Hidden semi-Markov models for rainfall-related insurance claims

BY Yue Shi, Antonio Punzo, Håkon Otneim and
Antonello Maruotti

DISCUSSION PAPER

NHH

Institutt for foretaksøkonomi
Department of Business and Management Science

FOR 17/2023

ISSN: 2387-3000

November 2023

Hidden semi-Markov models for rainfall-related insurance claims

Yue Shi[1*], **Antonio Punzo**[2], **Håkon Otneim**[1], **Antonello Maruotti**[3]

[1]Department of Business and Management Science, Norwegian School of Economics
[2]Department of Economics and Business, University of Catania
[3]Department GEPLI, LUMSA University

September 6, 2023

Abstract

We analyze the temporal structure of a novel insurance dataset about home insurance claims related to rainfall-induced damage in Norway, and employ a hidden semi-Markov model to capture the non-Gaussian nature and temporal dynamics of these claims. By exploring a wide range of candidate distributions and evaluating their goodness-of-fit as well as commonly used risk measures, we identify a suitable model for effectively modeling insurance losses stemming from rainfall-related incidents. Our findings highlight the importance of considering the temporal aspects of weather-related insurance claims and demonstrate that the proposed hidden semi-Markov model adeptly captures this feature. Moreover, the model estimates reveal a concerning trend: the risks associated with heavy rain in the context of home insurance have exhibited an upward trajectory between 2004 and 2020, aligning with the evidence of a changing climate. This insight has significant implications for insurance companies, providing them with valuable information for accurate and robust modeling in the face of climate uncertainties. By shedding light on the evolving risks related to heavy rain and their impact on home insurance, our study offers essential insights for insurance companies to adapt their strategies and effectively manage these emerging challenges. It underscores the necessity of incorporating climate change considerations into insurance models and emphasizes the importance of continuously monitoring and reassessing risk levels associated with rainfall-induced damage. Ultimately, our research contributes to the broader understanding of climate risk in the insurance industry and supports the development of resilient and sustainable insurance practices.

Keywords: Mixtures, Non-Gaussian distributions, EM algorithm, Risk measures, Rainfall data
JEL Codes: C02, C40, C60

[*] **Corresponding author.** *E-mail address: Yue.Shi@nhh.no*

1. Introduction

The proper modeling of losses is of great relevance for insurance companies, as it ensures efficient management of the organization in calculating precise values for premiums, reserves, and risk measures, in accordance with regulatory frameworks such as Solvency II and Basel II/III (Brazauskas and Kleefeld, 2016).

However, the emergence of climate change has introduced new challenges. To effectively adapt to the rapidly changing climate, the European Union has initiated the Green Deal, aiming to achieve a climate-neutral Europe by 2050. As part of this initiative, the EU taxonomy has been implemented, providing clearly-defined and standardized criteria for sustainable business activities. It encourages companies to adopt more climate-friendly practices and requires large companies in the EU to assess and report their sustainability performance and alignment with the taxonomy. Consequently, insurance companies must consider climate change and incorporate forward-looking models into their operations.

The increasing frequency and severity of extreme weather events suggest that weather-induced losses and associated claims will rise. This necessitates the quantification of climate risk and its integration into financial risk models. However, risk teams in the insurance industry face challenges in adapting their models to address the uncertainties posed by climate change. Traditionally, they have relied on simplistic Gaussian models to analyze losses. However, it is increasingly evident that classical actuarial models may no longer suffice to adequately assess and manage risks in this context. Therefore, it is crucial to explore new statistical approaches that can effectively capture the complexities associated with climate-related risks.

The major challenge in modeling insurance losses lies in accurately identifying a probability density function that describes the typical distribution characteristics. These features can be summarized as follows: a positive support (Klugman et al., 2012), right skewness (Eling, 2012; Adcock et al., 2015), and heavy tails (Hogg and Klugman, 1983; Pigeon and Denuit, 2011). Regarding the modeling of right-skewed and heavy-tailed insurance losses, the literature has pursued three major approaches, aiming to achieve parsimonious yet sufficiently flexible models with excellent goodness-of-fit properties across the entire data support. The first approach focuses on introducing or investigating distributions to find the best fit for the available data, comparing a wide range of candidate distributions. Notable examples can be found in works by Vernic (2006); Klugman et al. (2012); Eling (2012, 2014); Bhati and Ravi (2018). Another approach aims to improve the tail behavior of the model by considering compound distributions (Cooray and Cheng, 2015; Punzo et al., 2018).

Recently, much attention has been devoted to the finite mixture approach (McLachlan and Peel, 2000). Finite mixture models offer another method for modeling heavy-tailed and skewed losses. These flexible (semi-)parametric models can capture the general features of loss data and account for multimodality resulting from heterogeneity. This approach has been explored in various publications within the actuarial literature. For instance, Bernardi et al. (2012) proposes finite mixtures of skew-normal distributions within a Bayesian analysis framework. Verbelen et al. (2015) develops finite mixtures of Erlang distributions and utilizes the EM algorithm for model estimation. Gómez-Déniz et al. (2013) suggests a gamma mixture with the generalized inverse Gaussian distribution to fit a well-known Danish fire dataset. Miljkovic and Grün (2016) extends the distribution of finite mixture models to more general forms, such as the Burr, Gamma, Inverse

Burr, Inverse Gaussian, lognormal, Weibull, and GB2 distributions (Chan et al., 2018). Other examples using frequentist (Punzo et al., 2018; Miljkovic and Fernández, 2018; Abu Bakar et al., 2018; Bignozzi et al., 2018) and Bayesian (Hong and Martin, 2017, 2018) paradigms are available in the literature.

Extending the existing finite mixture models in the actuarial science literature, we propose a novel approach for modeling time-series loss data with positive support using a non-Gaussian hidden semi-Markov model. This approach considers flexible one-, two-, three-, and four-parameter conditional distributions. To estimate the model parameters, we employ an ad-hoc Expectation-Maximization (EM) algorithm within a maximum likelihood framework. Specifically, we modify the Baum-Welch algorithm, which is a special case of the Expectation-Maximization (EM) algorithm widely used in the context of hidden Markov models (Zucchini et al., 2017). This work can also be viewed as an extension of the work by Bulla (2011), where the time spent in a certain hidden state of the semi-Markov chain can follow any count distribution (Bulla and Bulla, 2006; Barbu and Limnios, 2009; Bulla et al., 2010; Yu, 2015).

The statistical approach we propose is motivated by the analysis of heavy rainfall-related claims in home insurance products. Home insurance is among the insurance lines most strongly exposed to climate risk, and previous studies have highlighted the significance of rainfall as an indicator of home insurance risk (Gradeci et al., 2019; Cheng et al., 2012; Spekkers et al., 2013, 2014, 2015; Torgersen et al., 2015). As pointed out by Hanssen-Bauer et al. (2009) in their report "Climate in Norway 2100", heavy precipitation frequency and intensity are expected to increase due to climate change. Under the most severe climate projections from the Intergovernmental Panel on Climate Change (IPCC), the number of days with heavy precipitation in Norway is predicted to double, and annual precipitation is estimated to increase by approximately 18% by the end of this century.

By considering both the characteristics of insurance losses and the impact of climate change, our proposed non-Gaussian hidden semi-Markov model offers an innovative and flexible approach to address the complex nature of climate-related risks in insurance. We apply our approach to a novel insurance dataset from Norway. This dataset provides detailed information on insurance losses specifically caused by rainfall, enabling a comprehensive understanding of the underlying dynamics between heavy rain and insurance claims.

We aim to explore the implications of climate change on insurance and the importance of innovative approaches to modeling and managing climate risk. Our finding is consistent with the observed effects of climate change. In the following sections, we will present the modeling work and discuss the implications of our results for insurance companies in the context of climate change. The paper is organized as follows. We illustrate the Hidden semi-Markov models (HSMM) in Section 2 and further provide insights on maximum likelihood inference and show details on the implementation of the Baum-Welch version of the EM algorithm in Section 3. We then perform data analysis in Section 4. The conclusion with a discussion is set out in the last section.

2. Methodology

2.1. Hidden Semi-Markov models

In this section, we present an overview of hidden semi-Markov models (HSMMs) and their application in modeling heterogeneous and/or multimodal data. We start by defining the basic structure of an HSMM, which is a time-dependent mixture model that involves a finite-state semi-Markov chain and a sojourn-time distribution associated with each hidden state.

More specifically, let $\{Y_t; t = 1, \ldots, T\}$ denote a time series of length T, with each Y_t taking values in $\mathbb{R}^+$. Moreover, let $\{S_t; t = 1, \ldots, T\}$ be the underlying hidden process where S_t represents the state of membership of Y_t from a finite-state semi-Markov chain taking values on the state space $\{1, \ldots, k, \ldots, K\}$. The process $\{S_t; t = 1, \ldots, T\}$ is constructed as follows. A homogeneous Markov chain with K states models the transitions between different states, with initial probabilities $\pi_k = \Pr(S_1 = k)$ and transition probabilities

$$\pi_{k|j} = \Pr(S_{t+1} = k \mid S_t = j, S_{t+1} \neq j) \tag{1}$$

where $\sum_{k=1}^{K} \pi_{k|j} = 1$ and $\pi_{j|j} = 0$, i.e., the diagonal elements of the transition probability matrix are zeros. In (1), k refers to the current state, whereas j refers to the one previously visited; this convention will be used throughout the paper. We collect the initial probabilities in the K-dimensional vector $\boldsymbol{\pi}$, whereas the time-homogeneous transition probabilities are collected in the $K \times K$ transition matrix $\mathbf{\Pi}$.

The observation process is linked to the underlying hidden process by the emission distribution $b_k(y_t)$, which is the conditional distribution of Y_t given the unobserved (or hidden) state $S_t = k$, $k = 1, \ldots, K$, with $\sum_{y_t} b_k(y_t) = 1$, in the discrete case, and $\int_{y_t} b_k(y_t)\, dy_t = 1$ in the continuous case, $t = 1, \ldots, T$.

Furthermore, we explicitly model the sojourn-time distribution associated with each hidden state, which models the time the hidden process $\{S_t\}$ spends in the k-th state. More specifically, the sojourn-time distribution is defined as:

$$d_k(u) := \Pr(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0, \ldots, u-2 \mid S_{t+1} = k, S_t \neq k). \tag{2}$$

This distribution, which is state-specific, captures the duration of each state visit and is a key feature that distinguishes HSMMs from hidden Markov models (HMMs), where the sojourn-time distribution is implicitly geometric. The semi-Markovian hidden process S_t of an HSMM does not have the Markov property at each time t, but is Markovian at the times of state changes only.

2.1.1. Sojourn distributions

In analogy with O'Connell and Højsgaard (2011a), we adopt the shifted Poisson and gamma distributions as models for the sojourn distribution. Other commonly used sojourn distributions can be found in Zucchini et al. (2017, Chapter 12.6).

The shifted Poisson distribution has the probability mass function (pmf)

$$d_k(u; \lambda_k, \zeta_k) = \frac{e^{-\lambda_k} \lambda_k^{(u-\zeta_k)}}{(u-\zeta_k)!}, \quad u = \zeta_k, \zeta_k + 1, \ldots, \tag{3}$$

where $\lambda_k > 0$ and $\zeta_k \in \{1, 2, \ldots\}$ is the shift parameter. This parameter sets the minimum sojourn time in state k, $k = 1, \ldots, K$, and is an additional parameter with respect to the classical Poisson distribution. In practice, the shift parameter ζ_k is often fixed to 1 (see, e.g., Zucchini et al., 2017,

p. 178) and not estimated.

The gamma distribution has the probability density function (pdf)

$$d_k\left(u;\lambda_k,\zeta_k\right)=\frac{u^{\lambda_k-1}}{\Gamma\left(\lambda_k\right)\zeta_k^{\lambda_k}}\exp\left(-\frac{u}{\zeta_k}\right),\quad u>0, \tag{4}$$

where $\lambda_k > 0$ and $\zeta_k > 0$ are the shape and scale parameters, respectively.

2.1.2. Emission distributions

We consider several parametric members of the GAMLSS (generalized additive models for location scale and shape; Stasinopoulos et al., 2017) family of distributions. All the models of the family are defined with a maximum of 4 parameters, say μ, σ, ν, and τ, and with a probability density function (pdf) that can be denoted by $b\left(y;\mu,\sigma,\nu,\tau\right)$. In particular, we consider the following 11 candidate distributions for $b\left(y;\mu,\sigma,\nu,\tau\right)$: exponential (EXP), gamma (GA), log-normal (LOGNO), inverse Gaussian (IG), Weibull (WEI), Box-Cox Cole and Green (BCCG), generalized gamma (GG), generalized inverse Gaussian (GIG), Box-Cox t (BCT), Box-Cox power exponential (BCPE) and Generalized beta type 2 (GB2); for details about these distributions, see Rigby et al. (2014) and Stasinopoulos and Rigby (2017). These one-, two-, three- and four-parameter parametric distributions are commonly employed in modeling loss data and are thus used as basic building blocks to generate more flexible distributions by incorporating them into the HSMM framework.

2.2. Marginal risk measures: VaR and ES

The Value-at-Risk (VaR) is a widely used risk measure for a risky asset, defined as the $(1-\alpha)$-quantile of the asset return distribution at a given confidence level α. It represents the minimum loss in the worst $1-\alpha$ cases (Jorion, 1997). Mathematically, for a random variable Y with an absolutely continuous density function $f(y)$ and cumulative density function $F(y)$, the VaR at level $\alpha \in (0,1)$ can be calculated as $\text{VaR}_\alpha(Y) \equiv F^{-1}(1-\alpha)$. However, the VaR only reflects the maximal probable losses at a $\alpha \times 100\%$ chance and does not provide sufficient information about the tail behavior of a distribution. Hence, VaR may not provide a proper measurement of risk if the distribution has a heavy tail and if extreme losses are of major concern.

To address this limitation, Acerbi (2002) introduced spectral risk measures, with the Expected Shortfall (ES) being a popular choice. The ES represents the Conditional Tail Expectation (CTE) of Y conditioned on its VaR level. Specifically, it is expressed as $\text{ES}(Y) \equiv \text{CTE}[\text{VaR}_\alpha(Y)] = E[Y \mid Y > \text{VaR}_\alpha(Y)]$. It provides a more conservative measure of risk compared to VaR at the same confidence level α and is effective in analyzing the tail behavior of the distribution (Nadarajah et al., 2014). Since it has been widely recognized as a coherent measurement of risk in situations with a non-normal distribution of losses, CTE, also known as "Tail VaR" or "expected tail loss", is adopted by financial regulators in the U.S. and Canada as capital standards for the insurance industry.

In addition, it is worth noting that Bernardi (2013) demonstrated that for a mixture model, the ES can be expressed as a convex linear combination of component-specific ESs. This property of the ES is further explored in the context of multiple risk components by Bernardi et al. (2017).

3. Likelihood inference

To estimate $\boldsymbol{\vartheta}$, the vector of all parameters of the proposed HSMM, we utilize a version of the expectation-maximization (EM) algorithm (Baum et al., 1970), although other alternatives are available in the literature (see, e.g., Rydén, 2008; Bulla and Berzel, 2008; MacDonald, 2014). Once the number of latent states K has been assigned or fixed, the algorithm operates based on the complete-data likelihood. The complete data likelihood of the HSMM is defined as:

$$\mathcal{L}_c\left(\boldsymbol{\vartheta}\right) = \pi_{s_1^*} d_{s_1^*}\left(u_1\right) \left\{ \prod_{r=2}^{R-1} \pi_{s_r^* | s_{r-1}^*} d_{s_r^*}\left(u_r\right) \right\} \pi_{s_R^* | s_{R-1}^*} D_{s_R^*}\left(u_R\right) \prod_{t=1}^{T} b(y_t \mid S_t = k; \mu_k, \sigma_k, \nu_k, \tau_k) \quad (5)$$

where π, d and b represent the transition probability, the sojourn-time distribution and the emission distribution, respectively. s_r^* represents the rth visited state, u_r denotes the time spent in that state (i.e., the duration of the rth visit), and $R-1$ is the number of state changes up to time T. Guédon (2003) proposed using the survivor function

$$D_k\left(u\right) = \sum_{v \geq u} d_k\left(v\right). \quad (6)$$

Working on $\mathcal{L}_c\left(\boldsymbol{\vartheta}\right)$ in (5), the EM iterates the following steps, until convergence:

E-step: compute the conditional expected value of the complete-data log-likelihood given the observed data and $\boldsymbol{\vartheta}^{(h)}$, the current estimate of $\boldsymbol{\vartheta}$ at the hth iteration; and

M-step: maximize the preceding expected value with respect to $\boldsymbol{\vartheta}$.

In the E-step, we calculate the expected complete data log-likelihood given the observed data as:

$$Q\left(\boldsymbol{\vartheta} | \boldsymbol{\vartheta}^{(h)}\right) = E_{\boldsymbol{\vartheta}^{(h)}}\left\{ \ln\left[\mathcal{L}_c\left(\boldsymbol{\vartheta}\right)\right] \mid y_1, \ldots, y_T \right\}, \quad (7)$$

where the expectation operator E with the subscript $\hat{\boldsymbol{\vartheta}}$ indicates that this expectation is computed using $\hat{\boldsymbol{\vartheta}}$ for $\boldsymbol{\vartheta}$. The E-step involves the calculation of three quantities:

1. the probability
$$\gamma_t^{(h)}\left(k\right) = P_{\boldsymbol{\vartheta}^{(h)}}\left(S_t = k \mid y_1, \ldots, y_T\right) \quad (8)$$
of being in state k at time t given the observed sequence;

2. the probability
$$\xi_t^{(h)}\left(j, k\right) = P_{\boldsymbol{\vartheta}^{(h)}}\left(S_{t-1} = j, S_t = k \mid y_1, \ldots, y_T\right) \quad (9)$$
that the process left state j at time $t-1$ and entered state k at t given the observed sequence;

3. the expected number of times a process spends u time steps in state k,
$$\eta_t^{(h)}(k, u) = P_{\boldsymbol{\vartheta}^{(h)}}\left(S_{t-u+1} = k, \ldots, S_t = k \mid y_1, \ldots, y_T\right). \quad (10)$$

Therefore, Equation (7) can be expressed as:

$$Q\left(\boldsymbol{\vartheta} \mid \boldsymbol{\vartheta}^{(h)}\right) = \sum_{k=1}^{K} \gamma_1^{(h)}(k) \log \pi_k \tag{11}$$

$$+ \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{k \neq j} \frac{\xi_t^{(h)}(j,k)}{\gamma_t^{(h)}(j)} \log \pi_{k|j} \tag{12}$$

$$+ \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{u=1}^{U} \frac{\eta_t^{(h)}(k,u)}{\gamma_t^{(h)}(k)} \log d_k(u) \tag{13}$$

$$+ \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t^{(h)}(k) \log b(y_t \mid S_t = k; \mu_k, \sigma_k, \nu_k, \tau_k). \tag{14}$$

The M-step returns the maximum likelihood estimates for parameters as follows:

$$\pi_k = \frac{\gamma_1(k)}{\sum_{k=1}^{K} \gamma_1(k)} \tag{15}$$

$$\pi_{k|j} = \frac{\sum_{t=1}^{T} \xi_t(j,k)}{\sum_{k \neq j} \sum_{t=1}^{T} \xi_t(j,k)}. \tag{16}$$

The maximization of the sojourn distribution parameters depends on the chosen sojourn distribution. If the shifted Poisson in considered, for each state k, $k = 1, \ldots, K$, λ_k is estimated for each possible shift parameter value $\zeta_k \in \left\{1, \ldots, \min\left(u : \eta_{ku}^{(h)} > 0\right)\right\}$ by classical point estimation procedures from the quantities $\eta_{ku}^{(h)}$ (Johnson et al., 2005). The pair (λ_k, ζ_k) which gives the maximum value of (13), say $\left(\lambda_k^{(h+1)}, \zeta_k^{(h+1)}\right)$, is retained. Guédon (2003) states that this *ad hoc* procedure works well in practice and this statement is further corroborated by O'Connell and Højsgaard (2011b) via simulations. If the gamma is considered as sojourn distribution, then the maximization of (13) is performed following the methodology of Choi and Wette (1969) which is implemented by the `gammafit()` function of the R package **mhsmm** (O'Connell and Højsgaard, 2011b).

To maximize (14) we use the nonlinear maximization algorithms `optim()` and `nlminb()` available in the R package **gamlss**, which improves the fitting speed and allows us to obtain parameters estimates even if a closed form solution of the M-step equation is not available. For details about these algorithms, see Stasinopoulos et al. (2023).

3.1. Operational aspects: algorithm initialization, standard errors, and model selection

When implementing a hidden semi-Markov model (HSMM), certain operational aspects require careful attention. In the following we outline and discuss some of these aspects:

First, regarding the initialization of the EM Algorithm, the choice of initial values for the EM algorithm in mixture models is crucial. In our HSMM with K states, we initialize the EM algorithm using the approach described in Maruotti and Punzo (2021). Initially, we obtain a state partition using the K-means method, assuming the independence of the observations. From this partition, the off-diagonal elements of the transition probability matrix are computed as transition proportions. The initial parameters for the sojourn distribution are derived by computing the sequence of sojourn times in state k from the initial state sequence, and the sojourn probabilities are obtained from the corresponding relative frequencies. Finally, the initial parameters for the observed conditional distribution are estimated using the maximum likelihood method on the observations in state k.

Second, the EM algorithm may converge to local maxima of the log-likelihood function or encounter singularities at the edge of the parameter space, leading to unbounded log-likelihood values. To address convergence, we employ the Aitken acceleration method as a criterion to determine whether the log-likelihood is sufficiently close to its estimated asymptotic value. The Aitken acceleration at the hth iteration is computed by

$$A^{(h)} = \frac{l^{(h+1)} - l^{(h)}}{l^{(h)} - l^{(h-1)}}, \tag{17}$$

where $l^{(h+1)}$, $l^{(h)}$, and $l^{(h-1)}$ represent the log-likelihood at iterations $h+1$, h, and $h-1$, respectively. The asymptotic estimate of the log-likelihood at iteration $h+1$ is then obtained as

$$l_{\infty}^{(h+1)} = l^{(h)} + \frac{1}{1 - A^{(h)}}(l^{(h+1)} - l^{(h)}). \tag{18}$$

Convergence is considered achieved when the absolute difference between $l_{\infty}^{(h+1)}$ and $l_{\infty}^{(h)}$ is smaller than a predefined threshold ϵ.

Third, when maximizing the likelihood with respect to the parameters, it is necessary to address technical issues such as numerical underflow. For detailed information on how to handle these problems, refer to Zucchini et al. (2017).

It is crucial to underscore that while initialization strategies, standard errors computation and convergence to local maxima are pivotal operational aspects when implementing an HSMM, they represent merely the tip of the iceberg. Additional contemplations might become imperative, contingent upon the intricate interplay between the model's unique requisites and the idiosyncrasies of the dataset under examination. For example, depending on the dataset size and complexity, computational efficiency becomes significant. Techniques to optimize computations, such as parallel processing or specialized algorithms for large datasets, might be required. Simlarly, applying appropriate regularization techniques might prevent overfitting while ensuring model flexibility is a delicate balance.

4. Model application

To illustrate the flexibility of the proposed HSMM for data on positive support, it is now applied to the time series of home insurance claims caused by rainfall-induced water damage in Norway. The fitted model is then used to compute the VaR and the CTE, and these two estimates are compared with the empirical counterparts.

4.1. Data description

In this section, we study rainfall-related insurance claims in Norway. Insurance data are provided by one of the largest Norwegian insurance companies. We have policy and claim data for all private buildings insured by the company in Norway for the period between 2004 and 2020. This novel data set has a relatively long time span. The claim amounts are aggregated weekly over all municipalities in Norway and the value is normalized in Euro (EUR) 2021[1]. The number of observations is $n = 888$. For a fair comparison between years, we use the "loss costs (in euro)", which are the total claim amounts divided by the number of policies as the variable of interest (Y) in our analysis. The weekly temporal evolution of loss costs between 2004 and 2020 is displayed in Figure 1.

In Norway, claims for damage caused by defined natural hazards are covered by a national insurance pool called Norwegian Natural Perils Pool[2]. When a claim caused by weather events is made to the insurance companies, it will be compensated under this insurance scheme if the damage is deemed to be incurred by natural perils. Claims for damage by extreme weather events at a local scale, which is not covered to this natural perils insurance scheme, are then covered by insurance companies. Evidence based on long-term observations of weather and insurance records shows that local extreme weather events generally cause more insurance losses than large natural disasters (Lyubchich et al., 2019). Torgersen et al. (2015) also point out that in Norway, only 4% of water-related claims were defined as natural hazards from 2008 to 2011. Thus, we only focus on non-catastrophic weather events covered by the insurance company in this study.

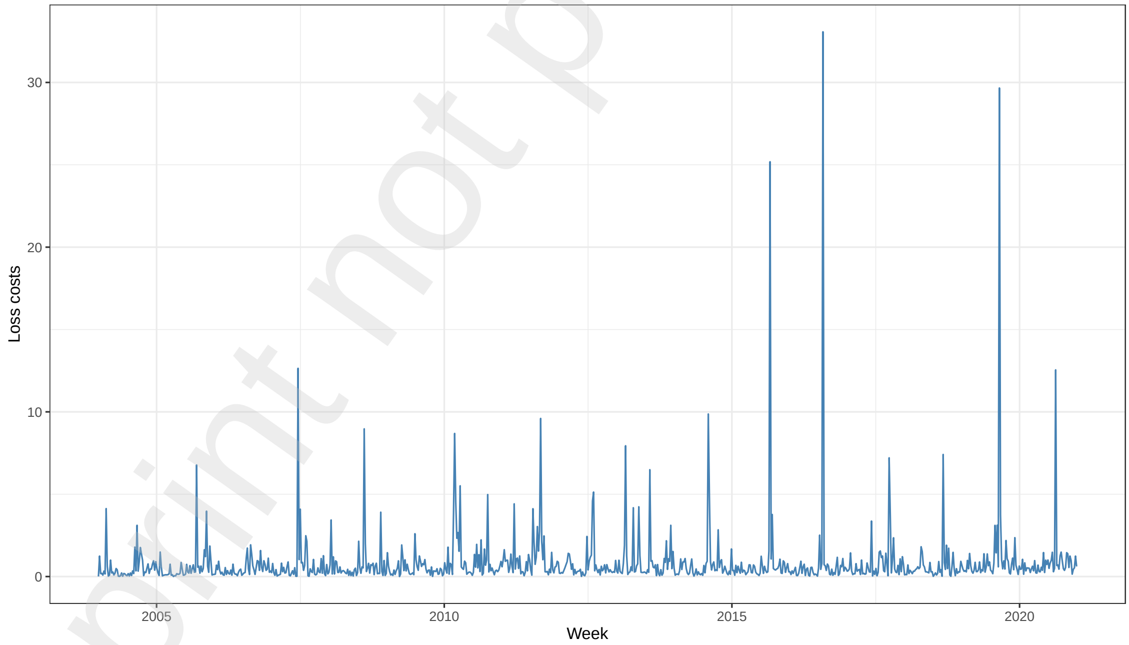

Figure 1. Weekly time series of loss costs (in euro) in rainfall-related home insurance claims across Norway between 2004–2020

Table 1 displays some quantiles from the empirical distribution of loss costs in rainfall-related

[1] According to the average exchange rate in 2021 provided by European Central Bank, EUR 1 = NOK 10.1633.

[2] "Norsk Naturskadepool" in Norwegian.

home insurance claims over all municipalities in Norway during the period 2004–2020. Obviously, the distribution is non-symmetric. The histogram (Figure 2) further shows that the empirical distribution is right-skewed, which is quite common in insurance (Furman, 2008 and Jeon and Kim, 2013). The distribution of insurance claims often has a heavy tail since an insurance company can experience incidents that have a low probability of occurrence but lead to extremely large losses (Ahn et al., 2012, Abu Bakar et al., 2015, and Tomarchio and Punzo, 2020). Correspondingly, incidents that result in smaller claims happen more often and therefore have a higher probability of occurrence. The positive support of the data suggests that classical real-valued distributions, like the Gaussian and t, are not suitable to describe the distribution of claim amounts since these models cause boundary bias, that is, allocation of probability mass outside the theoretical support (see, e.g., Bagnato and Punzo, 2013, Mazza and Punzo, 2014, 2015, and Tomarchio and Punzo, 2019).

Table 1. Some quantiles from the empirical distribution of loss costs (in euro) in Norway between 2004–2020

Quantiles	Loss costs
Min.	0.003
0.25	0.208
0.5	0.414
0.75	0.819
0.9	1.449
0.99	8.038
0.999	30.050
Max.	33.076

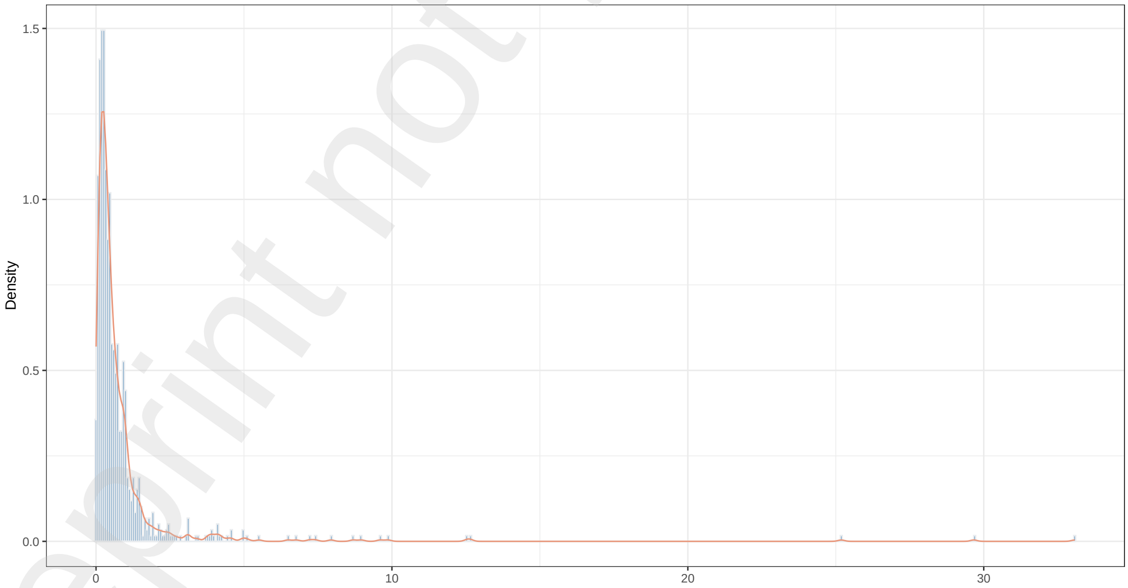

Figure 2. Histogram of loss costs in rainfall-related home insurance claims across Norway between 2004–2020. The red curve is the kernel density estimate – a smoothed version of the histogram.

4.2. Model fitting and selection

We fit the proposed HSMM to the weekly amounts of rainfall-related home insurance claims in Norway. For the sojourn distribution, initial attempts indicate that Gamma sojourns always perform better than the shifted Poisson sojourns [3]. Thus, we adopt the Gamma distribution as the sojourn-time distribution in this analysis.

Having the candidate models a differing number of parameters, we compare their goodness-of-fit via the Akaike information criterion (AIC; Akaike, 1974) that, in our formulation, is defined as

$$\text{AIC} = 2\ln\left[\mathcal{L}\left(\hat{\boldsymbol{\vartheta}}\right)\right] - 2m, \tag{19}$$

where m is the number of estimated parameters in the model and $\mathcal{L}\left(\hat{\boldsymbol{\vartheta}}\right)$ is the maximized value of the (observed-data) likelihood function. Under this formulation, the preferred model is the one with a higher AIC value.

4.2.1. Fitting the model to the whole period

We first fit the HSMMs having the gamma as the sojourn distribution to the full data set from 2004 to 2020 and examine different emission distributions for a different number of hidden states (K) with AIC and log-likelihood (LL). Results – in terms of the total number of parameters (m), maximized (observed-data) log-likelihood (LL), and AIC – are shown in Tables 2, 3, 4, and 5 for values of K ranging from 1 to 4.

Table 2. Total number of parameters (m), maximized (observed-data) log-likelihood (LL), and AIC for HSMMs having the gamma as sojourn distribution and varying in terms of emission distribution. The table refers to the case $K = 1$.

Emission distribution	m	LL	AIC
EXP	3	-731.28	-1464.56
GA	4	-716.95	-1437.89
LOGNO	4	-574.16	-1152.32
IG	4	-697.79	-1399.59
WEI	4	-678.29	-1360.59
BCCG	5	-573.72	-1153.43
GG	5	-573.76	-1153.52
GIG	5	-658.37	-1322.75
BCT	6	-552.07	-1112.14
BCPE	6	-555.05	-1118.10
GB2	6	-551.66	-1111.32

Generally, we can see that the result, in terms of LL, improves as the number of hidden states rises but it also leads to more computation. Taking these factors into consideration, among the HSMMs having the Gamma as the sojourn distribution, the AIC chooses the GB2 distribution for the emission distribution and $K = 3$ hidden states. The estimated parameters for the best-AIC model are summarized in Table 6. The density plots for estimated emission distributions and sojourn distributions are displayed in Figure 3 and Figure 4, respectively. According to the mean and 95% quantile for emission distributions in Table 7, the order of the three states from

[3] Results for the HSMMs having the shifted Poisson as the sojourn distribution can be found in Tables 14, 15, 16, and 17 of Appendix A.1.

Table 3. Total number of parameters (m), maximized (observed-data) log-likelihood (LL), and AIC for HSMMs having the gamma as sojourn distribution and varying in terms of emission distribution. The table refers to the case $K = 2$.

Emission distribution	m	LL	AIC
EXP	4	-561.89	-1137.78
GA	6	-540.99	-1099.97
LOGNO	6	-538.07	-1094.14
IG	6	-557.22	-1132.44
WEI	6	-542.16	-1102.32
BCCG	8	-568.33	-1158.66
GG	8	-508.70	-1039.40
GIG	8	-509.38	-1040.77
BCT	10	-508.04	-1042.09
BCPE	10	-532.67	-1091.33
GB2	10	-506.38	-1038.75

Table 4. Total number of parameters (m), maximized (observed-data) log-likelihood (LL), and AIC for HSMMs having the gamma as sojourn distribution and varying in terms of emission distribution. The table refers to the case $K = 3$.

Emission distribution	m	LL	AIC
EXP	5	-541.82	-1111.64
GA	8	-516.01	-1066.01
LOGNO	8	-495.19	-1024.38
IG	8	-517.58	-1069.15
WEI	8	-517.28	-1068.56
BCCG	11	-555.61	-1151.22
GG	11	-489.41	-1018.82
GIG	11	-493.32	-1026.64
BCT	14	-496.60	-1039.20
BCPE	14	-495.45	-1036.89
GB2	14	-485.04	-1016.08

Table 5. Total number of parameters (m), maximized (observed-data) log-likelihood (LL), and AIC for HSMMs having the gamma as sojourn distribution and varying in terms of emission distribution. The table refers to the case $K = 4$.

Emission distribution	m	LL	AIC
EXP	6	-540.83	-1127.66
GA	10	-496.85	-1047.71
LOGNO	10	-487.27	-1028.55
IG	10	-488.87	-1031.73
WEI	10	-505.92	-1065.83
BCCG	14	-504.43	-1070.85
GG	14	-484.41	-1030.83
GIG	14	-487.40	-1036.80
BCT	18	-479.82	-1029.65
BCPE	18	-487.60	-1045.20
GB2	18	-477.14	-1024.28

high to low risk is State 1, State 3, and State 2. Furthermore, we carry out a comparison to the actual empirical values of the mean and the quantiles within each of the predicted states. A good enough agreement between empirical end estimated values of mean and 95% quantile appears by comparing Table 7 with Table 8.

Table 6. Estimated parameters for the best-AIC HSMM (gamma sojourn, GB2 emission distribution, and $K = 3$)

Parts of the HSMM		State 1	State 2	State 3
Initial probabilities ($\boldsymbol{\pi}$)		0.000	1.000	0.000
Transition matrix ($\boldsymbol{\Pi}$)	State 1	0.000	0.000	1.000
	State 2	1.000	0.000	0.000
	State 3	0.276	0.724	0.000
Sojourn distributions (Gamma)	λ_k	13.181	5.742	7.036
	ζ_k	0.757	3.071	3.361
Emission distributions (GB2)	μ_k	0.882	0.430	0.267
	σ_k	2.615	1.324	0.544
	ν_k	0.466	1.028	12.458
	τ_k	0.459	2.175	9.741

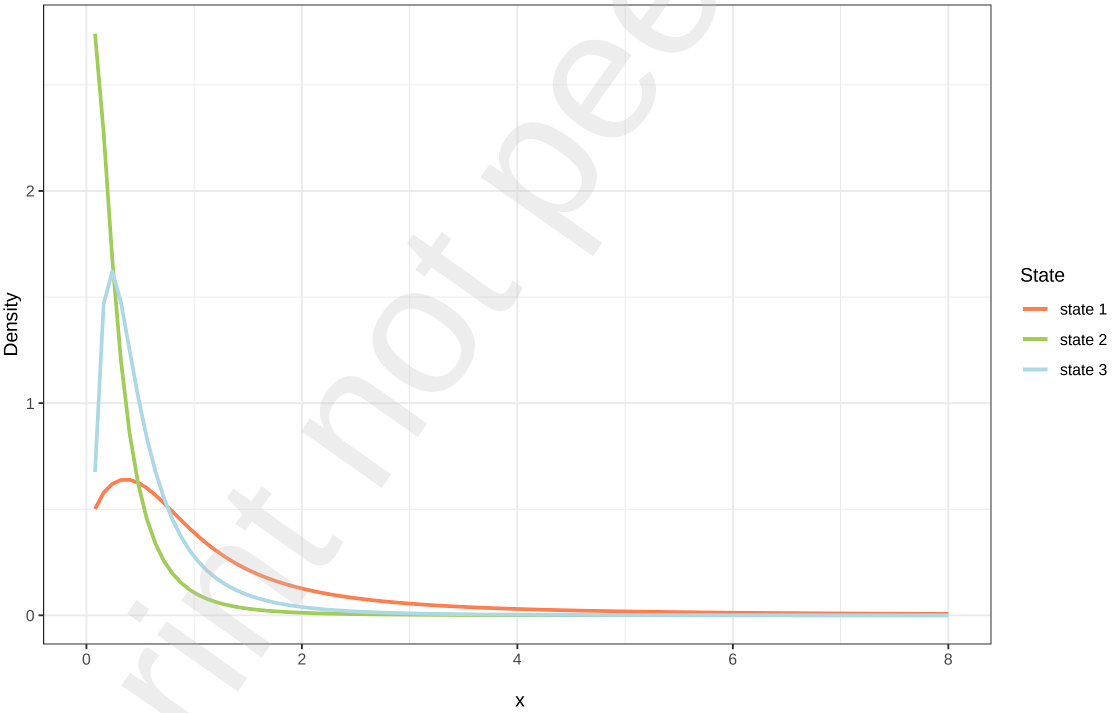

Figure 3. Density plots for estimated emission distributions

Table 7. Estimated mean and 95% quantile from the emission distributions for the period 2004–2020

	State 1	State 2	State 3
Mean	3.363	0.330	0.597
95% quantile	7.211	0.994	1.628

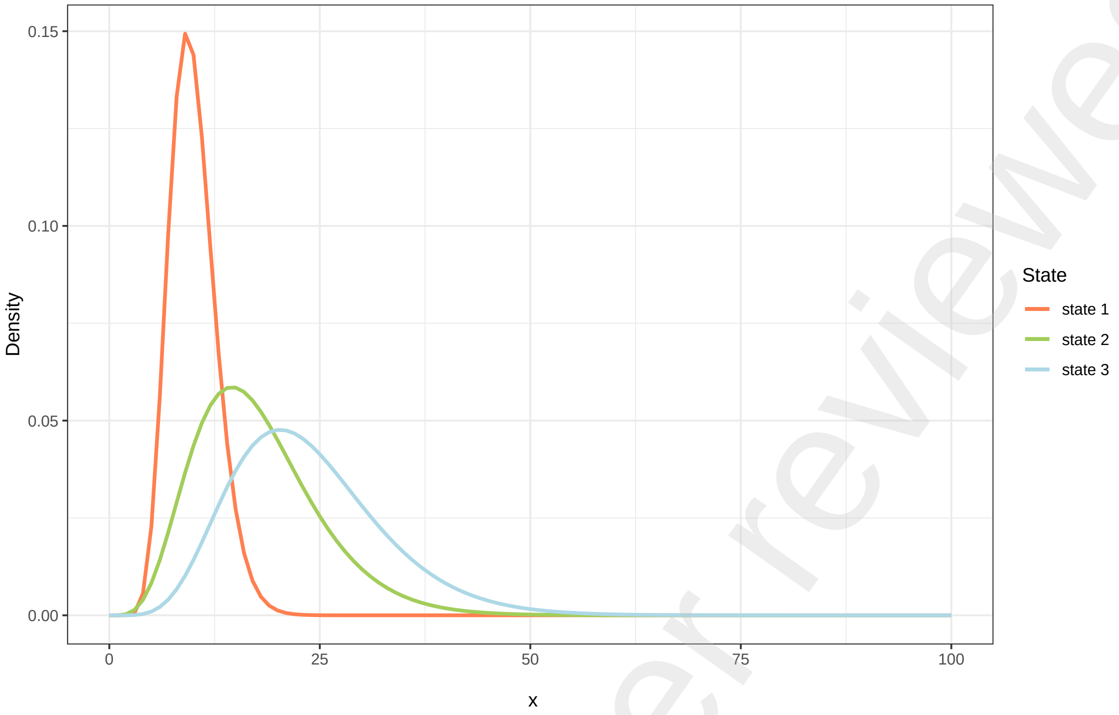

Figure 4. Density plots for estimated sojourn distributions

Table 8. Actual empirical mean and 95% quantile for the predicted states for the period 2004–2020

	State 1	State 2	State 3
Mean	2.140	0.317	0.595
95% quantile	7.860	0.857	1.400

Figure 5 clearly demonstrates that our proposed model is a good choice for modeling rainfall-related insurance claims. Almost all large claims are captured by state 1, which is the riskiest state. At the same time, we can see that the model contains two fairly similar states (states 2 and 3) for normal claim sizes, and state 3 gradually takes over. Since state 3 corresponds to slightly higher risk than state 2, this fact indicates that there is a development over time towards increased risk.

It is well known that climate change leads to more heavy rain in Norway, which in turn causes a lot of damage to private property. According to the report *Climate in Norway 2100*, it is projected that under the RCP8.5 scenario - the most severe climate projections from the Intergovernmental Panel on Climate Change (IPCC), the number of days with heavy precipitation in Norway will double, and the annual precipitation will increase by approximately 18% by the end of this century. Kundzewicz et al. (2017) point out that high-intensity rainfall is causing large damage to infrastructure and buildings in Norway.

Our finding adds new evidence to the literature that there is a sign of growing risk in weather-related insurance claims due to climate change. Motivated by the finding that there might be a shift over time for the distribution of normal claims, we split the long time series into two subperiods (2004–2012 and 2013–2020) and fit the model with two hidden states to the first and last half of time series respectively.

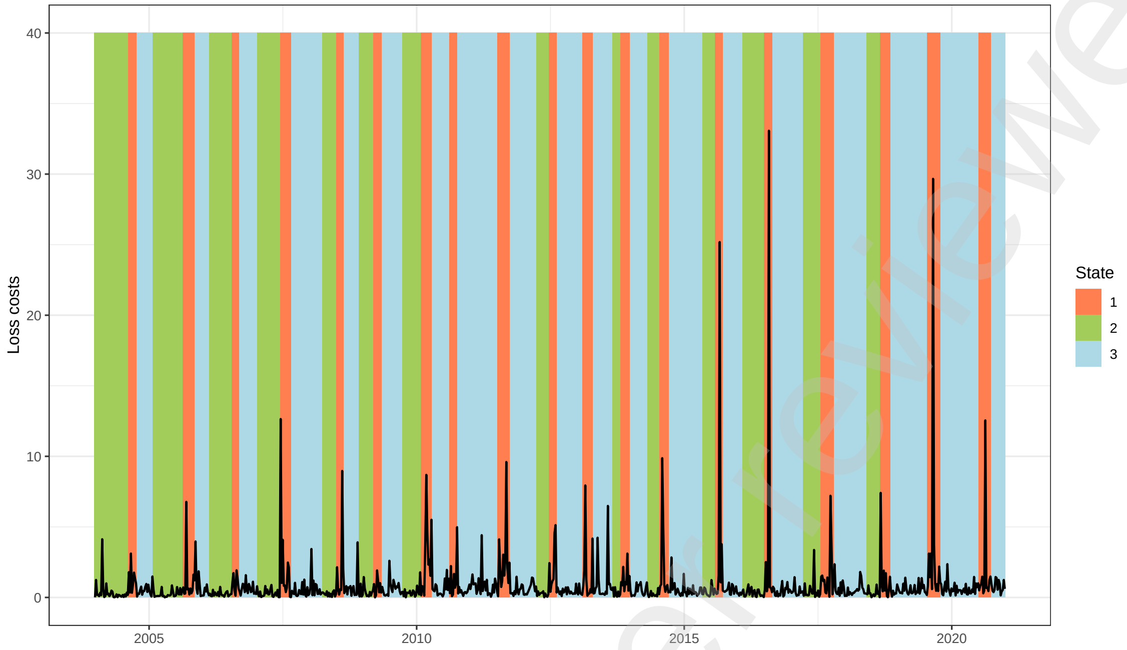

Figure 5. Time series plot with estimated states in the background (2004–2020)

4.2.2. Fitting the model to two sub-periods

To further confirm our finding above, we conduct the same analysis on the two sub-periods as we do above for the entire period, using the Gamma distribution as the sojourn distribution and the Generalized Beta type 2 distribution as the emission distribution. We employ a model with two hidden states: one for normal claims and one for extremes. Again, we find that the high-risk state continues to effectively capture all large claims.

For the period 2004–2012, according to the mean and 95% quantile for emission distributions in Table 9, state 1 corresponds to higher risk than state 2. From Figure 6 we can find that state 1 is visited more frequently during the period 2004–2012, and almost all extreme observations are captured by state 1. For the period 2013–2020, as shown in Table 10, state 1 is riskier than state 2. Looking at Figure 7, all large claims appear in this high-risk state. In addition, we also discover that risk has been increasing over time, not just for normal claims, but also for extreme claims. This finding supports our conclusion in the previous section when fitting the model to the whole period, which highlights the need for insurers to continually update their risk management strategies to adapt to the changing climate.

4.3. Computation of the risk measures

To investigate the appropriateness of the fitted model in reproducing empirical risk measures, we compute the Value at Risk (VaR) and conditional tail expectation (CTE), using the estimated parameters, at the 90%, 95%, and 99% confidence levels. For the entire period 2004–2020, according to Table 6 in Section 4.2.1, we generate a test sequence of length 1,000,000 from a three-state

Table 9. Estimated mean and 95% quantile from the emission distributions (left) and actual empirical mean and 95% quantile for the period 2004–2012 for the predicted states (right) for the period 2004–2012

	State 1	State 2
Mean	1.324	0.366
95% quantile	4.116	1.047

	State 1	State 2
Mean	1.360	0.338
95% quantile	4.520	0.922

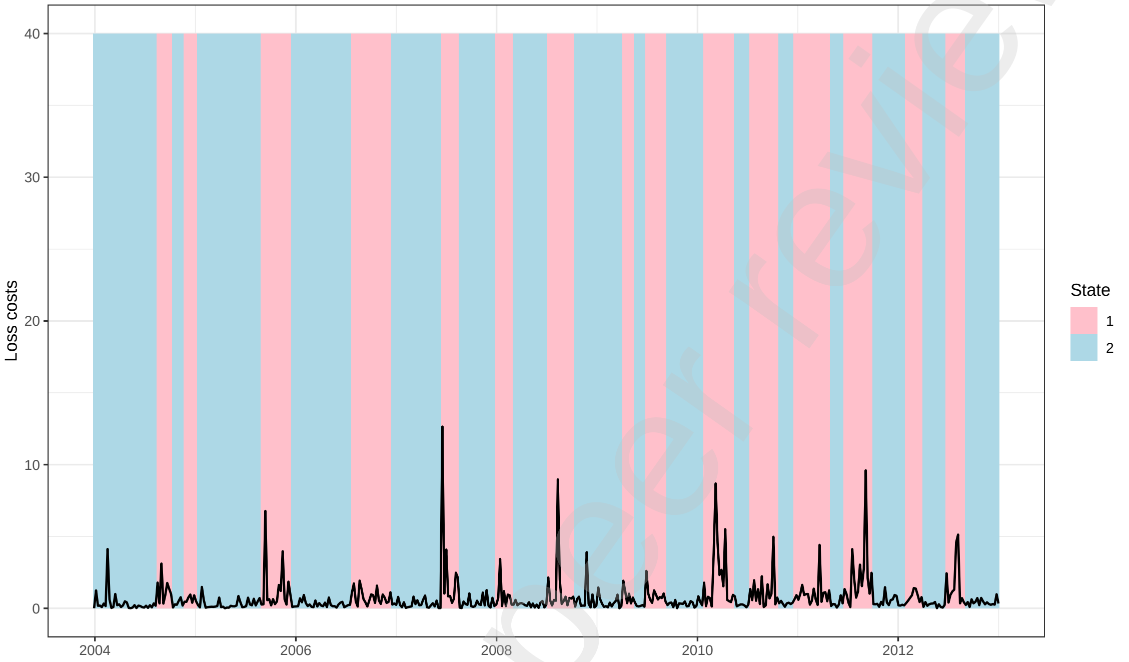

Figure 6. Time series plot with estimated states in the background (2004–2012)

HSMM with the estimated GB2 distributions. Table 11 summarizes the estimation results from the test series and the empirical value from the true series. Generally, we can see that our model estimates the risk measures close to the empirical values, but we also find a discrepancy between the estimated 99% CTE and the empirical value. In addition, we calculate these two risk measures using simulated data of the same length from three-state HSMMs with the other considered emission distributions. We can easily detect that compared to GB2 distribution, the other candidates give a more conservative estimation at the higher risk level (see 99% VaR and 99% CTE). This indicates that GB2 can be a better choice if extreme losses are of major concern and/or if insurance companies prefer to adopt a safer risk management strategy. For further reference, estimated results using GB2 distribution and the empirical values for the two sub-periods are given in Table 12 and Table 13, respectively.

Table 10. Estimated mean and 95% quantile from the emission distributions (left) and actual empirical mean and 95% quantile for the predicted states (right) for the period 2013–2020

	State 1	State 2
Mean	3.231	0.418
95% quantile	6.152	1.094

	State 1	State 2
Mean	1.970	0.390
95% quantile	6.960	0.975

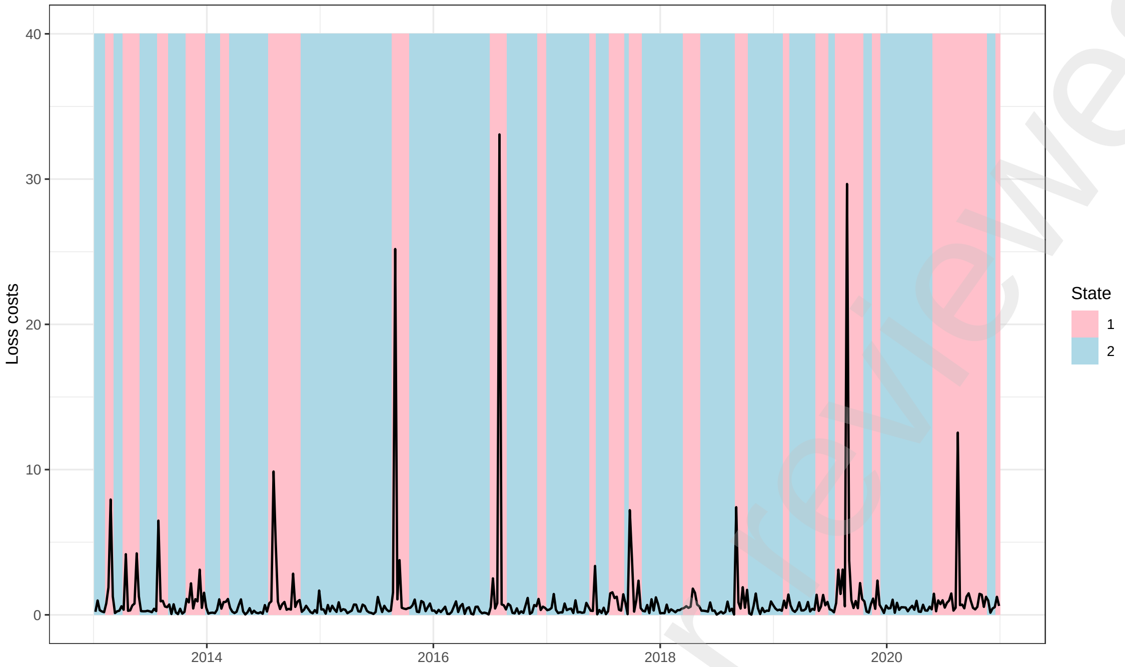

Figure 7. Time series plot with estimated states in the background (2013–2020)

Table 11. Risk measures for the period 2004–2020

	VaR			CTE		
	90%	95%	99%	90%	95%	99%
Empirical Value	1.449	2.475	8.038	4.297	6.716	16.694
GB2	1.478	2.165	6.461	6.339	10.922	41.382
EXP	1.492	1.768	2.365	1.877	2.138	2.714
GA	1.504	1.825	2.584	1.971	2.296	3.054
LOGNO	1.563	2.139	4.203	2.607	3.107	4.278
IG	1.651	2.209	3.832	2.386	2.780	3.674
WEI	1.530	1.932	2.959	2.111	2.459	3.267
BCCG	1.617	2.140	4.139	2.730	3.380	5.099
GG	1.495	2.046	4.226	2.761	3.353	4.799
GIG	1.474	2.151	5.021	2.964	3.535	4.750
BCT	1.419	2.182	6.381	4.579	5.668	8.231
BCPE	1.552	2.125	3.905	2.471	2.895	3.853

Table 12. Risk measures for the period 2004–2012

	VaR			CTE		
	90%	95%	99%	90%	95%	99%
Empirical Value	1.488	2.474	5.889	3.452	5.001	9.336
GB2	1.425	1.997	4.141	2.613	3.560	7.049

5. Concluding Remarks

This study contributes to the literature in three significant aspects. Firstly, we utilize a novel insurance dataset on rainfall-related claims obtained from a leading insurance company in Norway. This unique dataset covers private building claims from all municipalities in Norway, spanning a

Table 13. Risk measures for the period 2013–2020

	VaR			CTE		
	90%	95%	99%	90%	95%	99%
Empirical Value	1.426	2.392	9.561	5.194	8.676	22.069
GB2	1.528	2.441	8.846	7.506	13.127	49.098

long time period of 17 years. As the dataset has not been previously used for research purposes, it presents an exceptional opportunity for studying the impact of rainfall on insurance claims and offers a new perspective for investigating the effects of weather-related claims.

Secondly, we propose an effective model for modeling rainfall-related insurance claims, which is of great importance to the insurance industry. Although our study only focuses on insurance losses caused by rainfall, the methods could also be applied to other weather-related claims. The model can provide a more accurate estimation of potential insurance losses and can be used to improve risk management strategies to deal with the pressure posed by the fast-changing climate.

Lastly, our model estimates suggest that the risks associated with heavy rain in home insurance appear to have been increasing during the period 2004–2020, which is consistent with a changing climate. The findings offer useful insights to insurance companies for accurate modeling in the face of climate change. By incorporating these findings into their risk management framework, insurance companies can better prepare for future losses and minimize the impact of weather events on insured buildings.

Acknowledgments

This work has been supported by the Norwegian Research Council through the Climate Futures research center, grant number 309562. It has been partially supported by MUR, grant number 2022XRHT8R - *The SMILE project: Statistical Modelling and Inference to Live the Environment.* This research also contributes to the PNRR GRInS Project.

Appendix

A.1. HSMMs having the shifted Possion as sojourn distribution.

Here are the results of the total number of parameters (m), maximized (observed-data) log-likelihood (LL), and AIC for HSMMs having the shifted Poisson as sojourn distribution, when the number of hidden states (K) ranges from 1 to 4.

Table 14. The table refers to the case $K = 1$.

Emission distribution	m	LL	AIC
EXP	3	-731.28	-1464.56
GA	4	-716.95	-1437.89
LOGNO	4	-574.16	-1152.32
IG	4	-697.79	-1399.59
WEI	4	-678.29	-1360.59
BCCG	5	-573.72	-1153.43
GG	5	-573.76	-1153.52
GIG	5	-658.37	-1322.75
BCT	6	-552.07	-1112.14
BCPE	6	-555.05	-1118.10
GB2	6	-551.66	-1111.32

Table 15. The table refers to the case $K = 2$.

Emission distribution	m	LL	AIC
EXP	4	-574.35	-1162.70
GA	6	-565.95	-1149.91
LOGNO	6	-545.81	-1109.62
IG	6	-567.37	-1152.74
WEI	6	-567.11	-1152.23
BCCG	8	-571.50	-1165.01
GG	8	-545.14	-1112.27
GIG	8	-550.20	-1122.40
BCT	10	-551.68	-1129.36
BCPE	10	-548.87	-1123.75
GB2	10	-549.07	-1124.14

Table 16. The table refers to the case $K = 3$.

Emission distribution	m	LL	AIC
EXP	5	-564.47	-1156.94
GA	8	-537.42	-1108.84
LOGNO	8	-513.83	-1061.66
IG	8	-519.60	-1073.20
WEI	8	-540.10	-1114.21
BCCG	11	-571.50	-1165.01
GG	11	-499.20	-1038.40
GIG	11	-500.60	-1041.21
BCT	14	-504.97	-1055.95
BCPE	14	-517.90	-1081.80
GB2	14	-499.07	-1044.14

Table 17. The table refers to the case $K = 4$.

Emission distribution	m	LL	AIC
EXP	6	-550.91	-1147.81
GA	10	-505.61	-1065.22
LOGNO	10	-499.86	-1053.71
IG	10	-490.20	-1034.40
WEI	10	-510.49	-1074.98
BCCG	14	-521.10	-1104.19
GG	14	-489.35	-1040.71
GIG	14	-492.61	-1047.21
BCT	18	-489.25	-1048.50
BCPE	18	-495.02	-1060.04
GB2	18	-487.35	-1044.69

References

Abu Bakar, S., N. A. Hamzah, M. Maghsoudi, and S. Nadarajah (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics 61*, 146–154.

Abu Bakar, S., S. Nadarajah, and Z. A. K. Adzhar (2018). Loss modeling using burr mixtures. *Empirical Economics 54*, 1503–1516.

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance 26*(7), 1505–1518.

Adcock, C., M. Eling, and N. Loperfido (2015). Skewed distributions in finance and actuarial science: a review. *The European Journal of Finance 21*(13-14), 1253–1281.

Ahn, S., J. H. Kim, and V. Ramaswami (2012). A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics and Economics 51*(1), 43–52.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Bagnato, L. and A. Punzo (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. *Computational Statistics 28*(4), 1571–1597.

Barbu, V. S. and N. Limnios (2009). *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*, Volume 191. Springer Science & Business Media.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics 41*(1), 164–171.

Bernardi, M. (2013). Risk measures for skew normal mixtures. *Statistics & Probability Letters 83*(8), 1819–1824.

Bernardi, M., A. Maruotti, and L. Petrella (2012). Skew mixture models for loss distributions: a bayesian approach. *Insurance: Mathematics and Economics 51*(3), 617–623.

Bernardi, M., A. Maruotti, and L. Petrella (2017). Multiple risk measures for multivariate dynamic heavy–tailed models. *Journal of Empirical Finance 43*, 1–32.

Bhati, D. and S. Ravi (2018). On generalized log-moyal distribution: a new heavy tailed size distribution. *Insurance: Mathematics and Economics 79*, 247–259.

Bignozzi, V., C. Macci, and L. Petrella (2018). Large deviations for risk measures in finite mixture models. *Insurance: Mathematics and Economics 80*, 84–92.

Brazauskas, V. and A. Kleefeld (2016). Modeling severity and measuring tail risk of norwegian fire claims. *North American Actuarial Journal 20*(1), 1–16.

Bulla, J. (2011). Hidden markov models with t components. increased persistence and other aspects. *Quantitative Finance 11*(3), 459–475.

Bulla, J. and A. Berzel (2008). Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics 23*(1), 1–18.

Bulla, J. and I. Bulla (2006). Stylized facts of financial time series and hidden semi-markov models. *Computational statistics & data analysis 51*(4), 2192–2209.

Bulla, J., I. Bulla, and O. Nenadić (2010). hsmm—an r package for analyzing hidden semi-markov models. *Computational Statistics & Data Analysis 54*(3), 611–619.

Chan, J., S. Choy, U. Makov, and Z. Landsman (2018). Modelling insurance losses using contaminated generalised beta type-ii distribution. *ASTIN Bulletin: The Journal of the IAA 48*(2), 871–904.

Cheng, C. S., Q. Li, G. Li, and H. Auld (2012). Climate change and heavy rainfall-related water damage insurance claims and losses in ontario, canada. *Journal of Water Resource and Protection 4*(2), 49–62.

Choi, S. C. and R. Wette (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics 11*(4), 683–690.

Cooray, K. and C.-I. Cheng (2015). Bayesian estimators of the lognormal–pareto composite distribution. *Scandinavian Actuarial Journal 2015*(6), 500–515.

Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics 51*(2), 239–248.

Eling, M. (2014). Fitting asset returns to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics 59*, 45–56.

Furman, E. (2008). On a multivariate gamma distribution. *Statistics & Probability Letters 78*(15), 2353–2360.

Gómez-Déniz, E., E. Calderín-Ojeda, and J. M. Sarabia (2013). Gamma-generalized inverse gaussian class of distributions with applications. *Communications in Statistics-Theory and Methods 42*(6), 919–933.

Gradeci, K., N. Labonnote, E. Sivertsen, and B. Time (2019). The use of insurance data in the analysis of surface water flood events–a systematic review. *Journal of Hydrology 568*, 194–206.

Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics 12*(3), 604–639.

Hanssen-Bauer, I., H. Drange, E. Førland, L. Roald, K. Børsheim, H. Hisdal, D. Lawrence, A. Nesje, S. Sandven, A. Sorteberg, et al. (2009). Climate in norway 2100. *Background information to NOU Climate adaptation (In Norwegian: Klima i Norge 2100. Bakgrunnsmateriale til NOU Klimatilplassing), Oslo: Norsk klimasenter*.

Hogg, R. V. and S. A. Klugman (1983). On the estimation of long tailed skewed distributions with actuarial applications. *Journal of Econometrics 23*(1), 91–102.

Hong, L. and R. Martin (2017). A flexible bayesian nonparametric model for predicting future insurance claims. *North American Actuarial Journal 21*(2), 228–241.

Hong, L. and R. Martin (2018). Dirichlet process mixture models for insurance loss data. *Scandinavian Actuarial Journal 2018*(6), 545–554.

Jeon, Y. and J. H. Kim (2013). A gamma kernel density estimation for insurance loss data. *Insurance: Mathematics and Economics 53*(3), 569–579.

Johnson, N. L., A. W. Kemp, and S. Kotz (2005). *Univariate Discrete Distributions*. Wiley Series in Probability and Statistics. Wiley.

Jorion, P. (1997). *Value at risk: the new benchmark for controlling market risk*. Irwin Professional Pub.

Klugman, S. A., H. H. Panjer, and G. E. Willmot (2012). *Loss models: from data to decisions*, Volume 715. John Wiley & Sons.

Kundzewicz, Z. W., E. J. Førland, and M. Piniewski (2017). Challenges for developing national climate services–poland and norway. *Climate Services 8*, 17–25.

Lyubchich, V., N. K. Newlands, A. Ghahari, T. Mahdi, and Y. R. Gel (2019). Insurance risk assessment in the face of climate change: Integrating data science and statistics. *Wiley Interdisciplinary Reviews: Computational Statistics 11*(4), e1462.

MacDonald, I. L. (2014). Numerical maximisation of likelihood: A neglected alternative to em? *International Statistical Review 82*(2), 296–308.

Maruotti, A. and A. Punzo (2021). Initialization of hidden markov and semi-markov models: A critical evaluation of several strategies. *International Statistical Review 89*(3), 447–480.

Mazza, A. and A. Punzo (2014). **DBKGrad**: An R package for mortality rates graduation by discrete beta kernel techniques. *Journal of Statistical Software 57*(Code Snippet 2), 1–18.

Mazza, A. and A. Punzo (2015). Bivariate discrete beta kernel graduation of mortality data. *Lifetime data analysis 21*(3), 419–433.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Miljkovic, T. and D. Fernández (2018). On two mixture-based clustering approaches used in modeling an insurance portfolio. *Risks 6*(2), 57.

Miljkovic, T. and B. Grün (2016). Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics 70*, 387–396.

Nadarajah, S., B. Zhang, and S. Chan (2014). Estimation methods for expected shortfall. *Quantitative Finance 14*(2), 271–291.

O'Connell, J. and S. Højsgaard (2011a). Hidden semi markov models for multiple observation sequences: the mhsmm package for r. *Journal of Statistical Software 39*, 1–22.

O'Connell, J. and S. Højsgaard (2011b). Hidden semi Markov models for multiple observation sequences: The **mhsmm** package for R. *Journal of Statistical Software 39*(4), 1–22.

Pigeon, M. and M. Denuit (2011). Composite lognormal–pareto model with random threshold. *Scandinavian Actuarial Journal 2011*(3), 177–192.

Punzo, A., L. Bagnato, and A. Maruotti (2018). Compound unimodal distributions for insurance losses. *Insurance: Mathematics and Economics 81*, 95–107.

Punzo, A., A. Mazza, and A. Maruotti (2018). Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions. *Journal of Applied Statistics 45*(14), 2563–2584.

Rigby, B., M. Stasinopoulos, G. Heller, and V. Voudouris (2014). The distribution toolbox of gamlss. *The GAMLSS Team*.

Rydén, T. (2008, 12). Em versus markov chain monte carlo for estimation of hidden markov models: a computational perspective. *Bayesian Anal. 3*(4), 659–688.

Spekkers, M., F. Clemens, and J. Ten Veldhuis (2015). On the occurrence of rainstorm damage based on home insurance and weather data. *Natural Hazards and Earth System Sciences 15*(2), 261–272.

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

Spekkers, M., M. Kok, F. Clemens, and J. Ten Veldhuis (2014). Decision-tree analysis of factors influencing rainfall-related building structure and content damage. *Natural Hazards and Earth System Sciences 14*(9), 2531–2547.

Stasinopoulos, M. and B. Rigby (2017). ***gamlss.dist**: Distributions for Generalized Additive Models for Location Scale and Shape*. R package Version 5.0-4 (2017-12-11).

Stasinopoulos, M., B. Rigby, V. Voudouris, C. Akantziliotou, M. Enea, and D. Kiose (2023). Package 'gamlss.dist'. *Available online: http://www. gamlss. org (accessed on 16 July 2021)*.

Stasinopoulos, M. D., R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani (2017). *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press.

Tomarchio, S. D. and A. Punzo (2019). Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society Series A: Statistics in Society 182*(4), 1247–1266.

Tomarchio, S. D. and A. Punzo (2020). Dichotomous unimodal compound models: application to the distribution of insurance losses. *Journal of Applied Statistics 47*(13–15), 2328–2353.

Torgersen, G., J. T. Bjerkholt, K. Kvaal, and O. G. Lindholm (2015). Correlation between extreme rainfall and insurance claims due to urban flooding–case study fredrikstad, norway. *Journal of Urban and Environmental Engineering 9*(2), 127–138.

Verbelen, R., L. Gong, K. Antonio, A. Badescu, and S. Lin (2015). Fitting mixtures of erlangs to censored and truncated data using the em algorithm. *ASTIN Bulletin: The Journal of the IAA 45*(3), 729–758.

Vernic, R. (2006). Multivariate skew-normal distributions with applications in insurance. *Insurance: Mathematics and economics 38*(2), 413–426.

Yu, S.-Z. (2015). *Hidden Semi-Markov models: theory, algorithms and applications*. Morgan Kaufmann.

Zucchini, W., I. L. MacDonald, and R. Langrock (2017). *Hidden Markov models for time series: an introduction using R*. CRC press.

NHH

NORGES HANDELSHØYSKOLE
Norwegian School of Economics

Helleveien 30
NO-5045 Bergen
Norway

T +47 55 95 90 00
E nhh.postmottak@nhh.no
W www.nhh.no