Review report 2 on "Short-Term Management of Water-Damage Claim Risk Using Ensemble Precipitation Forecasts

The manuscript examines whether ensemble precipitation forecasts can help an insurer anticipate days with unusually high numbers of water-damage claims in Bergen and Oslo. The focus on a practical operational threshold is useful, and I appreciate the comparison between models trained on realised precipitation and models trained directly on forecast information. Linking the predictions to a decision problem also gives the study a clearer purpose than a comparison of classification scores alone. I recommend major revision. The results suggest that precipitation forecasts contain useful information about claim surges, but the manuscript does not yet describe the validation well enough to assess the strength of that finding. There is also an inconsistency in the definition of Strategy 1, several errors in the interpretation of verification metrics, and insufficient support for the stated economic benefits.

My main comments are below.

1. Section 2 and Figure 6: Please state the exact training, validation and test periods, with the number of surge days in each city. The stated 2013–2021 period and the total of 2,921 days also need reconciliation. Explain how missing dates and exclusions were handled.

2. Section 3.2 and Table 2: Strategy 1 is defined by averaging claim probabilities across all ensemble members, whereas Table 2 describes using the ensemble median. These are different calculations for a nonlinear model. Please clarify which was implemented and make the description consistent.

3. Lines 147–151 and 456–459: Please distinguish forecast initialisation, actual availability and the target day. The Birk example includes predictions for the issue day, which needs clarification given the stated one-to-four-day horizon. Reporting claim-prediction skill separately by lead time would also be useful.

4. Figure 4: The number of surge days is small, and consecutive days may belong to the same storm. Please explain the bootstrap procedure and preserve prediction–outcome pairs and temporal dependence. The wide confidence intervals also call for caution when comparing models.

5. Lines 265–284: Predicting the majority class every day does not produce a high ROC AUC; a constant prediction gives an AUC of 0.5. Please correct this explanation. PRAUC is appropriate here, but a Brier skill score would provide useful information about probability accuracy.

6. Section 4.2 and Figure 5: A constant probability estimated from training data is not necessarily perfectly calibrated in the test period. Please correct this statement and report the calibration-bin definitions, sample sizes and uncertainty. Explain how bins with no events were treated before taking logarithms.

7. Section 4.3: Preparing staff and preventing property damage are different decisions. Please define what the loss L represents and explain the assumption that action removes it entirely. The reported 30–50% savings should be presented as potential benefits under the assumed decision model. A sensitivity check for partial effectiveness would strengthen this section.

8. Section 2.1: Please clarify the geographical selection of claims and whether the insured portfolio changed during the study. Since claims are assigned by damage-occurrence date, also explain how

closely this target represents the incoming claims workload that the proposed system is intended to anticipate.

9. Sections 2.2–2.3: Daily precipitation averaged over nine grid cells may smooth local rainfall intensities relevant to urban damage. Please discuss this limitation, including antecedent wetness and the inclusion of snowfall. A limited comparison with independent rainfall observations during major claim episodes would be helpful.

10. Section 3.3 and Appendix A: More information is needed on model fitting, hyperparameter selection and regularisation, particularly for the CNNs given the small event sample. Please confirm that these choices were made independently of the final test results.

11. Section 5: Please give the training dates used for Birk and clarify whether later years contributed to model fitting. For the 2023 cases, report actual claim counts and discuss the missed events more directly. These selected examples alone provide limited evidence of performance beyond the training period.

12. Section 4.4: Please include the numerical robustness results in a short supplement. Also describe the observed model as a retrospective reference rather than a guaranteed upper bound on skill. Similar scores do not establish that the forecasts contain nearly all useful weather information.