensionality of the factor copula and exacerbated by the discreteness in the outcome, we propose a two-stage estimation procedure for this more general setting. The first stage estimates parameters in the marginal models assuming working independence and the second stage estimates parameters in the dependence model via a composite likelihood approach.

4.1. Stage I: Estimating parameters in marginals

In the first stage, we estimate the parameters in the marginal models in the spirit of the inference functions for margins method (Joe and Xu, 1996). Specifically, the log-likelihood function under the working independence assumption is

$$l_1(\boldsymbol{\theta}_1) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{1}(y_{ij}=0)\log f_{ij}(y_{ij}|y_{ij}=0,\boldsymbol{x}_{ij}) + \mathbf{1}(y_{ij}>0)\ \log f_{ij}(y_{ij}|y_{ij}>0,\boldsymbol{x}_{ij}), \tag{4.2}$$

where $f_{ij}(\cdot|\boldsymbol{x}_{ij},\boldsymbol{\theta}_1)$ is the marginal density (under (3.4) or (3.6) for the censored and two-part mixture GB2 strategies, respectively) and $\mathbf{1}(\cdot)$ is the indicator function. The estimators for the parameters $\boldsymbol{\theta}_1$ are obtained by

$$\tilde{\boldsymbol{\theta}}_1 = \arg\max_{\boldsymbol{\theta}_1} l_1(\boldsymbol{\theta}_1). \tag{4.3}$$

4.2. Stage II: Estimating parameters in dependence

In the second stage, we estimate parameters in the factor copula fixing the estimates for parameters in the marginal regression. The discreteness and lack of balance in the data make the inference functions for margins challenging to implement. To further improve computational efficiency, we consider the composite likelihood method for dependence parameter estimation (Lindsay, 1988). Composite likelihood methods have been prevalent in spatial analysis for reducing the computational burden in estimation (see Varin *et al.*, 2011 for a survey). The composite likelihood function results from the product of a collection of component likelihoods, with the responses within a component assumed to be dependent, but orthogonal across components. We consider the special case of pairwise composite likelihood, which takes the form:

$$l_2(\boldsymbol{\theta}_2) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\sum_{j<j'} w(y_{ij},y_{ij'})\ \log f(y_{ij},y_{ij'}|\boldsymbol{x}_i,\tilde{\boldsymbol{\theta}}_1), \tag{4.4}$$

where $\tilde{\boldsymbol{\theta}}_1$ are the marginal parameter estimates in the first stage, $f(y_{ij},y_{ij'})$ denotes the bivariate likelihood function for $(Y_{ij},Y_{ij'})$, and w is the weight given to the pair. The bivariate joint density is

$$f(y_{ij},y_{ij'}|\boldsymbol{x}_i,\tilde{\boldsymbol{\theta}}_1) = \begin{cases} C(F_{ij}(y_{ij}|\tilde{\boldsymbol{\theta}}_1),F_{ij'}(y_{ij'}|\tilde{\boldsymbol{\theta}}_1)), & y_{ij}=0, y_{ij'}=0 \\ C^{(1)}(F_{ij}(y_{ij}|\tilde{\boldsymbol{\theta}}_1),F_{ij'}(y_{ij'}|\tilde{\boldsymbol{\theta}}_1))\,f_{ij}(y_{ij}|\tilde{\boldsymbol{\theta}}_1), & y_{ij}>0, y_{ij'}=0 \\ C^{(2)}(F_{ij}(y_{ij}|\tilde{\boldsymbol{\theta}}_1),F_{ij'}(y_{ij'};\tilde{\boldsymbol{\theta}}_1))\,f_{ij'}(y_{ij'}|\tilde{\boldsymbol{\theta}}_1), & y_{ij}=0, y_{ij'}>0 \\ c(F_{ij}(y_{ij}|\tilde{\boldsymbol{\theta}}_1),F_{ij'}(y_{ij'}|\tilde{\boldsymbol{\theta}}_1))\,f_{ij}(y_{ij}|\tilde{\boldsymbol{\theta}}_1)f_{ij'}(y_{ij'}|\tilde{\boldsymbol{\theta}}_1), & y_{ij}>0, y_{ij'}>0, \end{cases} \tag{4.5}$$

where $C^{(h)}(u_1,u_2)=\dfrac{\partial}{\partial u_h}C(u_1,u_2)$ for $h=1,2$ is the partial derivative of the bivariate copula with respect to the h-th component. Thus, the dependence parameter estimates are computed as

$$\tilde{\boldsymbol{\theta}}_2 = \arg\max_{\boldsymbol{\theta}_2} l_2(\boldsymbol{\theta}_2). \tag{4.6}$$

The pairwise formulation (4.4) is critical to our application in two aspects. First, it lowers the computational burden via dimension reduction. Notably, the high-dimensional integration due to the discreteness in the loss payments within a storm is reduced to a maximum of two dimensions. Second, regardless of the computational burden, the likelihood method is not straightforward to implement due to the lack of balance in the number of claims across storms. The pairwise formulation avoids this issue since all multivariate distributions have a dimension of two.

The above two-stage estimator is an extension of the multilevel composite likelihood in Zhao *et al.* (2021) to the context of replicated spatial data. The estimator is consistent and asymptotically normal, where the asymptotic covariance is given by the inverse of the Godambe information matrix (Godambe, 1960). In many cases, including ours, the asymptotic covariance cannot be solved analytically, although consistent estimates can be obtained via jackknife (Joe and Xu, 1996) or other subsampling techniques (Heagerty and Lumley, 2000).

In implementing the composite likelihood method, an efficient weight can improve both statistical and computational efficiency (see, for instance, Joe and Lee, 2009 and Bai *et al.*, 2012). Denote by

$d(s_{ij}, s_{ij'})$ the Euclidean distance between observations y_{ij} and $y_{ij'}$. We use the weight defined below for (4.4):

$$w(y_{ij}, y_{ij'}) = \begin{cases} 1, & \text{if } d(s_{ij}, s_{ij'}) \leq d \\ 0, & \text{if } d(s_{ij}, s_{ij'}) > d, \end{cases}$$

for a fixed spatial lag d. In data analysis, the optimal value of d is selected to result in the most informative set of estimating equations by minimizing the trace of the inverse of the Godambe information matrix, so as to minimize the asymptotic variance of the estimator (Bevilacqua *et al*., 2012). Numerical experiments have also suggested that shorter distances often result in greater efficiency (Varin *et al*., 2011; Bai *et al*., 2012).

5. Data analysis

We implement the proposed spatial factor copula regression model to analyze the hail property insurance claims data described in Section 2. We use the hailstorms in 2011–2014 as the training data for model development and hold out the hailstorms in 2015 for model validation and prediction. There are 709 and 130 storms (with 21,937 and 1945 corresponding claims) in the training and hold-out data, respectively.

5.1. Estimation results

We fit the factor copula regression model using the proposed two-stage estimation. To summarize, we incorporate the covariates from Section 2 in GB2 marginal claim payment distributions and consider both the deductible-censored and two-part mixture model strategies for accommodating the excess zeros from Section 3.1. Due to insufficient variability in the cosine of the wind direction (see Table 1), we exclude it in the model fitting. Pairwise spatial distances between locations are taken as the great-circle distance in kilometers under the haversine method assuming a spherical earth. Thus, the parameter ψ representing the practical range and the spatial lag d are also interpreted in kilometers.

The tuning parameter d in the composite likelihood weight function is selected based on the criterion of minimizing the trace of the inverse of the Godambe information matrix. We estimate the Godambe information matrix using a delete-subset jackknife approach to alleviate the computational burden (Joe, 2014). Specifically, we randomize the in-sample storms into 18 subsets and re-estimate the full model parameters on each delete-subset dataset. We consider 35 different candidate values for d ranging from 0.25 to 25 km. The optimal spatial lag d selected is 1 and 1.25 km for the censored and two-part mixture GB2 marginal models, respectively.

Table 2 presents the parameter estimates, with standard errors estimated via the delete-subset jackknife. The hail size and wind speed weather characteristics are significant and positively associated with claim payments. Several property characteristics are statistically significant, with older properties experiencing higher claim payments (while exhibiting a lower probability of positive payment in the two-part model). Outbuildings and multi-family properties are both associated with lower claim payments compared to single-family properties. In addition, properties with metal and wood roofing experience higher claim payments compared to the commonly used asphalt (while wood roofing is associated with a lower probability of positive payment in the two-part model). Masonry construction, considered to be sturdier than frame construction, is also associated with lower insurance losses. The association parameters under both models indicate significant correlation introduced due to the unobserved storm-specific effect, with ρ estimates of 0.27 and 0.23 under the censored and two-part models, respectively. In addition, the respective practical range estimates of 1.89 and 1.55 km indicate that the spatial correlation between claimants within a hailstorm is low beyond those distances. The κ nugget effect estimates suggest that there remains short-scale variability in the losses.

Table 2. *Estimation results for the proposed factor copula regression model with two alternative strategies for zero inflation.*

	Censored GB2			Two-part mixture GB2					
				Logit			GB2		
	Est.	Std. error		Est.	Std. error		Est.	Std. error	
Marginal GB2 parameters									
σ	0.083	0.007	***	–	–		0.090	0.001	***
α_1	0.060	0.005	***	–	–		0.113	0.002	***
α_2	0.343	0.034	***	–	–		0.331	0.006	***
Coefficients									
Intercept	1.494	0.151	***	1.711	0.462	***	0.548	0.110	***
Hail size	0.179	0.005	***	0.768	0.043	***	0.147	0.009	***
Wind speed	0.003	0.001	*	0.030	0.013	**	0.005	0.003	*
Wind direction (sin)	0.005	0.026		−0.240	0.072	***	0.057	0.040	
Exposure density	−0.027	0.026		0.565	0.070	***	−0.043	0.025	*
Season									
Summer (reference)									
Fall	−0.110	0.030	***	−0.167	0.106		−0.131	0.033	***
Spring	−0.039	0.014	**	−0.052	0.085		−0.028	0.025	
Winter	0.006	0.027		0.237	0.104	**	−0.007	0.039	
Log building age	0.050	0.004	***	−0.106	0.025	***	0.054	0.005	***
Property type									
Single family (reference)									
Outbuilding	−0.312	0.108	**	−0.493	2.640		−0.264	0.070	***
Multi-family	−0.313	0.020	***	−0.275	0.054	***	−0.350	0.018	***
Roof type									
Asphalt (reference)									
Slate	0.654	0.070	***	0.351	0.228		0.504	0.078	***
Metal	0.536	0.016	***	0.312	0.053	***	0.483	0.022	***
Tile	0.020	0.037		−0.511	0.056	***	−0.150	0.057	***
Wood	0.218	0.011	***	−0.493	0.034	***	0.232	0.009	***
Flat	−0.140	0.042	***	−0.047	0.087		−0.294	0.077	***
Other	0.365	0.038	***	0.639	0.162	***	0.325	0.046	***
Construction type									
Frame (reference)									
Masonry	−0.015	0.012		−0.156	0.037	***	−0.035	0.009	***
Other	0.180	0.033	***	0.426	0.106	***	0.114	0.036	***
Reporting lag	−0.001	0.000	***	−0.001	0.000	***	−0.001	0.000	***
Log coverage amount	0.636	0.012	***	0.001	0.041		0.686	0.010	***
Log deductible	–	–		−0.191	0.010	***	0.028	0.003	***
Dependence parameters									
ψ	1.894	0.269	***	–	–		1.545	0.113	***
ρ	0.272	0.026	***	–	–		0.227	0.009	***
κ	0.233	0.017	***	–	–		0.271	0.017	***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

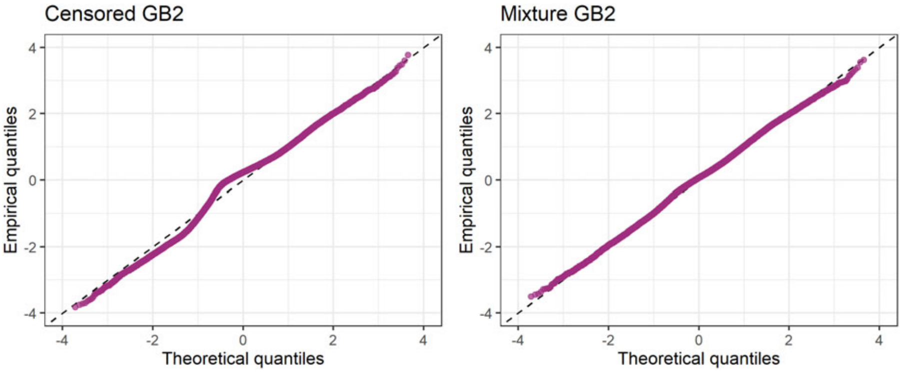

Figure 3. *Q-Q plots of Cox–Snell residuals for the censored (left) and two-part mixture (right) marginal models.*

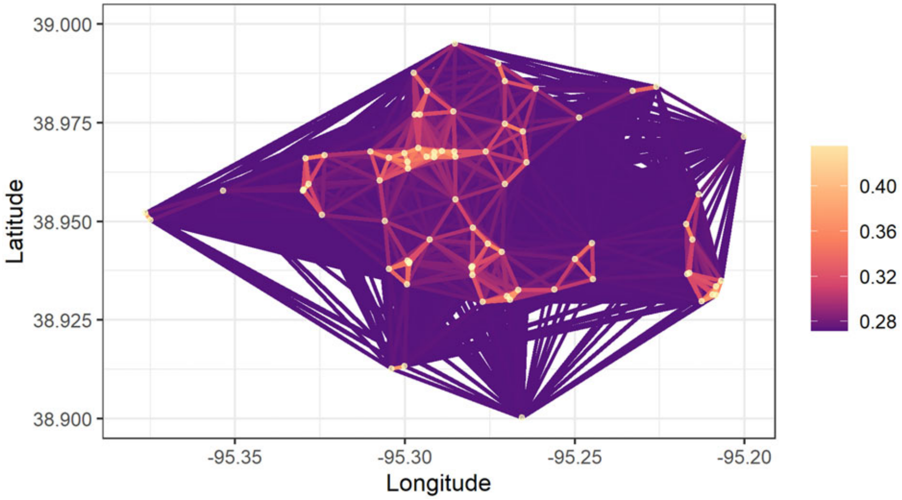

Figure 4. *Example of calibrated dependence illustrated on claims from the storm in Figure 1 as a connected graph, where the colored edges indicate the strength of the pairwise correlations.*

To assess the fit of the marginal models, we consider the Cox–Snell residuals (Cox and Snell 1968) based on the probability integral transform (PIT) (Dawid 1984) of the response using their fitted distributions. In particular, due to the probability mass at zero in the claim payments, the standard PIT is not uniformly distributed. We instead consider the more general randomized PIT approach (Czado *et al*. 2009; Rüschendorf 2009) and use the transformation in Yang and Shi (2019) for semicontinuous observations. Figure 3 displays the Q-Q plots of the generalized Cox–Snell residuals for the censored and two-part mixture models. The Q-Q plots follow the 45-degree line closely, suggesting that GB2 regressions flexibly capture the skewed, heavy-tailed losses. The slight hump near the zero quantile in the censored model suggests that the distributional assumptions of the more flexible two-part mixture marginal better reflects the training data compared to the censored marginal model.

To visualize the calibrated dependence based on the fitted copula parameters in Table 2, Figure 4 maps an example for a group of claims within the storm depicted in Figure 1. The claim locations are plotted as a connected graph, where the colored edges indicate the strength of the fitted pairwise correlations. Correlations are stronger between claims that are geographically closer to each other, although claims within the storm still exhibit low positive correlation at any distance.

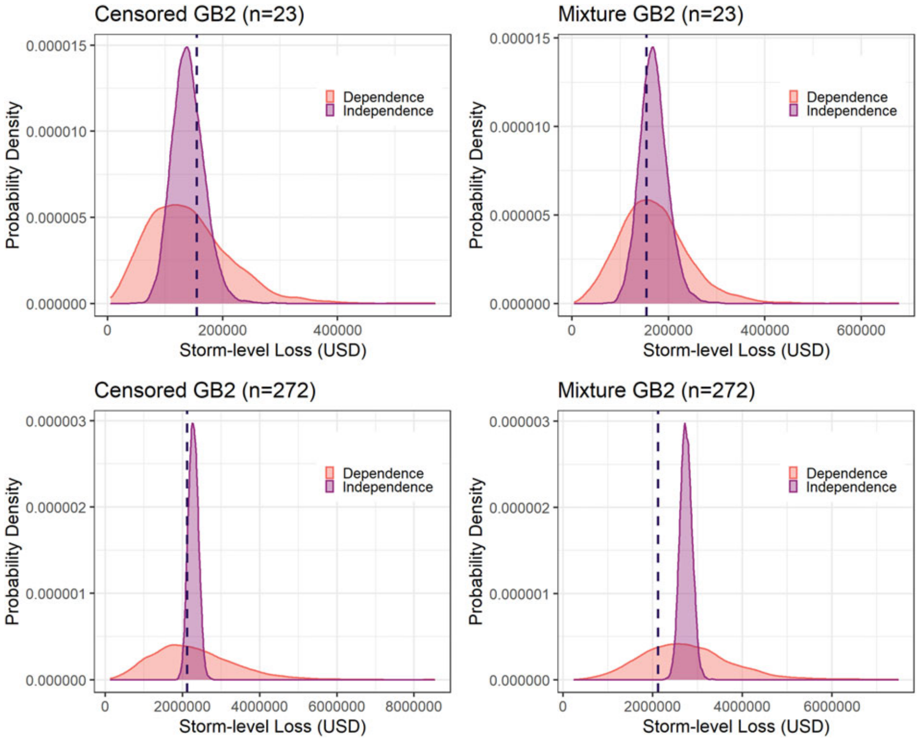

Figure 5. *Predictive distributions of storm-level losses under dependence and independence models for the censored (left) and two-part mixture (right) models, with the observed loss in black.*

5.2. Prediction

We evaluate the dependence assumptions of the proposed factor copula model using the hold-out hailstorms and their associated claims by simulating their predictive distributions. To demonstrate the value of the factor copula, we compare the predictive distributions from the factor copula regression model to those from the corresponding regression models assuming independence between claims. Specifically, based on the fitted model estimates in Table 2, we simulate 5000 realizations of the joint losses for each hold-out storm under the factor copula and under the independence copula to obtain predictive distributions of storm-level losses. For example, Figure 5 displays the predictive distributions of the storm-level losses for two hold-out hailstorms with different numbers of claims n under the factor copula dependence (orange) and independence (purple) models for the censored (left) and the two-part mixture (right) strategies of handling the probability mass at zero. The predictive distributions of storm-level losses under the factor copula have heavier tails and exhibit more right skew compared to those under independence, with the effect becoming more pronounced with higher claim frequency.

To assess the distributional assumptions of the full factor copula model on the hold-out hailstorms, we examine the PIT of the storm-level losses and the coverage probabilities of the prediction intervals. Denote the storm-level losses by $S_i = \sum_{j=1}^{n_i} Y_{ij}$. Based on the simulated predictive distributions, the empirical distribution function of S_i is $\hat{H}_i(s) = (1/b) \sum_{k=1}^{b} \mathbf{1}(s_{ik} \leq s)$, where s_{ik} are the simulated realizations of S_i for $k = 1, \ldots, b$ and $b = 5000$ replications in our setting. The predictive distribution is probabilistically calibrated if the PIT $\hat{H}_i(S_i)$ has a standard uniform distribution. Figure 6 displays the uniform Q-Q plots of the PIT of the hold-out storm losses for the censored (top) and two-part mixture (bottom) factor models (left), compared to their counterparts under independence (right). The Q-Q plots suggest that the independence assumption does not capture the storm-level loss distributions effectively.

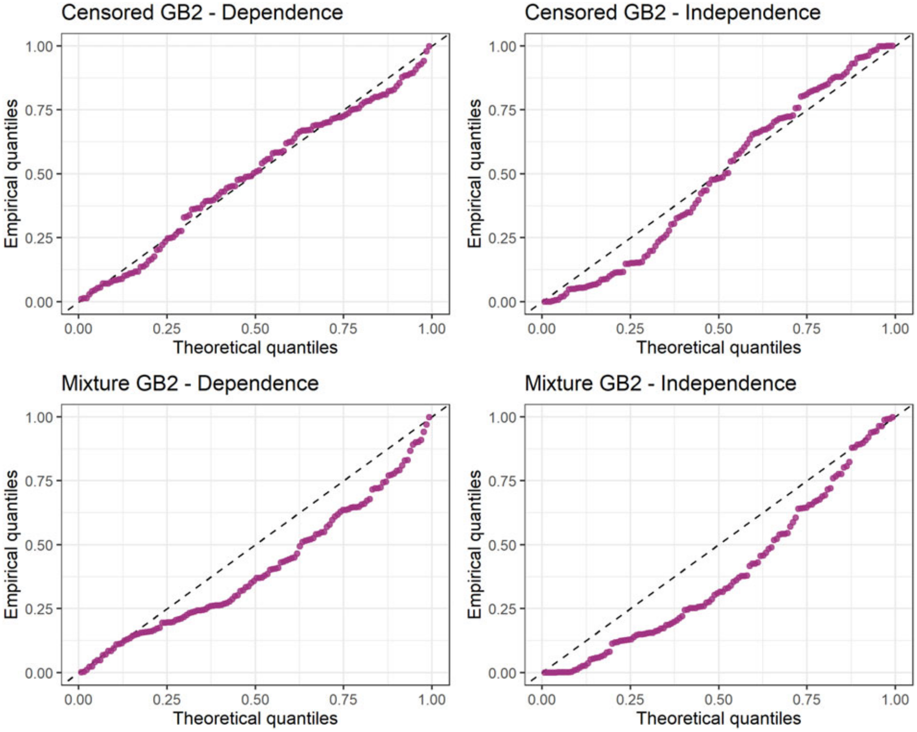

Figure 6. Q-Q plots of PIT of storm-level losses on hold-out sample for the censored (top) and two-part mixture (bottom) dependence models (left), compared to the independence case (right).

The two-part mixture GB2 dependence model similarly does not appear appropriate, showing departures from the 45-degree line. Thus, the censored GB2 factor copula model is better probabilistically calibrated to the hold-out sample compared to the two-part mixture GB2, which may have overfitted the marginal distribution on the training data in Section 5.1.

We further assess out-of-sample calibration by examining the coverage probabilities of the prediction intervals for storm-level losses. We build prediction intervals for a range of levels based on the predictive distributions of S_i. For $\alpha \in (0, 1)$, let $L_{i,PI}(\alpha)$ and $U_{i,PI}(\alpha)$ be the lower and upper bounds of the $100(1-\alpha)\%$ prediction interval of S_i. Then $\hat{L}_{i,PI}(\alpha) = \hat{H}_i^{-1}(\alpha/2)$ and $\hat{U}_{i,PI}(\alpha) = \hat{H}_i^{-1}(1-\alpha/2)$ so that $\Pr(\hat{L}_{i,PI}(\alpha) < S_i < \hat{U}_{i,PI}(\alpha)) \approx 1-\alpha$ with approximately $\alpha/2$ probability in each tail. The coverage probability is the probability that the observed storm-level loss lies in its prediction interval, which we estimate empirically across the hold-out hailstorms. If the distributional assumptions of the model are appropriate, the $100(1-\alpha)\%$ prediction interval should achieve a coverage probability of at least $1-\alpha$. Figure 7 summarizes the coverage probabilities for the prediction intervals for the censored (left) and two-part mixture (right) models. The dependence models appear to obtain the desired coverage probabilities near the 45-degree line, compared to the lower coverage probabilities of the independence models. The censored GB2 dependence model coverage probabilities follow the 45-degree line more closely than the two-part GB2 dependence model, suggesting that, similar to the PIT Q-Q plots, the censored GB2 factor copula model is more distributionally appropriate for the hold-out sample. We proceed with the censored GB2 factor copula model in the applications.

6. Application

Characterizing the dependence between insured losses stemming from a common storm allows insurers to make efficient claims management decisions, especially when claims in a concentrated area are

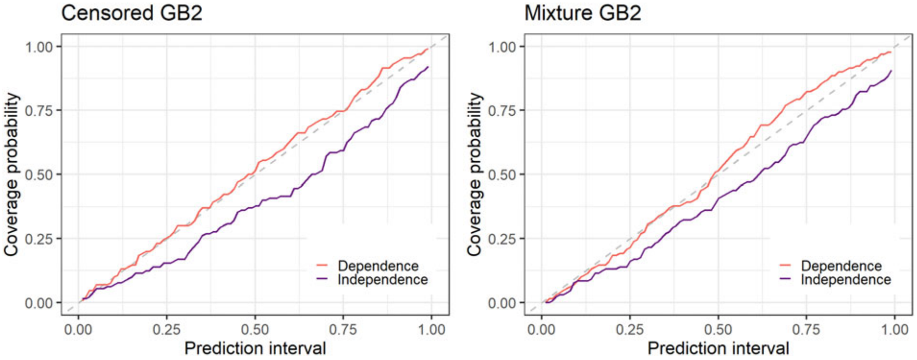

***Figure 7.** Prediction interval coverage probabilities under dependence and independence models.*

spatially correlated. We demonstrate an application in the context of a treaty reinsurance contract on hail property damage claims. In particular, we consider an aggregate excess-of-loss (XL) contract, a non-proportional arrangement where the treaty covers all accumulated losses consequential to a given event that exceed the attachment point (aggregate reinsurance deductible). Aggregate XL reinsurance is particularly well suited for managing property claims from weather-related events as it protects insurers from the accumulation of risks whose individual losses may not be very large, but whose total payout may be substantially amplified by positive correlation. The spatial and aspatial association in our hail property claims exemplify the positive correlation in this setting. In fact, special cases of aggregate XL with high cover are a common method for insurers to manage natural catastrophe-related property losses for primary perils such as earthquakes (Parodi, 2014).

In the presence of an aggregate XL reinsurance treaty that covers event-level losses, insurers are interested in predicting their retained losses to provide claims management insights. Following an observed storm, the insurer can leverage the weather, property, and policy information to obtain a predictive distribution of their retained losses to support their financial planning activities, such as setting case reserves and making risk retention and transfer decisions.

Using the predictive distributions of the hold-out hailstorms in 2015, we compare the factor copula model that captures the spatial and aspatial dependence among losses to the independence assumption for the aggregate XL contract. As an example, for an attachment point of $1 million, Figure 8 shows the cumulative distribution functions of the predictive distributions of the storm-level losses that are retained by the insurer under the dependence and independence models for four storms with varying numbers of claims n, two of which were previously shown in Figure 5. Insurers setting case reserves based on predictive distributions of their retained losses would substantially underestimate the risk if assuming independence, leading to potentially insufficient reserves. To illustrate, consider an insurer that sets the case reserves for each storm equal to the 75% value-at-risk of their retained losses. Based on the distributions in Figure 8, the dependence model implies a $4734 (8.9%) higher case reserve than the independence model for the storm with $n = 5$ claims, a $24,517 (15.5%) higher reserve for the storm with $n = 23$ claims, and a $134,366 (15.5%) higher reserve for the storm with $n = 102$ claims. There is no difference for the storm with $n = 272$ claims due to both of the models predicting a very high (at least 75%) probability of the storm-level loss exceeding the $1 million attachment point.

Considering a range of attachment points, Figure 9 displays the expected retained losses under the dependence and independence models for the same storms as those shown in Figures 5 and 8. At a given attachment point, the dependence model suggests lower expected retained losses (i.e., reinsurance deductible cost) than the independence model. Thus, for a given level of desired retention by the insurer, the dependence model suggests selecting a higher attachment point than under the independence model to manage claim costs. Even when incorporating the same rich covariate information, insurers must

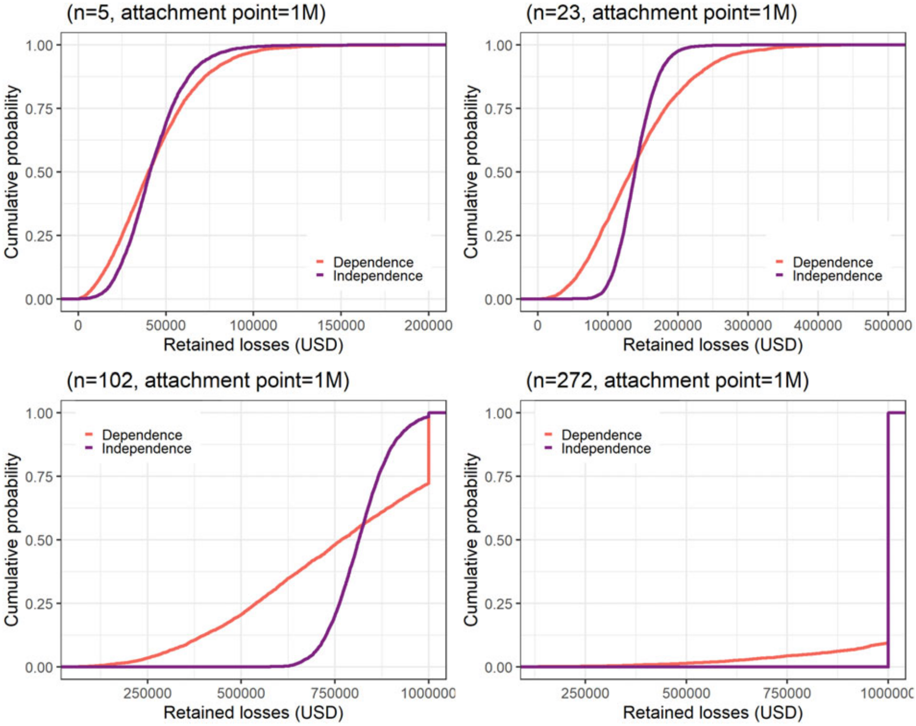

Figure 8. Predictive distributions of retained losses for a common attachment point of $1 million under dependence and independence models. Storms include those shown in Figure 5.

account for the dependence between losses arising from the same storm in their claims management decisions.

Our application involving an aggregate XL reinsurance treaty reveal several implications for insurers' claims management and risk transfer decisions involving excess layers of coverage. First, ignoring the spatial and aspatial dependence among losses originating from a common storm greatly mischaracterizes the distributions of the insurer's retained and ceded losses, which impacts financial planning decisions such as setting case reserves. Second, for a desired level of retention by the insurer, the dependence model suggests that the insurer can seek a higher attachment point than under the independence model.

7. Conclusion

Property damage from hail and other severe convective storm perils is responsible for a substantial portion of US weather-related insurance losses against the backdrop of changing weather and socioeconomic patterns and growing property exposure. The intense, localized nature of convective storms produces a concentration of spatially correlated claim outcomes whose dependence must be characterized to support effective claims management and reinsurance decisions.

We provide a method for insurers to predict the financial impact of spatially dependent property insurance claims arising out of a common storm event. Our copula-based regression for replicated spatial data features a spatial factor copula that captures the spatial dependence between properties that decays with distance, as well as the aspatial dependence from being jointly affected by the same storm irrespective of location. While retaining the interpretation of latent sources of dependence, the framework also allows us to flexibly leverage the suite of well-studied univariate methods of incorporating granular weather and claim information to model skewed, heavy-tailed insurance losses. We

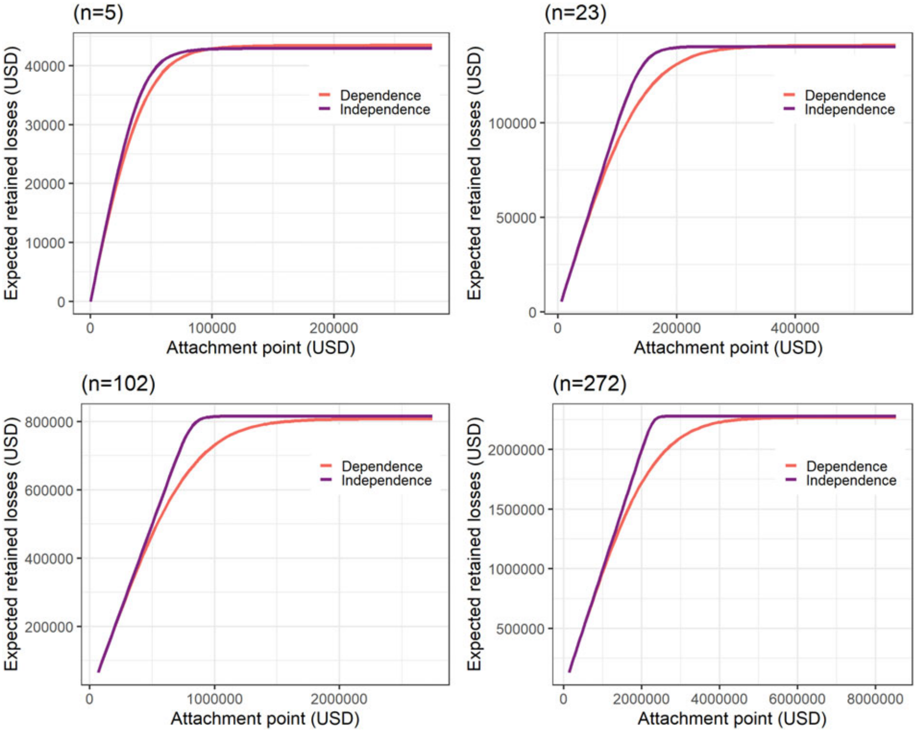

Figure 9. *Expected retained losses for a range of attachment points under dependence and independence models. Storms correspond to those shown in Figure 8.*

propose a likelihood-based estimation procedure to relieve the computational burden associated with high dimensionality and exacerbated by the zero-inflated loss outcomes. To demonstrate the estimation and prediction, we analyze a unique hail claims dataset constructed with hail radar maps and property exposures. We further highlight the implications of the spatial and aspatial dependence on insurer claims management decisions through an application involving an aggregate excess-of-loss reinsurance treaty.

While our proposed method focuses on the analysis of replicated spatial data to model joint property losses, we recognize the importance of combining actuarial models of claim occurrence (e.g., frequency, location, timing, and severity) with climate science models of storm occurrence and weather scenarios to support a broader array of actuarial applications. Our spatial factor copula framework assumes that losses from different storms are independent conditional on the occurrence of claims and the observed heterogeneity. Future work may explore the potential relationship between the frequency and geographical distribution of claims with their severity. Potentially evolving hail patterns due to climate change suggest further consideration of temporal trends across storms. In addition, while the Gaussian factor copula has the ability to accommodate flexible correlation specifications and unbalanced numbers of claims, it is restricted by its symmetry and lack of tail dependence. Growing concern about managing risk associated with the highest loss potentials suggests that future work should explore more flexible dependence characterization.

Our application setting involves hail property loss data, although our model is more broadly applicable to other severe storm settings where a concentrated group of properties are simultaneously affected by an event that introduces potential for positively correlated losses to rapidly accumulate. Our work demonstrates how data and analytics can support our understanding of the dependence in claims arising from common weather events, as well as their effects on claims management decisions. As (re)insurers focus on modeling and monitoring a growing range of natural catastrophe perils, the increasing availability of detailed weather and claims information offer continual opportunities to expand their capacity to accept and manage weather and climate risk.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/asb.2025.7.

Acknowledgements. The authors would like to thank the anonymous referees and the Editor for their thoughtful comments and suggestions that improved the paper. Lisa Gao acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (RGPIN-03389-2023).

Competing interests. The authors have no potential competing or conflicting interests to declare.

References

Al-Daffaie, K. and Khan, S. (2017) Logistic regression for circular data. *AIP Conference Proceedings*, **1842**(1), 030022-1–030022-9.

Antonio, K. and Plat, R. (2014) Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, **2014**(7), 649–669.

Aon (2020) Reinsurance market outlook.

Aon (2021) 2021 weather, climate and catastrophe insight.

Badescu, A.L., Chen, T., Lin, X.S. and Tang, D. (2019) A marked cox model for the number of ibnr claims: Estimation and application. *ASTIN Bulletin: The Journal of the IAA*, **49**(3), 709–739.

Badescu, A.L., Lin, X.S. and Tang, D. (2016) A marked cox model for the number of ibnr claims: Theory. *Insurance: Mathematics and Economics*, **69**, 29–37.

Bai, Y., Song, P.X.-K. and Raghunathan, T. (2012) Joint composite estimating functions in spatiotemporal models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(5), 799–824.

Barthel, F. and Neumayer, E. (2012) A trend analysis of normalized insured damage from natural disasters. *Climatic Change*, **113**(2), 215–237.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2012) Estimating space and space-time covariance functions for large data sets: A weighted composite likelihood approach. *Journal of the American Statistical Association*, **107**(497), 268–280.

Brown, T.M., Pogorzelski, W.H. and Giammanco, I.M. (2015) Evaluating hail damage using property insurance claims data. *Weather, Climate, and Society*, **7**(3), 197–210.

Changnon, S.A. (1999) Data and approaches for determining hail risk in the contiguous united states. *Journal of Applied Meteorology*, **38**(12), 1730–1739.

Changnon, S.A. (2009) Increasing major hail losses in the us. *Climatic Change*, **96**(1), 161–166.

Chen, H., MacMinn, R. and Sun, T. (2015) Multi-population mortality models: A factor copula approach. *Insurance: Mathematics and Economics*, **63**, 135–146.

Cox, D.R. and Snell, E.J. (1968) A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, **30**(2), 248–265.

Czado, C., Gneiting, T. and Held, L. (2009) Predictive model assessment for count data. *Biometrics*, **65**(4), 1254–1261.

Dawid, A.P. (1984) Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, **147**(2), 278–290.

Diggle, P.J. and Ribeiro, P.J. (2007) *Model-based Geostatistics*. New York, NY, USA: Springer.

Frees, E.W. (2015) Analytics of insurance markets. *Annual Review of Financial Economics*, **7**(1), 253–277.

Gao, L. and Shi, P. (2022) Leveraging high-resolution weather information to predict hail damage claims: A spatial point process for replicated point patterns. *Insurance: Mathematics and Economics,* **107**, 161–179.

Godambe, V.P. (1960) An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, **31**(4), 1208–1211.

Guttorp, P. and Gneiting, T. (2006) Studies in the history of probability and statistics xlix on the matérn correlation family. *Biometrika*, **93**(4), 989–995.

Haug, O., Dimakos, X.K., Vårdal, J.F., Aldrin, M. and Meze-Hausken, E. (2011) Future building water loss projections posed by climate change. *Scandinavian Actuarial Journal*, **2011**(1), 1–20.

Heagerty, P.J. and Lumley, T. (2000) Window subsampling of estimating functions with application to regression models. *Journal of the American Statistical Association*, **95**(449), 197–211.

Henckaerts, R., Côté, M.-P., Antonio, K. and Verbelen, R. (2021) Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, **25**(2), 255–285.

Hua, L., Xia, M. and Basu, S. (2017) Factor copula approaches for assessing spatially dependent high-dimensional risks. *North American Actuarial Journal*, **21**(1), 147–160.

Huang, S., Zhang, J. and Zhu, W. (2024) Storm cat bond: Modeling and valuation. *North American Actuarial Journal*, **28**(4), 1–26.

Insurance Information Institute (2020a) Facts and statistics: Homeowners and renters insurance. [Online; accessed 21. Jun. 2022].

Insurance Information Institute (2020b) Severe convective storms: Evolving risks call for innovation to reduce costs, drive resilience.

Joe, H. (2014) *Dependence Modeling with Copulas*. New York, NY: CRC Press.

Joe, H. and Lee, Y. (2009) On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, **100**(4), 670–685.

Joe, H. and Xu, J.J. (1996) *The estimation method of inference functions for margins for multivariate models*. *Working paper*, Department of Statistics, University of British Columbia.

Krupskii, P. and Genton, M.G. (2017) Factor copula models for data with spatio-temporal dependence. *Spatial Statistics*, **22**, 180–195.

Krupskii, P., Huser, R. and Genton, M.G. (2018) Factor copula models for replicated spatial data. *Journal of the American Statistical Association*, **113**(521), 467–479.

Krupskii, P. and Joe, H. (2013) Factor copula models for multivariate data. *Journal of Multivariate Analysis*, **120**, 85–101.

Lee, G.Y. (2017) General insurance deductible ratemaking. *North American Actuarial Journal*, **21**(4), 620–638.

Lindsay, B.G. (1988) Composite likelihood methods. *Contemporary mathematics*, **80**(1), 221–239.

Lyubchich, V. and Gel, Y. (2017) Can we weather proof our insurance? *Environmetrics*, **28**(2), e2433.

Matérn, B. (1986) *Spatial Variation*. New York, NY: Springer Verlag.

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E. and Houston, T.G. (2012) An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, **29**(7), 897–910.

Nikoloulopoulos, A.K. and Joe, H. (2015) Factor copula models for item response data. *Psychometrika*, **80**, 126–150.

NOAA, Department of Defense, Federal Aviation Adminstration, and United States Navy (1998) *Automated Surface Observing System (ASOS) User's Guide*.

NOAA National Centers for Environmental Information (NCEI) (2022) Billion-dollar weather and climate disasters. [Online; accessed 21. Jun. 2022].

Oh, D.H. and Patton, A.J. (2017) Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics*, **35**(1), 139–154.

Parodi, P. (2014) *Pricing in General Insurance*. New York, NY: CRC Press.

Patton, A.J. (2006) Modelling asymmetric exchange rate dependence. *International Economic Review*, **47**(2), 527–556.

Pigeon, M., Antonio, K. and Denuit, M. (2013) Individual loss reserving with the multivariate skew normal framework. *ASTIN Bulletin: The Journal of the IAA*, **43**(3), 399–428.

Rüschendorf, L. (2009) On the distributional transform, sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, **139**(11), 3921–3927.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M. and Meze-Hausken, E. (2013) A bayesian hierarchical model with spatial variable selection: The effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**(1), 85–100.

Shi, P. (2014) Fat-tailed regression models. In *Predictive Modeling Applications in Actuarial Science* (eds. E. Edward, G. Meyers and R.A. Derrig). Cambridge: Cambridge University Press.

Shi, P., Feng, X., Boucher, J.-P. (2016) Multilevel modeling of insurance claims using copulas. *Annals of Applied Statistics*, **10**(2), 834–863.

Shi, P., Fung, G.M. and Dickinson, D. (2022) Assessing hail risk for property insurers with a dependent marked point process. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **185**(1), 302–328.

Shi, P. and Shi, K. (2017) Territorial risk classification using spatially dependent frequency-severity models. *ASTIN Bulletin: The Journal of the IAA*, **47**(2), 437–465.

Shi, P. and Zhao, Z. (2020) Regression for copula-linked compound distributions with applications in modeling aggregate insurance claims. *Annals of Applied Statistics*, **14**(1), 357–380.

Sklar, M. (1959) Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de l'Université de Paris*, **8**, 229–231.

Spekkers, M., Kok, M., Clemens, F. and Ten Veldhuis, J. (2013) A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences*, **17**(3), 913–922.

Swiss Re Institute (2021) Natural catastrophes in 2020: Secondary perils in the spotlight, but don't forget primary-peril risks.

Tufvesson, O., Lindström, J. and Lindström, E. (2019) Spatial statistical modelling of insurance risk: A spatial epidemiological approach to car insurance. *Scandinavian Actuarial Journal*, **2019**(6), 508–522.

Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica*, **11**(1), 5–42.

Wüthrich, M.V. (2018) Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, **2018**(6), 465–480.

Yang, L. and Shi, P. (2019) Multiperil rate making for property insurance using longitudinal data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **182**(2), 647–668.

Zhao, Z., Shi, P. and Feng, X. (2021) Knowledge learning of insurance risks using dependence models. *INFORMS Journal on Computing*, **33**(3), 1177–1196.