https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

Short-Term Management of Water-Damage Claim Risk Using Ensemble Precipitation Forecasts

Håkon Otneim [1], Etienne Dunn-Sigouin[2], Sondre Hølleland[1], Mahsa Gorji[1], and Geir Drage Berentsen[1]

[1]Department of Business and Management Science, NHH Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway
[2]Bjerknes Centre for Climate Research, NORCE Norwegian Research Centre AS, Jahnebakken 5, 5007 Bergen, Norway

Correspondence: Håkon Otneim (hakon.otneim@nhh.no)

Abstract. Insurers are increasingly challenged by weather-related claims arising from property damage, yet they lack adequate tools for near-term planning because traditional actuarial models do not incorporate real-time weather forecasts. This study demonstrates that incorporating ensemble precipitation forecasts improves short-range (1-4 days ahead) predictions of property insurance claim counts, thereby enabling proactive risk management. We present a forecasting framework for two Norwegian cities, Bergen and Oslo, using precipitation forecasts to predict days exceeding operationally significant thresholds. The models are evaluated by their forecast skill and reliability in predicting claim surges, as well as by their economic value in a cost-loss decision context, illustrating the potential reduction in expected costs when early warning triggers are in place. Results show that weather-informed models substantially outperform baseline models based on climatology, improving discrimination of claim events and yielding up to 30-50% reduction in expected daily costs under various ideal warning scenarios. Two case studies of extreme events highlight how weather forecasts translated into early claims warnings could guide resource allocation and customer advisories. Overall, the presented framework highlights the practical benefit of integrating meteorological forecast information into insurance operations and offers a template for insurers to enhance climate resilience through improved risk communication and short-term decision support.

1 Introduction

Climate change is reshaping the risk landscape for property insurers. Rising greenhouse gas concentrations are altering the frequency, intensity, and spatial distribution of extreme precipitation and other weather extremes. The Intergovernmental Panel on Climate Change (IPCC, 2023) concludes that changes in extreme rainfall and flood risk are already detectable in many regions and are projected to intensify under continued warming. At the same time, increasing asset exposure and vulnerability due to economic development amplify losses (Mills, 2005). Observed trends in weather-related losses therefore reflect both changing hazards and socioeconomic development. For insurers, this combination challenges the traditional assumption that future risk can be inferred from historical experience alone. It raises concerns about the long-term insurability of some perils and regions, see, for instance, Keys and Mulder (2024) and Flavelle and Rojanasakul (2024).

While much attention has focused on "headline" catastrophe events, insurers also face mounting pressure from frequent, low-severity events that cumulatively generate substantial losses and operational strain. Mills (2005) and subsequent studies

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

document the growing cost of recurrent weather-related damages, including water intrusion and pluvial flooding in urban environments. Industry reports and academic work alike emphasize that such high-frequency events can erode profitability and service quality even when they fall short of what is traditionally considered a catastrophe (The Geneva Association, 2021; Lyubchich et al., 2019b). Regulatory initiatives, such as the EU taxonomy and climate-risk disclosure frameworks, further push insurers to demonstrate robust risk management across both catastrophic and non-catastrophic weather perils (European Commission, 2020).

The Norwegian housing insurance market provides a useful case study of these challenges. Finance Norway (2024) reports that a significant fraction of weather-related housing damage claims in recent years stem from frequent, low-severity rainfall events that cause water intrusion, overloaded drainage, and sewer backflow. A state-mandated Natural Perils Pool distributes the substantial financial burden of major natural hazards among insurers. However, individual companies still bear the operational and administrative costs of handling both ordinary and pool-covered water damage claims (Norwegian Natural Perils Pool, 2024). Recent studies document increasing precipitation in Norway and rising trends in both ordinary and water-related home insurance claims (Haug et al., 2011; Scheel and Hinnerichsen, 2012; Lyubchich et al., 2019b), further highlighting the need for forward-looking tools that can support day-to-day operations, in addition to the more traditional actuarial focus on long-term issues such as pricing, capital allocation, and portfolio management.

A substantial actuarial and statistical literature has analyzed the impact of weather and climate on property insurance claims. Early work on Norwegian data employed generalized linear models to relate water-damage claim counts and severities to meteorological conditions and to project future losses under climate change scenarios; see Haug et al. (2011) for an early effort, and Heinrich-Mertsching et al. (2023) for a more recent high-resolution analysis of the long-term impact of climate risk on home insurance. Bayesian hierarchical models with spatial structure have been used to identify weather-sensitive regions and quantify the effect of meteorological covariates on claim frequency (Scheel et al., 2013); see also Wahl et al. (2022) for a similar perspective on risk premiums. Subsequent studies applied extreme-value and peaks-over-threshold methods to model water-related insurance losses and their dependence on extreme rainfall (Rohrbeck et al., 2018; Shi, 2025; Hettiarachchi et al., 2018), while yet other studies extend this line of work using hidden semi-Markov models for temporal clustering of rainfall-related claims and retrospective reconstructions of claim histories from high-resolution weather data (Shi et al., 2025b, a). Parallel efforts in North America and elsewhere have used regression and machine learning methods to forecast weather-induced home insurance claims, with a general focus on attribution and long-horizon climate impacts rather than short-term operational decisions (Lyubchich et al., 2019a; Dey et al., 2021).

Insurers and researchers have also begun to develop impact models that use ensemble weather predictions to forecast near-term damages and losses (Merz et al., 2020). For example, ensemble wind forecasts have been combined with empirical vulnerability models to predict winter storm losses and to quantify the uncertainty and skill of such forecasts at the event scale (Pardowitz et al., 2016; Jaison et al., 2025). Operational impact-forecasting systems now exist for building damage from winter storms, using ensemble wind forecasts to produce probabilistic damage maps several days in advance (Röösli et al., 2021). Studies comparing insurer loss data with modelled building damage suggest that combining claim records with physical event footprints can substantially improve risk assessment; see Welker et al. (2021) for a case study from Switzerland, and

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

Moemken et al. (2024) for a more general European perspective on wind storm damages. Together with emerging work on forecast-based insurance design and parametric products (see, for instance, Abrego-Perez and Nuñez-Mora (2026)), these developments suggest that short- to medium-range weather forecasts could be used to improve insurance risk management.

From a risk management perspective, most existing models remain focused on (i) retrospective tools for understanding historical weather-claim relationships, (ii) forward-looking approaches that price long-term risk under varying climate change scenarios, or (iii) event-focused systems anticipating rare, high-impact hazards like storms and floods. However, the routine operational problem faced by property insurers is different. On any given day, they must decide whether to anticipate a surge in weather-related claims, adjust staffing and logistics, and, where possible, issue targeted warnings or mitigation advice to policyholders. This gap is especially relevant for Norwegian property insurance, since more frequent but lower-impact rain-related damages constitute a significant share of their costs. This is a short-term risk management problem in which the key perspective is not on expected annual losses but rather on the probability that the number of claims tomorrow or over the next few days will exceed a level that stresses operations.

Modern ensemble weather prediction systems are well-suited to supporting such decisions. By perturbing initial conditions, ensemble forecasts provide a probabilistic description of future weather and have become standard in operational meteorology for risk-based decision support. However, care must be taken when integrating such forecasts in downstream applications, such as short-term insurance risk management, since they may contain biases that require correction. In parallel, the verification and economic-evaluation literature provides a rich toolbox for assessing forecast discrimination, reliability, and decision value, including proper scoring rules and cost-loss-based measures of potential economic value (Jolliffe and Stephenson, 2012; Murphy, 1993; Katz and Murphy, 1997).

This paper brings these ideas together by examining whether ensemble precipitation forecasts can anticipate short-term surges in water-related home insurance claims. We focus on two Norwegian metropolitan areas and frame the problem as a rare-event binary classification task, predicting the daily probability that precipitation-driven claim counts exceed a practically relevant threshold. Rather than aiming to model the entire distribution of claim counts or to price long-term risk, our goal is explicitly operational: To evaluate whether weather-informed probabilistic forecasts improve short-term risk management decisions by more skilfully predicting insurance claim surges than models that do not incorporate weather information.

Our contributions are threefold. First, we develop probabilistic models to predict insurance claim surges that combine observed and forecast precipitation, employing both generalized additive models and machine-learning approaches across different training and testing strategies. Second, we evaluate these forecasts using metrics appropriate for low-base-rate events, including measures of discrimination, calibration, and precision-recall, and compare these predictions against baselines derived from claims-only models without weather-forecast information. Third, we quantify the potential economic value of these probabilistic claim forecasts using a cost-loss framework adapted from the weather forecast literature, interpreting claim predictions as inputs to a stylized operational decision problem in which an insurer chooses whether to take preventive or preparatory action.

By situating short-term claim prediction within a risk-management and decision-theoretic framework and explicitly linking ensemble precipitation forecasts to the operational challenges posed by frequent water-damage claims, we aim to complement

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

existing actuarial work on climate-related insurance risk. The results illustrate how ensemble precipitation forecasts, suitably processed and evaluated, can provide actionable probabilistic information for day-to-day insurance operations, and they highlight both the potential and the limitations of weather-forecast-based approaches to managing weather-related insurance risk.

2 Data

2.1 Precipitation-related home insurance claims

The claims data are provided by one of Norway's largest private housing insurance companies. We analyse daily counts of home insurance claims related to precipitation-driven water damage in the Oslo and Bergen metropolitan areas for the period 2013–2021. The claims include damages caused by water intrusion through roofs, walls, and foundations, as well as overloaded drainage and sewer backflow. Figure 1a shows the time series of claim counts on a logarithmic y-scale.

Over the 2,921 days in our sample, a total of 2,829 claims were recorded: 1,543 in Oslo and 1,286 in Bergen. The empirical distributions are highly zero-inflated: in Oslo, 76% of days have zero claims and 94% have at most one claim. In Bergen, the corresponding figures are 81% and 96%. Nevertheless, both cities exhibit heavy right tails, with single-day maxima of 220 claims in Oslo and 291 in Bergen.

Norwegian insurers participate in the Natural Perils Pool, a collective reinsurance arrangement for natural disasters. This policy covers damages from a predefined list of hazards, including storms, floods, and landslides. All insurers pay a premium for objects that are also insured against fire (such as private homes), and the pool reimburses covered claims proportionally to each insurer's market share. From the perspective of our study, which focuses on short-term risk management and operational planning, it is immaterial whether a given claim is ultimately borne by the individual insurer or by the pool. The insurer still handles all customer communication and administrative processing, and short-term surges in pool-covered claims have the same operational impact as ordinary water-damage claims. We therefore include both ordinary and pool-covered claims, distinguishing between them only descriptively. In total, 934 claims (33%) in the dataset are eventually covered by the Natural Perils Pool. The proportion differs between cities: in Oslo, only 14% of claims are pool-covered, whereas in Bergen, 56% are. This reflects differences in local hazard profiles and exposure.

An important practical feature of these data is the potential delay in reporting and registration. Policyholders do not always report damages immediately after occurrence, and claims must be registered and quality-checked in the insurer's systems before they appear as final counts in the database. In our analysis, we align claims with the date of occurrence as recorded by the insurer, which is the date on which the damage is deemed to have occurred and on which administrative processing typically begins. For operational forecasting, however, the fully reconciled number of claims on day t is not known when a forecast for day $t+1$ is issued, even though many of the associated customer contacts and case files are already being processed. This delay in the availability of reliable daily totals has direct implications for the use of lagged claims as predictors in real-time models, an issue we return to in Section 4.4.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

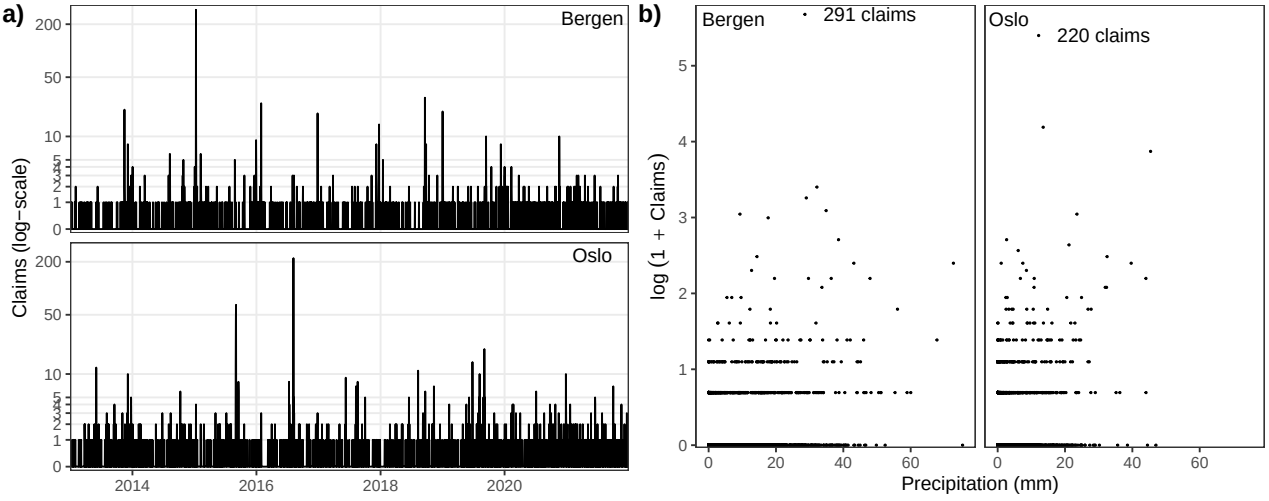

Figure 1. a) Precipitation-related insurance claims for Bergen and Oslo. Note that the y-scale is logarithmic by a $\log(1 + \text{claims})$ transformation. b) Relationship between the number of daily claims and the observed precipitation in the two cities as scatter plots.

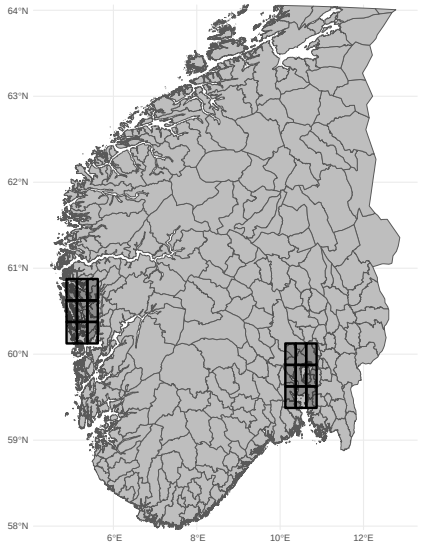

Figure 2. Map of southern Norway, with the nine grid cells covering the Bergen metropolitan area displayed on the western coast, while the nine grid cells covering the Oslo metropolitan area displayed on the eastern edge of the map.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

2.2 Precipitation and relation to insurance claims

We define observed precipitation as the daily accumulated total precipitation (rain and snow) from 2013-2021 at a horizontal resolution of $0.25° \times 0.25°$ from the ERA5 reanalysis (Hersbach et al., 2023). Although ERA5 is not a direct observation but a model-based reconstruction produced through the assimilation of diverse observational datasets into a numerical weather prediction system, it provides spatially complete and dynamically consistent fields at the same grid resolution as the forecasts analysed below. For our purposes, ERA5 offers an operationally realistic and internally consistent representation of meteorological conditions throughout the study period. For each metropolitan area and each day, we compute a spatial average over a $0.75° \times 0.75°$ region centred on the urban domain, corresponding to a 3×3 block of grid cells shown in Figure 2.

Figure 1b illustrates the relationship between daily claim counts and corresponding daily precipitation in the two metropolitan areas. Although high precipitation does not invariably result in large numbers of claims, elevated claim counts are often associated with higher amounts of precipitation. This asymmetric dependence suggests that rainfall is a necessary but not sufficient condition for water intrusion damages, with differences in exposure and vulnerability modulating the impacts. These characteristics motivate the simple binary formulation of the prediction problem introduced in section 3, where the objective is to discriminate between operationally quiet and operationally stressful days rather than to predict exact claim counts.

2.3 Precipitation forecasts

We use ensemble forecasts of daily accumulated total precipitation from 2013-2021 from the European Centre for Medium-Range Weather Forecasts (Buizza et al., 2018) obtained from the MARS archive (European Centre for Medium-Range Weather Forecasts (ECMWF), 2024). Each forecast consists of 51 ensemble members with a horizontal resolution of $0.25° \times 0.25°$. All members and lead-times are spatially averaged over the same $0.75° \times 0.75°$ grid box as the observations. Figure 3a shows an example precipitation forecast for Bergen initialised on the 15^{th} of February, 2015, where the grey lines depict the forecast ensemble members, and the green line shows the observed precipitation.

Forecasts are initialised twice per week, on Mondays and Thursdays at 00 UTC. Due to bandwidth constraints on the MARS archive, we downloaded these bi-weekly forecasts for the full 2013–2021 period, rather than a denser set of initialisation times over a shorter period. For each calendar day t in our sample, we select the forecast with the shortest lead time available at the start of that day. Specifically, this means that most days are associated with precipitation forecasts with lead times of 1–4 days. For two missing forecast cycles, we rely on lead times up to 7 days.

Before using the precipitation forecasts to predict insurance claims, we first evaluate their skill at predicting observed precipitation using two standard scoring rules from meteorology: the deterministic Mean Squared Error Skill Score (MSESS) and the probabilistic Continuous Ranked Probability Skill Score (CRPSS) (Jolliffe and Stephenson, 2012). The MSESS is based on the mean squared error of the ensemble-mean forecast $\hat{y}_n$ relative to the verifying observation y_n for all forecast-observation pairs N,

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2, \tag{1}$$

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

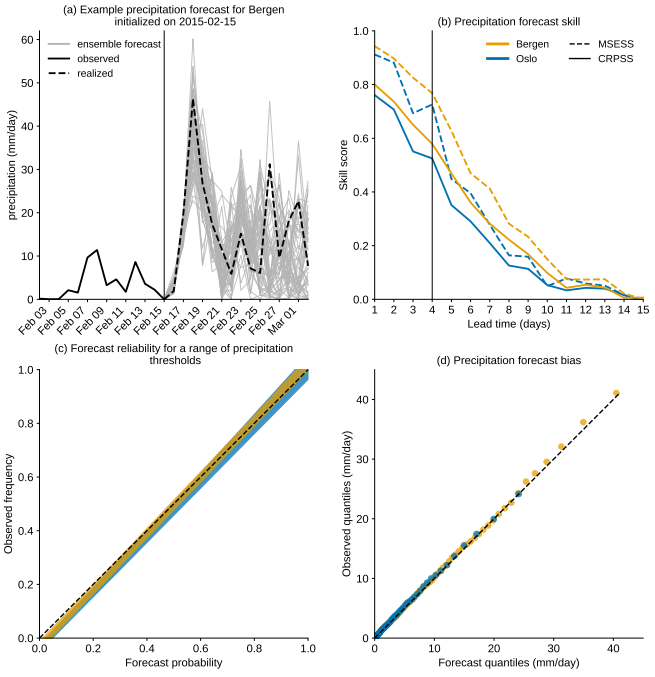

Figure 3. Summary of precipitation forecast performance in Bergen and Oslo. a) Example ensemble precipitation forecast for Bergen initialised on 15 February 2015, showing the 51 ensemble members and the corresponding observed precipitation. b) Forecast skill as a function of lead time, quantified using the deterministic Mean Square Error Skill Score (MSESS) and the probabilistic Continuous Ranked Probability Skill Score (CRPSS). c) Forecast reliability (calibration) across representative precipitation thresholds (0, 5, 10, 15, 25 and 30 mm/day). d) Quantile–quantile (Q–Q) relationship between forecast and observed precipitation for lead times 1–4 days.

which is converted into a skill score by benchmarking against an observational climatological reference forecast,

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{clim}}}. \tag{2}$$

To assess the full predictive distribution, we use the CRPSS, which measures the integrated squared distance between the forecast cumulative distribution function $F_n(x)$ to the step function associated with the verifying observation for all precipitation thresholds x,

$$\mathrm{CRPS} = \frac{1}{N} \sum_{n=1}^{N} \int_{-\infty}^{\infty} \left(F_n(x) - 1\{x \geq y_n\}\right)^2 \mathrm{d}x. \tag{3}$$

and define the corresponding skill score relative to a reference forecast based on the observational climatological distribution,

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}}{\mathrm{CRPS}_{\mathrm{clim}}}. \tag{4}$$

By construction, a value of one denotes a perfect forecast, zero indicates no improvement over climatology, and negative values imply performance worse than the climatological baseline.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

As shown in Figure 3b, forecast skill is highest during the first few days after initialization and declines thereafter, reflecting the rapid loss of predictability in mid-latitude precipitation (Dunn-Sigouin et al., 2025). The CRPSS is consistently lower than the MSESS because it evaluates the full predictive distribution and therefore constitutes a more stringent benchmark for forecast performance. Skill is generally higher in Bergen than in Oslo, likely reflecting differences in prevailing precipitation regimes: more predictable large-scale frontal systems in the former versus more localized and less predictable convective events in the latter. Because our claim-prediction framework relies on operationally actionable guidance, we focus on lead times of 1–4 days (denoted by the vertical black line), during which both deterministic and probabilistic precipitation skill remain substantial.

Figure 3c,d further evaluate systematic biases in forecast precipitation amounts and probabilities. Figure 3c summarises forecast reliability across a range of precipitation thresholds (0, 5, 10, 15, 25, and 30 mm/day), spanning the observed precipitation range in both cities for lead times 1–4 days. Reliability (or calibration) describes how closely forecast probabilities match observed frequencies on average: for example, events predicted with a $p\%$ probability should occur on roughly $p\%$ of occasions (See also Section 4.2, where we report similar reliability diagrams for the forecasted claim events). In both Bergen and Oslo, the shaded envelopes lie close to the diagonal $y = x$ line, indicating that forecast probabilities are generally well calibrated and show little systematic over- or under-confidence.

Figure 3d presents the quantile-quantile (Q-Q) relationship between forecast and observed precipitation for lead days 1-4. The points closely follow the diagonal $y = x$ line in both cities, indicating only small overall systematic biases. A slight underestimation is evident in the upper tail in Bergen, where the highest observed quantiles exceed the corresponding forecast quantiles by approximately 5–10%.

Overall, the precipitation forecasts used in this study exhibit good calibration and only minor biases at the considered lead times. We therefore do not apply bias correction prior to using the precipitation forecasts for claim prediction. This decision is further supported by the fact that models trained and tested with forecast precipitation yield similar claim prediction skill to those using only observed precipitation (see Section 4), suggesting weather forecast bias plays a minor role.

3 Modelling strategies

3.1 Definitions and choice of threshold

Let C_{it} denote the number of precipitation-related claims on day $t \in \{1, \ldots, T\}$ in city i (Oslo or Bergen). As discussed above, the empirical distribution of C_{it} is highly zero-inflated and heavy-tailed. From the insurer's operational perspective, the precise count on a given day is less important than whether the claims volume exceeds a level that requires extraordinary measures, such as reallocating staff or issuing targeted warnings.

We therefore define a threshold u and consider the binary outcome

$$Y_{it} = \begin{cases} 0, & C_{it} \leq u, \\ 1, & C_{it} > u. \end{cases}$$

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

In our analysis, we use $u = 2$. In Table 1 we tabulate the empirical cumulative distribution of C_{it} for each city. With $u = 2$, the event $Y_{it} = 1$ occurs on 1.6% of days in Bergen and 2.4% of days in Oslo. This choice reflects discussions with our insurance partner: on most days, the company can handle up to two water-damage claims per city without special preparation, whereas days with three or more claims are perceived as operationally challenging, particularly when they occur in clusters.

Table 1. Empirical cumulative probability mass function of claims (as percentages).

	≤ 0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Bergen	80.8	96.0	98.4	99.1	99.3	99.4
Oslo	76.0	93.9	97.6	98.6	99.1	99.3

We therefore frame the problem as a rare-event classification task rather than as a count regression. A count model for C_{it} could in principle be used to derive exceedance probabilities for any threshold, including $u = 2$, see Rohrbeck et al. (2018) for a suitable modelling strategy in this direction. However, in our application, the decision to take preventive or preparatory action is triggered by crossing a specific, operationally meaningful threshold. Directly modelling Y_{it} allows us to focus model capacity on the parts of the distribution most relevant to risk management and to evaluate performance using metrics tailored to highly imbalanced binary outcomes.

We have, for the sake of completeness, checked that choosing $u = 1$ would not materially change the conclusions of this paper. Choosing $u = 3$ would lead to yet more imbalance, particularly in the case of Bergen, requiring a careful adaptation of the methodological approach, which is outside the scope of this paper, as this threshold is deemed of lesser practical interest. See Section 4.4 for more details on our robustness checks.

3.2 Strategies for translating weather forecasts into insurance risk

The central methodological question is how to translate ensemble precipitation into predictions of Y_{it}. We consider two complementary strategies that rely solely on information available to the insurer at the time of forecasting.

Strategy 1 (S1): Train on observed precipitation, predict using ensemble weather forecast. In the first strategy, we fit models that describe the conditional probability

$$p(y|x) = P(Y_{it} = 1|X_{it} = x),$$

where X_{it} denotes a scalar measure of *observed* daily precipitation for city i and day t. In our case, X_{it} is the spatially averaged daily accumulated precipitation in the relevant grid box. The models are trained separately for each city using data from the training period, so that the fitted relationships represent how observed precipitation translates into operational claim risk in each city.

At forecast time, we no longer know the realised precipitation X_{it} on future days. Instead, we have a 51-member ensemble $\{X_{it}^{(j)} : j = 1, \ldots, 51\}$ which, in the standard meteorological interpretation, is treated as a set of exchangeable draws from a predictive distribution $F_{X_{it}}$ for daily precipitation conditional on today's information. Given an estimated conditional response

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

function fitted to observed precipitation, we define the Strategy 1 forecast probability of a claims surge as

$$\widehat{p}_{it}^{\mathrm{S1}} = \widehat{P}(Y_{it} = 1) = \frac{1}{51} \sum_{j=1}^{51} \widehat{p}\left(Y_{it} = 1 \mid X_{it} = X_{it}^{(j)}\right).$$

Under the assumption that the ensemble members are samples from a predictive distribution for X_{it}, this quantity is the standard Monte Carlo estimator of

$$\int \widehat{p}(Y_{it} = 1 \mid x)\, \mathrm{d}F_{X_{it}}(x),$$

that is, of the law-of-total-probability expression for the forecast probability induced by the conditional model $\widehat{p}(\cdot \mid x)$ and the predictive distribution of X_{it}. In practice, both the conditional response $\widehat{p}$ and the predictive distribution for X_{it} are estimated, and the ensemble is finite, so the resulting forecast $\widehat{p}_{it}^{\mathrm{S1}}$ should be understood as a model-based approximation to the true predictive probability. Conceptually, we propagate uncertainty in future precipitation using a model trained on observed precipitation.

Strategy 2 (S2): Train directly on past weather forecasts, predict using future weather forecasts. In the second strategy, we train models directly on ensemble-based predictors rather than on observed precipitation. For each day t and city i, we construct a set of predictors from the ensemble, including scalar summaries such as the ensemble mean, median, minimum, and maximum precipitation, as well as the full vector of 51 ensemble members ordered by their value (the *ranked forecasts*).

The aim is to let the statistical model learn how the distribution of forecast precipitation, rather than only its central tendency, relates to the probability of a claims surge. For instance, a forecast where a few ensemble members predict very heavy rain while the median remains moderate may imply a different risk level than a forecast where all members predict similar, moderate amounts.

Both strategies assume the relationship between precipitation and insurance claims remains stable over time, whether derived from observed (S1) or forecasted (S2) precipitation. This assumption may fail due to changes in variability of observed precipitation, periodic upgrades to weather forecasting systems that modify their predictive skill, or changes in societal vulnerability and exposure. Despite this, each approach offers distinct operational advantages. Training on observations requires substantially less data and can be implemented more rapidly, particularly under bandwidth or availability constraints imposed by major weather forecast providers. In contrast, training directly on weather forecast output internalizes systematic model biases within the predictive framework, thereby eliminating the (possible) need for additional downstream bias correction when generating future claim predictions.

3.3 Forecast models

Table 2 summarizes the set of forecasting models explored in this study, organized according to the two strategies outlined above. We include a variety of model classes, ranging from a standard logistic regression and generalized additive models (Hastie and Tibshirani, 1990), to more flexible machine learning approaches, including the Lasso (Tibshirani, 1996), xgBoost (Chen and Guestrin, 2016), and Convolutional Neural Networks (Goodfellow et al., 2016). This broad range enables us to benchmark traditional statistical methods against more complex models that capture nonlinearities and high-dimensional pat-

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

Table 2. Modeling strategy, where S1 and S2 refer to the model belonging to Strategy 1 or Strategy 2 as described above.

	Name	Description	S1	S2
Reference models	observed	Logistic regression trained and tested on observed precipitation		
	seasonal	Logistic regression as a generalized additive model with cyclic p-splines for day of year, without forecast information		
	unconditional	Constant probability of observing $Y = 1$, equal to the observed proportion in the training data		
Generalized linear/ additive models	observed-forecast	Logistic regression trained on observed rain, forecasted using median of ensemble	x	
	observed-forecast-gam	Generalized additive model trained on observed rain, forecasted using median of ensemble	x	
	saturated	Logistic regression with all ranked forecasts as predictors		x
	stepwise	Logistic regression with stepwise backward model selection		x
	lasso	Lasso-penalized logistic regression with all ranked forecasts		x
Machine learning methods	xgboost	XGBoost using minimum, median, and maximum forecast for each day as predictors		x
	cnn	Convolutional neural network with all ranked forecasts		x

terns in ensemble forecast data. We note that our suite of models does not constitute an exhaustive list of possibilities, and that there remains potential for improvement through a more comprehensive exploration of the space of statistical and machine learning methods, as well as by allocating more computational resources to training and tuning the models.

Two models serve as baselines at each end of the spectrum of predictive performance; The `observed` model, trained and tested on observed precipitation, which represents an upper bound in terms of having a perfect weather forecast available for training the models, as well as the `unconditional` model that predicts a constant probability equal to the observed claim frequency in the training data, without any weather forecast input. We also include the `seasonal` model as a third reference, which describes purely seasonal patterns in insurance claims, independent of meteorological input.

We then present two statistical models trained on observed precipitation and tested using the median ensemble forecast, as well as a range of ensemble-based models that differ in how they summarize or incorporate the 51-member forecast data. Models such as `lasso`, `stepwise`, and `saturated` vary primarily in the level of predictor selection or regularization applied to the whole ensemble, while the `cnn` model is tasked with learning the spatial or ordered structure across the ensemble

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

members. We select a limited set of predictors for xgBoost (the minimum, median, and maximum ensemble members), as this yields slightly better results than using the entire ensemble. The neural network architecture is documented in A.

To assess the robustness of our conclusions, we fit the suite of models presented in Table 2 under the following three variations, none of which alter the conclusions in the paper: (i) by changing the threshold u, (ii) by fitting a joint model for both cities, using the city as an additional predictor, and (iii) by using the lagged value of of the claim indicator $Y_{i,t-1}$ as an additional predictor. See Section 4.4 for a further discussion of these checks.

4 Results

4.1 Predictive power

Our primary objective in predicting future claim events is to accurately distinguish occurrences from non-occurrences, known as *discrimination* (Murphy, 1993). Common verification metrics for binary classification include the area under the Receiver Operating Characteristic (ROC) curve (AUC), which assesses a model's ability to rank events across decision thresholds, and the Brier score, which summarizes the overall accuracy of probabilistic predictions for binary events by penalizing deviations between predicted probabilities and observed outcomes. However, in this study, we report the Area Under the Precision-Recall Curve (PRAUC), which we find more appropriate for assessing discrimination for two key reasons.

First, short-term insurance claim prediction involves a substantial class imbalance, with the number of non-claim days far exceeding that of claim days. In such settings, the AUC can give an overly optimistic impression of model performance, as even a high number of false positives may appear negligible when normalized by the overwhelming number of non-events (Jolliffe and Stephenson, 2012). As a result, all models listed in Table 2 achieve high ROC AUC values simply by correctly identifying the majority class "no-claim" every day. In contrast, the Precision-Recall (PR) curve focuses on the minority class (claim surges), making it more sensitive to correctly identifying rare but operationally important events. A similar limitation applies to the Brier score, which is likewise dominated by the large number of correctly predicted non-events under strong class imbalance.

Second, the practical costs associated with prediction errors are not necessarily the same between false negatives and false positives, as discussed in the preceding section. False positives (FP, predicting claims where none occur) may lead to unnecessary warnings or resource allocation, potentially causing customer fatigue or trust erosion (the so-called "cry wolf" effect; see LeClerc and Joslyn (2015)). Conversely, false negatives (FN, failing to predict actual claims) can result in substantial financial losses and missed opportunities to mitigate damage. In this context, evaluation metrics that emphasise performance on the event class, such as PRAUC, more directly reflect the decision-relevant trade-offs faced by insurers.

The PR curve addresses this by explicitly summarizing the trade-off between precision and recall, defined as

$$\text{Precision} = \frac{TP}{TP+FP}, \qquad \text{Recall} = \frac{TP}{TP+FN},$$

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

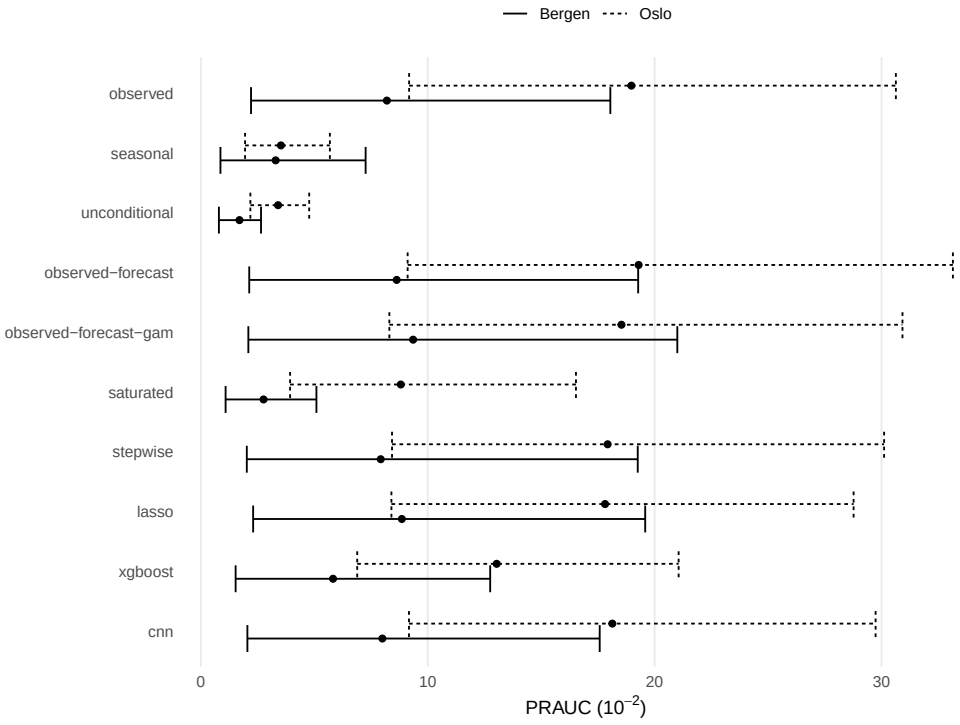

Figure 4. PRAUC for all models and the two metropolitan areas. The 95% confidence intervals are bootstrapped with 1000 bootstrap samples from the predicted probabilities on the test set.

where, in addition to the false positives and false negatives, we need the number of true positives (TP) for a given threshold. The PRAUC summarizes this curve across all possible classification thresholds.

The out-of-sample predictive power for all models in Table 2 is presented in Figure 4. Lower PRAUC values indicate worse discrimination, while higher values indicate better discrimination. The results are presented in the same order as listed in the table. All results are reported here, along with their 95% confidence intervals derived from bootstrapping the predicted probabilities on the test set. The solid lines represent results from Bergen, and the dashed lines represent results from the Oslo area. There are two main results in this figure that we comment on below:

Models that incorporate weather information better predict claim surges than those that do not. The PRAUC metric reveals a clear separation between the two baseline scenarios: the `observed` baseline represents an upper bound on performance, while the `unconditional` baseline, which ignores meteorological input features entirely, performs substantially worse. This sharp contrast underscores the importance of incorporating weather information into prediction models. Among models that include weather forecast information, the `saturated` model consistently underperforms, likely due to over-fitting or poor generalization. Models grouped under strategies 1 and 2, as described in Section 3.2, generally achieve performance on

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

par with the upper bound `observed` model, suggesting that even relatively simple or regularized approaches can effectively extract predictive signal from the ensemble forecast data.

On the other hand, more flexible models, such as xgBoost and Convolutional Neural Networks, do not outperform simpler methods. This may reflect challenges in model tuning, limited sample size, or a lack of sufficiently complex signals in the data that would benefit from higher model capacity. This suggests that the specific model is less important than the quality of the input weather forecast, which is skilful for predictions over the next few days (See Figure 3). Overall, model performance in this case appears to depend more on effective regularization and appropriate model complexity than on raw flexibility.

Claim surges in Oslo are better predicted than those in Bergen. The out-of-sample PRAUC reveals systematic differences in model performance between the two cities, with consistently higher scores observed in Oslo compared to Bergen. As shown in Figure 4, the predictions for Oslo consistently yield higher PRAUC values across models compared with the predictions for Bergen. One plausible explanation is that precipitation patterns in Oslo exhibit greater temporal variability, alternating between dry spells and intense rainfall, whereas those in Bergen tend to be more frequent but less variable. This contrast may lead to clearer predictive signals in Oslo, which the models can better capture.

4.2 Reliability

A secondary, yet important, objective in predicting future claim events is ensuring that predictions are *reliable* (or *calibrated*), meaning that predicted probabilities closely match observed frequencies. In other words, for all instances where a model predicts an event to occur with probability $p\%$, we expect that event to be realized in approximately $p\%$ of these cases. This matters in decision-making contexts, such as insurance forecasting, where predicted risk levels must be trusted to guide action.

It is crucial to recognize that the two concepts of discrimination and reliability are distinct. A model may exhibit high discrimination, effectively differentiating between events and non-events, yet produce predictions that systematically over- or underestimate the actual occurrence probabilities. Conversely, a model can provide well-calibrated (reliable) probabilities but still fail to strongly distinguish occurrences from non-occurrences. Our `unconditional` forecast, which predicts the observed frequency of a claim event every day regardless of the weather forecast, is therefore a reliable forecast without much discriminatory power. Thus, a high-quality forecast for risk assessment and damage prevention must have both high discriminative ability and reliable probabilities.

We assess the reliability of the prediction models listed in Table 2 by grouping the predicted daily probabilities of observing the high-claim outcome into bins and computing the cumulative observed frequencies of this outcome within each bin (DeGroot and Fienberg, 1982). A perfectly reliable forecast would yield points aligned along the diagonal in the resulting plot.

However, the pronounced class imbalance in the claims data requires special consideration. If we were to create ten equally spaced probability bins ranging from 0% to 100%, nearly all predicted values would cluster in the lowest bins, leaving the remaining bins mostly empty. This is a common issue in imbalanced settings that leads to high variance in calibration estimates (Nixon et al., 2019). To better address this issue, we visualize reliability using logarithmic binning of predicted probabilities. Logarithmic scaling effectively spreads out smaller probability values, mitigating distortions caused by extreme class imbalance and providing a clearer representation of how accurately our models capture rare, high-claim events.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

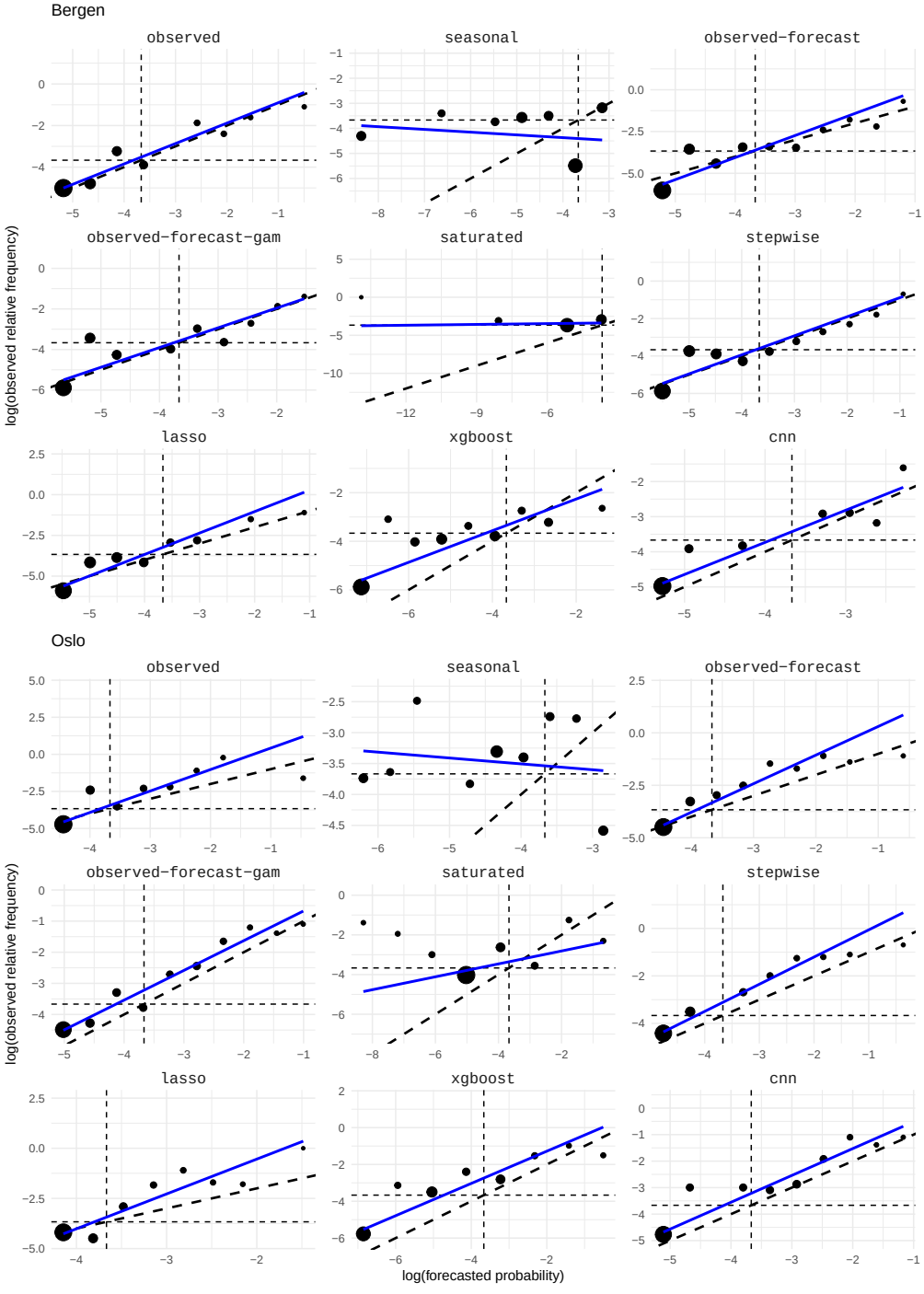

Figure 5. Log-reliability curves for the models listed in Table 2. The top three rows report reliability for the models estimated for the Bergen area, while the bottom three rows report reliability for the models estimated for the Oslo area.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

See Figure 5 for the resulting log-reliability plots. The top three rows report results from each of the models applied to Bergen, and the bottom three rows report results from each of the models applied to Oslo. Note that the baseline model `observed` exhibits high reliability, particularly in the case of Bergen. The opposite baseline model `unconditional` is not included in this figure, as that would only result in one single point in the plot (indicating a perfectly reliable model in the sense that the point $(\widehat{p},\widehat{p})$, where $\widehat{p}$ is the observed unconditional frequency of claim events, is located exactly on the diagonal).

Forecasts that exhibit higher discrimination are also more reliable. Models previously identified as having lower predictive power (`seasonal`, `saturated`, and to a certain extent the `xgboost`, as discussed in the preceding subsection) also demonstrate lower reliability. Conversely, models with strong discriminatory ability typically yield reliable predictions across both cities and both modeling strategies. Specifically, for Bergen, the `observed-forecast` model from Strategy 1 and the lasso model applied directly to the full ensemble forecasts (Strategy 2) are particularly reliable. In Oslo, we observe a similar pattern; in addition to the `observed-forecast` model, the `observed-forecast-gam` model from Strategy 1 and the `cnn` from Strategy 2 also produce forecasts closely aligned with the ideal diagonal line. This indicates reliable estimates of risk, which, in turn, can inform the insurer's operational decision-making, such as issuing warnings to customers and recommending precautionary measures to help reduce potential damage

4.3 Potential economic value

A third objective in predicting future claim events is to ensure that the predictions are valuable for decision-making, since forecast quality alone does not guarantee improved decisions (Murphy, 1993). While several approaches exist for evaluating decision value, and different decisions can be made based on a given forecast, here we assess the *potential economic value* of the predictions using a standard cost–loss decision framework adapted from the weather forecast literature (Katz and Murphy, 1997). The framework considers a user who must decide each day whether to take preventive action based on the forecast probability that the number of claims exceeds a given threshold, here set to two claims per day.

Let p_t denote the probability of a forecast model listed in Table 2 on day t. Preventive action incurs a cost C, while failing to act on a day when the event occurs results in a loss L. Assuming a risk-neutral decision-maker, the optimal decision rule is to take action whenever the expected loss from inaction, $p_t L$, exceeds the cost of action C, which is equivalent to acting when

$$p_t \geq \frac{C}{L} \tag{5}$$

Applying this rule across all T days yields a binary sequence of actions $a_t \in \{0,1\}$, where $a_t = 1$ denotes action and $a_t = 0$ denotes inaction. Let $y_t = 1$ if the claim threshold is exceeded on day t, and $y_t = 0$ otherwise. The expected cost (EC) associated with the forecast is then

$$EC = \frac{1}{T}\sum_{t=1}^{T}\left[C\,a_t + L\,(1-a_t)\,y_t\right]$$

The expected cost can be written directly in terms of binary forecast outcomes. Taking action contributes to expected cost through both true positives and false positives, while losses arise only from false negatives. Setting $L = 1$ allows costs to be

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

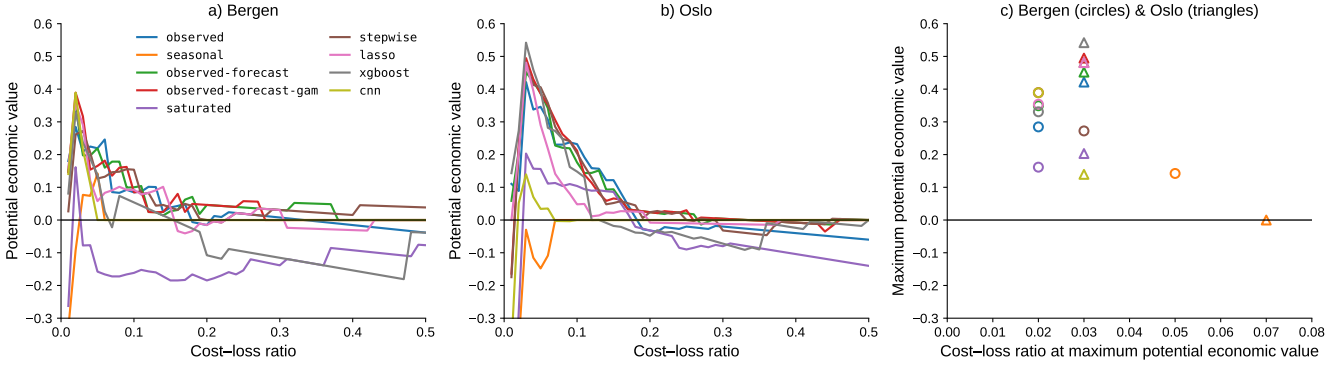

Figure 6. Potential economic value of claim forecasts relative to the unconditional baseline forecast during the test period (2020-2021) in Bergen a) and Oslo b) municipalities. c) Cost-loss at maximum potential economic value for all claim forecasts in a) and b).

expressed as a function of the cost-loss ratio $C/L = C$ alone,

$$EC = C \cdot P(\text{true positives} + \text{false positives}) + P(\text{false negatives})$$

This decomposition makes explicit how forecast discrimination and reliability (Sections 4.1 and 4.2) influence economic performance by controlling correct actions, unnecessary actions, and missed events.

To benchmark performance relative to decisions made without weather information, we define the potential economic value (PEV) of a forecast relative to the unconditional forecast,

$$PEV = 1 - \frac{EC}{EC_{\text{unc}}}$$

Positive values of PEV indicate lower expected costs relative to the unconditional forecast, while negative values indicate higher expected costs.

Figure 6a,b shows the potential economic value of the forecasts listed in Table 2 as a function of the cost-loss ratio for Bergen and Oslo. Across both cities, weather-informed claim predictions achieve positive economic value over a range of decision-making contexts, demonstrating that incorporating weather information can substantially improve decisions relative to the unconditional baseline. Notably, predictions driven by weather forecasts (e.g., `observed-forecast-gam`, purple line) perform comparably to those using the weather that actually occurred (`observed`, red line), indicating that most economically relevant information is already available in the weather forecast.

Potential economic value is highest at low cost-loss ratios (Figure 6c), occurring close to the climatological frequency of days with more than two insurance claims in each city (approximately 2% and 3% in Bergen and Oslo, respectively). At this threshold, the unconditional forecast transitions between acting and not acting according to the decision rule in Equation (5), rendering baseline decisions particularly sensitive to errors. As a result, even modest improvements in identifying high-risk days through weather-informed forecasts can substantially reduce both unnecessary actions and missed events, yielding substantial gains in potential economic value relative to the unconditional baseline.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

The magnitude of these gains differs between the two cities. Maximum potential economic value reaches approximately 30-40% in Bergen and 40-50% in Oslo (Figure 6c), reflecting higher claim forecast quality in the latter city (Figure 4 and Figure 5). Despite substantial differences in statistical structure, several prediction models achieve comparable maximum potential economic value, indicating that weather forecast input is more important than the specific choice of prediction model, provided the underlying weather forecast is skilful.

By contrast, at higher cost-loss ratios, potential economic value declines towards zero and may even become negative (Figure 6a,b). In this regime, preventive action is costly relative to potential losses, and forecasts either converge towards inaction or are penalized severely for unnecessary action. Only forecasts of very high quality can yield positive potential economic value. Therefore, for the claim forecasts considered here, the economically optimal strategy is to forego preventive action and accept the risk of loss.

The interpretation of potential economic value is most intuitive in an operational setting. Consider an insurer responsible for managing day-to-day claims. The weather forecast indicates a high likelihood of a severe storm in the next few days, and the claim prediction model assigns a high probability to a surge in insurance claims. If the cost of acting on this information is low relative to the potential loss, for example, sending an automated SMS advising homeowners to clear gutters and check drainage, then acting on the forecast can substantially reduce expected costs. By contrast, if preventive action requires costly structural interventions, the same forecast may yield no, or even negative, economic benefit from the insurer's perspective. Only a highly skilled claim forecast can be relied on to guide expensive action.

The cost-loss framework adopted here necessarily relies on simplifying assumptions. In practice, preventive actions may only partially reduce risk rather than eliminate it entirely; intervention effectiveness may vary across households; and costs and losses may differ between events and customers. The framework also neglects behavioural responses, such as whether recipients act on warnings or adapt their behaviour over repeated events. These simplifications inevitably limit this framework.

Despite these limitations, the cost-loss framework provides a transparent and interpretable means of translating probabilistic forecast skill into decision-relevant quantities. By explicitly linking forecast probabilities to expected costs across decision contexts, it enables an assessment of the *potential* economic gains from weather-informed claim predictions. As such, it complements traditional metrics of forecast quality, such as discrimination and reliability, discussed in Sections 4.1 and 4.2.

4.4 Robustness checks

We briefly summarise the robustness analysis that complements the main results; the detailed results are not included in the paper.

Alternative thresholds. Repeating the analysis with threshold $u = 1$ (i.e., defining an event as "more than one claim in a day") increases the base rate of events and slightly improves discrimination metrics across all models. It also improves reliability for the more complex models `xgboost` and `saturated`. However, the relative ranking of models is largely unchanged: weather-informed models outperform the unconditional and seasonal baselines, and the S1 models together with the regularised S2 models achieve the best performance. For $u = 3$, the event becomes extremely rare, with too few exceedances in the training and test periods to robustly estimate and evaluate models. Performance metrics become unstable and highly

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

sensitive to random variation. We therefore do not pursue this threshold further and emphasise that $u = 2$ is both operationally relevant and statistically tractable for the data at hand.

Joint city models. We also investigate the possibility of pooling data from the two cities. To this end, we re-estimate all models in Table 2 on the combined dataset, including a city indicator as a fixed effect. The resulting joint models recover the main features of the two separate city-specific models: the estimated effect of precipitation on the log-odds of a claims surge is essentially the same across cities, while the city indicator primarily shifts the intercept in the regression-type models, reflecting different baseline frequencies of events. Discrimination and reliability are comparable to those obtained from the separate models, and the relative ranking of models remains unchanged.

From a purely statistical perspective, pooling data across cities increases the effective sample size and may stabilise parameter estimates. However, given the marked differences in climatology, topography, and building practices between Oslo and Bergen, and the fact that insurers often manage these markets separately, we are reluctant to impose a common precipitation–risk relationship. We therefore retain separate city-specific models in the main analysis and view the joint models as a robustness check.

Lagged claims as predictors. Finally, we assess the impact of residual temporal dependence in the claims process by including a one-day lagged claim-surge indicator $Y_{i,t-1}$ as an additional predictor. Recall that $Y_{it} = 1$ if the claim count on day t exceeds the threshold u and $Y_{it} = 0$ otherwise. For each model in Table 2, we fit an augmented version that includes $Y_{i,t-1}$ alongside the existing weather-based predictors. Across both cities, these extended models exhibit a modest but consistent improvement in discrimination and calibration, indicating that there is some dependence on surge days beyond that captured by daily precipitation and seasonality. Importantly, however, the relative ordering of models remains the same: the weather-informed models still dominate the unconditional and seasonal baselines, and the S1 and regularised S2 models remain the best performers within their respective strategy classes.

Because the fully reconciled number of claims on day $t-1$ is typically not known at the time a forecast for day t would be issued (see Section 2.1), the lagged indicator $Y_{i,t-1}$ is not available in real time for operational forecasting. The models with lagged claims should therefore be interpreted purely as diagnostic tools for assessing potential temporal dependence, rather than as candidates for deployment in a day-ahead warning system.

5 Case studies: Extreme weather events

We now turn to a closer assessment of model performance by examining two named high-precipitation events: the storm "Birk," which brought heavy rainfall to Western Norway on December 22-23, 2017, and "Hans," a more recent and severe precipitation event that affected Eastern Norway, as well as parts of Sweden, Finland, and the Baltic countries, on August 7, 2023.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

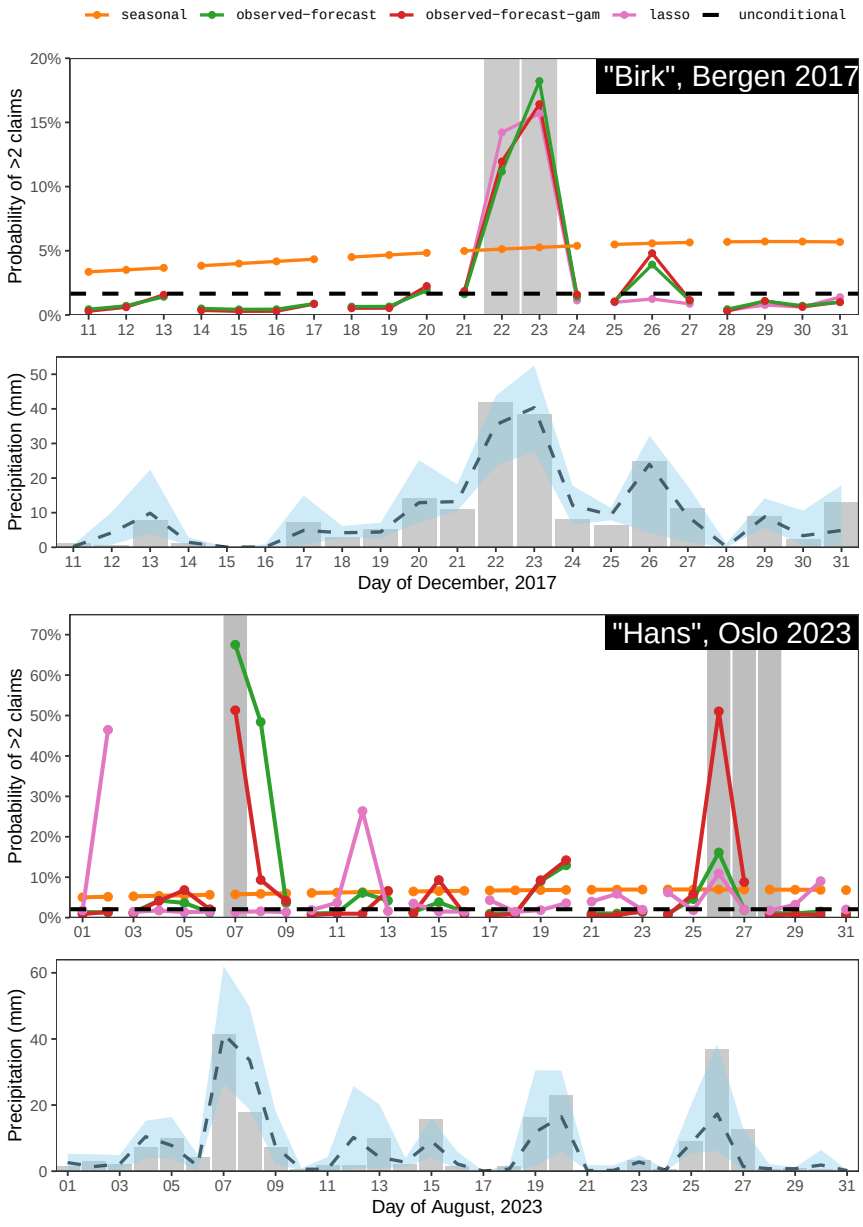

Figure 7. Observed-forecast model for days around the storm "Birk" in Bergen, 2017 (top panel) and the storm "Hans" that occurred near Oslo in 2023 (third panel from the top). The lines indicate the forecasted probabilities of more than two claims using different forecast models, where the line segments indicate a sequence of days for which the weather forecast was initiated on the first day. The grey bars indicate days on which we observed more than two claims. The second and fourth panel display the corresponding weather forecast median (dashed line) and 90% prediction interval (blue shaded region) as well as the observed daily precipitation (grey bars).

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

5.1 "Birk", Bergen, December 22-23 2017

The event "Birk" occurred within our sample period, but we have excluded December 2017 from the training data to ensure an out-of-sample evaluation. The event was well aligned geographically with our Bergen study area, making it a relevant test of model performance under extreme yet otherwise familiar conditions.

We can assess model performance during this event by examining the top two panels of Figure 7. The upper plot shows the predicted probability of observing more than two claims per day across a selection of our modeling strategies. The dashed purple line indicates the unconditional relative frequency of such events, serving as a baseline. Forecasts from the seasonal model are highlighted with red dots and connected by red lines for clarity.

Each set of forecasts initialized on the same day is connected by a solid line, illustrating the multi-day prediction horizon. For example, the first forecast shown, issued at 00:00 on Monday, December 11, covers outcomes for Monday through Wednesday. The subsequent forecast, published at 00:00 on Thursday, December 14, predicts conditions for Thursday through Sunday and is likewise represented by a connected sequence of points in the plot. This pattern is repeated throughout our sample space.

On December 22 and 23, we observe a notable increase in the predicted risk of experiencing more than two claims, compared to the lower baseline probabilities observed throughout most of the preceding period. All three non-baseline models shown in the plot forecast probabilities well above 10% for December 22, with the predicted risk rising further to exceed 15% for December 23. The actual outcome was that the insurer received more than two claims on both of these days, as indicated by the grey bars in the background of the plot.

The second panel in Figure 7 presents the corresponding weather forecasts over this period, expressed in millimetres of daily precipitation. The dashed line represents the median forecast from the ensemble for each day, while the blue shaded ribbon spans the 5th to 95th percentile range, capturing forecast uncertainty. Observed precipitation levels are shown as grey bars.

5.2 "Hans" and "Little Hans", Oslo, August 2023

We now turn to the extreme precipitation event “Hans,” which occurred on August 7, 2023, affecting large parts of Scandinavia and the Baltic countries. This event requires some additional context. Although it produced unusually heavy rainfall over Oslo, the most extreme weather occurred in the neighboring counties of Buskerud, Akershus, and Innlandet. The most significant property damage resulted from river flooding in those areas and largely fell outside our defined Oslo sampling region. To complement this case, we also include an additional event from August 26-27, 2023, informally referred to as "Little Hans" ("Lille Hans" in Norwegian), which, while less severe than "Hans", was more centrally located over Oslo and is thus more representative of our sampled area.

Importantly, the "Hans" events are out-of-time, occurring more than 2.5 years after the end of the training period. Moreover, for administrative reasons, the claim data for these cases pertain to the Oslo municipality rather than the specific nine grid cells used in our main analyses (see Figure 2 for a map showing the grid cells and the municipality borders).

Despite these additional challenges, as shown in the third panel of Figure 7, both models trained on observed weather (Strategy 1 in Table 2) correctly indicate an elevated risk of a high-claim day in Oslo on August 7. This aligns with the

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

actual outcome, as more than two claims were reported. In contrast, the lasso model, which was trained directly on the weather forecast, fails to identify heightened risk on this day. It does, however, predict increased risk on August 2 and August 12, though the final panel of the figure shows little forecasted precipitation for those days, suggesting potential overfitting to features in the forecast data and a lack of robustness to the passing of time from the training period, which ended 2.5 years before the event.

Just two weeks after "Hans," another episode of intense rainfall occurred in the Oslo area on August 26-27, leading to many water-related insurance claims. As shown in the bottom panel of Figure 7, the weather forecast did indicate some rainfall for these days, but the predicted levels were not markedly different or more severe than those for August 19-20. Most of the claim forecasting models did not respond strongly to this signal, except the `observed-forecast-gam` model, which flagged a notably elevated risk on August 26.

Despite changes in data structure and context, we find that several models continue to perform well during these periods. This suggests a degree of robustness in the modeling approach and its potential usefulness in operational forecasting and risk assessment beyond the original training context.

6 Conclusions

In this study, we have identified several key components essential for building predictive models of short-term weather-related insurance claims and for evaluating their performance. A central finding of our modeling strategy is that weather forecasts provide valuable information regarding future damage. This insight remains robust across various modeling frameworks, predictive models, and evaluation criteria. Our results lay the groundwork for insurers to develop operational models tailored to their specific portfolios and decision-making processes. While forecasts of rare, spatially clustered weather-induced losses must be evaluated with care, our framework helps insurers choose strategies that align with their perspective, whether prioritizing operational readiness or proactive customer mitigation.

To address this, we discuss our findings from the insurer's perspective and thoroughly evaluate the forecasting models in three key dimensions. First, we employ classical evaluation metrics to assess the models' ability to discriminate between events and non-events in advance. Second, we emphasize the importance of accurately quantifying risk so that forecasted probabilities of claim events can effectively inform subsequent risk modeling efforts. Finally, we illustrate how insurers can economically assess the forecasted probabilities to guide preventive actions, which can be costly but beneficial for reducing overall damage. We complement the analysis by providing two case studies: one out-of-sample and one out-of-time, which further demonstrate the robustness of our approach.

We place less emphasis on selecting a particular predictive model, opting instead to present results using a suite of standard models readily implementable with standard software. We conclude that careful and reasonable use of information from weather forecast ensembles is more important than the specific choice of model. Nevertheless, we recognize the potential for improving claim forecasts by expanding methodological approaches and investing additional computational resources in hyperparameter tuning. Such efforts may yield varying results based on geographical, temporal, and legal contexts. Future

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

research could also investigate how integrating additional risk factors, such as forecasts of wind, flooding, and other weather hazards, alongside detailed information on local vulnerabilities and risk exposures, can further enhance our understanding of the evolving risks faced by the insurance sector in a changing climate, and how it should better prepare for the future.

Code and data availability. The insurance data used in this study is not publicly available. However, code and a simulated dataset from the `observed-forecast` model is available at https://github.com/holleland/ForecastingInsuranceClaims.

Appendix A: CNN structure

In this appendix, we include the structure of the CNN models used for Bergen and Oslo, respectively.

Bergen

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 47, 16)            96

 max_pooling1d (MaxPooling1D  (None, 23, 16)           0
 )

 conv1d_1 (Conv1D)           (None, 17, 32)            3616

 max_pooling1d_1 (MaxPooling  (None, 8, 32)            0
 1D)

 conv1d_2 (Conv1D)           (None, 4, 80)             12880

 max_pooling1d_2 (MaxPooling  (None, 2, 80)            0
 1D)

 flatten (Flatten)           (None, 160)               0

 dense (Dense)               (None, 288)               46368
```

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026


```
 dropout (Dropout)           (None, 288)               0

 dense_1 (Dense)             (None, 1)                 289

=================================================================
Total params: 63,249
Trainable params: 63,249
Non-trainable params: 0
_________________________________________________________________
```

Oslo

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 47, 80)            480

 max_pooling1d (MaxPooling1D  (None, 23, 80)           0
 )

 conv1d_1 (Conv1D)           (None, 21, 112)           26992

 max_pooling1d_1 (MaxPooling  (None, 10, 112)          0
 1D)

 conv1d_2 (Conv1D)           (None, 4, 64)             50240

 max_pooling1d_2 (MaxPooling  (None, 2, 64)            0
 1D)

 flatten (Flatten)           (None, 128)               0

 dense (Dense)               (None, 416)               53664
```

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026


```
 dropout (Dropout)           (None, 416)               0

 dense_1 (Dense)             (None, 1)                 417

=================================================================
Total params: 131,793
Trainable params: 131,793
Non-trainable params: 0
_________________________________________________________________
```

Author contributions. Håkon Otneim: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. Etienne Dunn-Sigouin: Conceptualization, Methodology, Validation, Investigation, Resources, Software, Data curation, Writing – original draft. Sondre Hølleland: Methodology, Formal analysis, Validation, Writing – review & editing, Software, Data curation. Mahsa Gorji: Investigation, Writing – review & editing. Geir Drage Berentsen: Conceptualization, Writing – review & editing. All authors approved the final manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank Sondre Lykke Rødal for his preliminary analysis of the problem in his master's thesis with Mahsa Gorji. This work was supported by the Norwegian Research Council, which supported this work through the Centre for Research-Based Innovation *Climate Futures*, project number 309562; Some computations were performed on resources provided by the Sigma2 NS9873K project, as part of the National Infrastructure for High-Performance Computing and Data Storage in Norway.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

References

Abrego-Perez, A. L. and Nuñez-Mora, J. A.: Climate risk, policy, and insurance: a forecast-based model for weather index design in vulnerable economies, International Economics and Economic Policy, 23, 10, https://doi.org/10.1007/s10368-025-00695-3, 2026.

Buizza, R., Alonso Balmaseda, M., Brown, A., English, S. J., Forbes, R., Geer, A., Haiden, T., Leutbecher, M., Magnusson, L., Rodwell, M., Sleigh, M., Stockdale, T., Vitart, F., and Wedi, N.: The Development and Evaluation Process Followed at ECMWF to Upgrade the Integrated Forecasting System (IFS), ECMWF Technical Memorandum 829, European Centre for Medium-Range Weather Forecasts (ECMWF), https://doi.org/10.21957/xzopnhty9, 2018.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, Association for Computing Machinery, https://doi.org/10.1145/2939672.2939785, 2016.

DeGroot, M. H. and Fienberg, S. E.: The Comparison and Evaluation of Forecasters, Journal of the Royal Statistical Society. Series D (The Statistician), 32, 12–22, https://doi.org/10.2307/2987588, 1982.

Dey, A. K., Lyubchich, V., and Gel, Y. R.: Modeling Weather-induced Home Insurance Risks with Support Vector Machine Regression, arXiv preprint arXiv:2103.08761, https://doi.org/10.48550/arXiv.2103.08761, 2021.

Dunn-Sigouin, E., Kolstad, E. W., Wulff, O., Parker, D. J., and Keane, R. J.: Balancing accuracy versus precision: Enhancing the usability of sub-seasonal forecasts, Climate Services, https://doi.org/10.1016/j.cliser.2025.100594, 2025.

European Centre for Medium-Range Weather Forecasts (ECMWF): Access to forecasts, https://www.ecmwf.int/en/forecasts/accessing-forecasts, accessed: 2024-09-05, 2024.

European Commission: EU Taxonomy for Sustainable Activities, https://finance.ec.europa.eu/sustainable-finance/tools-and-standards/eu-taxonomy-sustainable-activities_en, accessed 2024-09-05, 2020.

Finance Norway: Klimarapport 2024, Tech. rep., Finance Norway, https://www.finansnorge.no/contentassets/2d9eee6b15d3417280ce8a3a7cd76976/klimarapport-2024.pdf, in Norwegian; English title: Climate Report 2024; accessed 2024-10-16, 2024.

Flavelle, C. and Rojanasakul, M.: Insurers Are Deserting Homeowners as Climate Shocks Worsen, The New York Times, https://www.nytimes.com/interactive/2024/12/18/climate/insurance-non-renewal-climate-crisis.html, 2024.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, Cambridge, MA, ISBN 9780262035613, https://www.deeplearningbook.org/, 2016.

Hastie, T. J. and Tibshirani, R. J.: Generalized Additive Models, Chapman and Hall, London, ISBN 9780412343902, 1990.

Haug, O., Dimakos, X. K., Vårdal, J. F., Aldrin, M., and Meze-Hausken, E.: Future Building Water Loss Projections Posed by Climate Change, Scandinavian Actuarial Journal, 2011, 1–20, https://doi.org/10.1080/03461230903266533, 2011.

Heinrich-Mertsching, C., Wahl, J. C., Ordoñez, A., Stien, M., Elvsborg, J., Haug, O., and Thorarinsdottir, T. L.: Assessing present and future risk of water damage using building attributes, meteorology, and topography, Journal of the Royal Statistical Society Series C: Applied Statistics, 72, 809–828, https://doi.org/10.1093/jrsssc/qlad043, 2023.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present, https://doi.org/10.24381/cds.adbb2d47, [data set], 2023.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

Hettiarachchi, S., Wasko, C., and Sharma, A.: Increase in flood risk resulting from climate change in a developed urban watershed–the role of storm temporal patterns, Hydrology and Earth System Sciences, 22, 2041–2056, https://doi.org/10.5194/hess-22-2041-2018, 2018.

IPCC: Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, ISBN 9781009325844, https://doi.org/10.1017/9781009325844, 2023.

Jaison, A., Michel, C., Sorteberg, A., and Breivik, Ø.: Towards Impact-Based Forecasting of Storm-Damages Using Locally Calibrated Damage Functions, Meteorological Applications, 32, e70 087, https://doi.org/10.1002/met.70087, 2025.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, John Wiley & Sons, 2 edn., ISBN 9780470660713, https://doi.org/10.1002/9781119960003, 2012.

Katz, R. W. and Murphy, A. H., eds.: Economic Value of Weather and Climate Forecasts, Cambridge University Press, ISBN 9780511608278, https://doi.org/10.1017/CBO9780511608278, 1997.

Keys, B. J. and Mulder, P.: Property Insurance and Disaster Risk: New Evidence from Mortgage Escrow Data, NBER Working Paper 32579, National Bureau of Economic Research, https://doi.org/10.3386/w32579, 2024.

LeClerc, J. and Joslyn, S.: The Cry Wolf Effect and Weather-Related Decision Making, Risk Analysis, 35, 385–395, https://doi.org/10.1111/risa.12336, 2015.

Lyubchich, V., Kilbourne, K. H., and Gel, Y. R.: Where Home Insurance Meets Climate Change: Making Sense of Climate Risk, Data Uncertainty, and Projections, Variance, 12, 278–292, https://variancejournal.org/article/127582-where-home-insurance-meets-climate-change-making-sense-of-climate-risk-data-uncertainty-and-projections, 2019a.

Lyubchich, V., Newlands, N. K., Ghahari, A., Mahdi, T., and Gel, Y. R.: Insurance Risk Assessment in the Face of Climate Change: Integrating Data Science and Statistics, Wiley Interdisciplinary Reviews: Computational Statistics, 11, e1462, https://doi.org/10.1002/wics.1462, 2019b.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I. V., Feser, F., Koszalka, I., Kreibich, H., et al.: Impact forecasting to support emergency management of natural hazards, Reviews of Geophysics, 58, e2020RG000 704, https://doi.org/10.1029/2020RG000704, 2020.

Mills, E.: Insurance in a climate of change, Science, 309, 1040–1044, https://doi.org/10.1126/science.1112121, 2005.

Moemken, J., Messori, G., and Pinto, J. G.: Windstorm losses in Europe–What to gain from damage datasets, Weather and Climate Extremes, 44, 100 661, https://doi.org/10.1016/j.wace.2024.100661, 2024.

Murphy, A. H.: What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, Weather and Forecasting, 8, 281–293, https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, 1993.

Nixon, J., Dusenberry, M., Zhang, L., Jerfel, G., and Tran, D.: Measuring Calibration in Deep Learning, CoRR, abs/1904.01685, https://doi.org/10.48550/arXiv.1904.01685, 2019.

Norwegian Natural Perils Pool: Årsrapport 2023, Tech. rep., Norwegian Natural Perils Pool, https://res.cloudinary.com/forsikringsdrift/image/upload/%C3%85rsrapport-2023-NNP-signed.pdf, in Norwegian; annual report for 2023, 2024.

Pardowitz, T., Osinski, R., Kruschke, T., and Ulbrich, U.: An Analysis of Uncertainties and Skill in Forecasts of Winter Storm Losses, Natural Hazards and Earth System Sciences, 16, 2391–2402, https://doi.org/10.5194/nhess-16-2391-2016, 2016.

Rohrbeck, C., Eastoe, E. F., Frigessi, A., and Tawn, J. A.: Extreme Value Modelling of Water-Related Insurance Claims, The Annals of Applied Statistics, 12, 246–282, https://doi.org/10.1214/17-AOAS1081, 2018.

https://doi.org/10.5194/egusphere-2026-1169
Preprint. Discussion started: 31 March 2026

Röösli, T., Appenzeller, C., and Bresch, D. N.: Towards operational impact forecasting of building damage from winter windstorms in Switzerland, Meteorological Applications, 28, e2035, https://doi.org/10.1002/met.2035, 2021.

Scheel, I. and Hinnerichsen, M.: The Impact of Climate Change on Precipitation-Related Insurance Risk: A Study of the Effect of Future Scenarios on Residential Buildings in Norway, The Geneva Papers on Risk and Insurance - Issues and Practice, 37, 365–376, https://doi.org/10.1057/gpp.2012.7, 2012.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., and Meze-Hausken, E.: 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, 85–100, https://doi.org/10.1111/j.1467-9876.2012.01039.x, 2013.

Shi, Y.: Assessing the dependence between extreme rainfall and extreme insurance claims: A bivariate peak over threshold method, Risk Analysis, 45, 2504–2520, https://doi.org/10.1111/risa.70033, epub 2025-04-16, 2025.

Shi, Y., Berentsen, G. D., and Otneim, H.: Insurance in a Changing Climate: A Retrospective Study of Water-Related Claims and Pricing Strategies in Norway, Discussion Paper 2025/3, Department of Business and Management Science, NHH Norwegian School of Economics, https://ideas.repec.org/p/hhs/nhhfms/2025_003.html, 2025a.

Shi, Y., Punzo, A., Otneim, H., and Maruotti, A.: Hidden semi-Markov models for rainfall-related insurance claims, Insurance: Mathematics and Economics, 120, 91–106, https://doi.org/10.1016/j.insmatheco.2024.11.008, 2025b.

The Geneva Association: Climate Change Risk Assessment for the Insurance Industry, Tech. rep., The Geneva Association, 2021.

Tibshirani, R.: Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58, 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x, 1996.

Wahl, J. C., Aanes, F. L., Aas, K., Frøyn, S., and Piacek, D.: Spatial modelling of risk premiums for water damage insurance, Scandinavian Actuarial Journal, 2022, 216–233, https://doi.org/10.1080/03461238.2021.1951346, 2022.

Welker, C., Röösli, T., and Bresch, D. N.: Comparing an insurer's perspective on building damages with modelled damages from pan-European winter windstorm event sets: a case study from Zurich, Switzerland, Natural Hazards and Earth System Sciences, 21, 279–299, https://doi.org/10.5194/nhess-21-279-2021, 2021.