August 10, 2026

Dear Editor and Reviewers,

We sincerely thank the editorial team and the reviewers for their time and constructive feedback during the review process.

Please find below our point-by-point response to the minor revision comments, as well as the revised manuscript and a version with the changes highlighted (attached) of our work titled "**Using machine learning for the prediction of flood-related 112 calls**".

Sincerely,
The authors

Point-by-Point Response

Introduction:
Thank you for the effort to improve the manuscript by addressing the comments from the editor and reviewers. The revised manuscript has improved greatly, and most of the reviewers' comments have been addressed.
However, there are still some concerns regarding the evaluation of the predictive skill of the ML approach, as well as the uncertainty analysis. Therefore, I ask the authors to carefully consider and address the three concerns below before the manuscript can be considered for publication.

Response:
We would like to once again thank the reviewers for their valuable feedback, which has significantly improved the clarity and overall quality of our manuscript. We also extend our gratitude to the editor for the additional comments provided and for managing the review process.
In our point-by-point responses below, we hope to have satisfactorily addressed the remaining points to make our manuscript suitable for publication.

Comment 1. Prediction without recent call history versus continuation prediction:
The rainfall-stage analysis does not directly show how the model performs when the recent call-history predictors are zero. Rainfall onset is not equivalent to the first reported impact, and previous calls may already be present in samples classified as rainfall "start."
Please report results separately for samples with no calls during the preceding three hours and samples with at least one call during that period. Sample counts, positive cases, CSI, recall, and precision should be provided. If this analysis is not included, the manuscript should be framed more clearly as impact nowcasting and tracking, and the abstract should state that recent emergency calls are used as predictors.

Response:
This comment raises a valuable point, and we now agree that reporting separate results for samples with and without calls received in the previous three hours introduces a different perspective to the rainfall stage analysis, and in general to the notion of how using recent impact data affects performance. We have conducted the suggested analysis, however we opted to include it in the appendix (Appendix A) rather than the main text as the findings do not introduce fundamentally new results, but rather compliment existing ones, again confirming both that our models benefit from utilizing this data and that the task of predicting the instance appearance of impacts is distinct from (and more challenging) predicting the following ones. Below, we provide the text for this new section, along with the supporting figure and table:

> " **Section title:** Presence of emergency calls in recent hours
> In Sects. 6.2.2, 6.2.3, and 6.2.4 we found that the use of recent calls as a predictor contributes significantly to model performance. In this section, we analyse this further by presenting metrics separately for cases where emergency calls were received in the preceding three hours (which serve as input features for the model) and cases where none were received. These results, alongside general model performance, are shown in Figure A1 as the average across temporal configurations, with error bars indicating the standard deviation.
> For the high-density group, results are highest in cases where calls were received, achieving a CSI of 41.84% compared to 11.10% without recent calls. This trend is consistent across all three metrics and aligns with the results of Sect. 6.2.4, as at the start of a rainfall event,

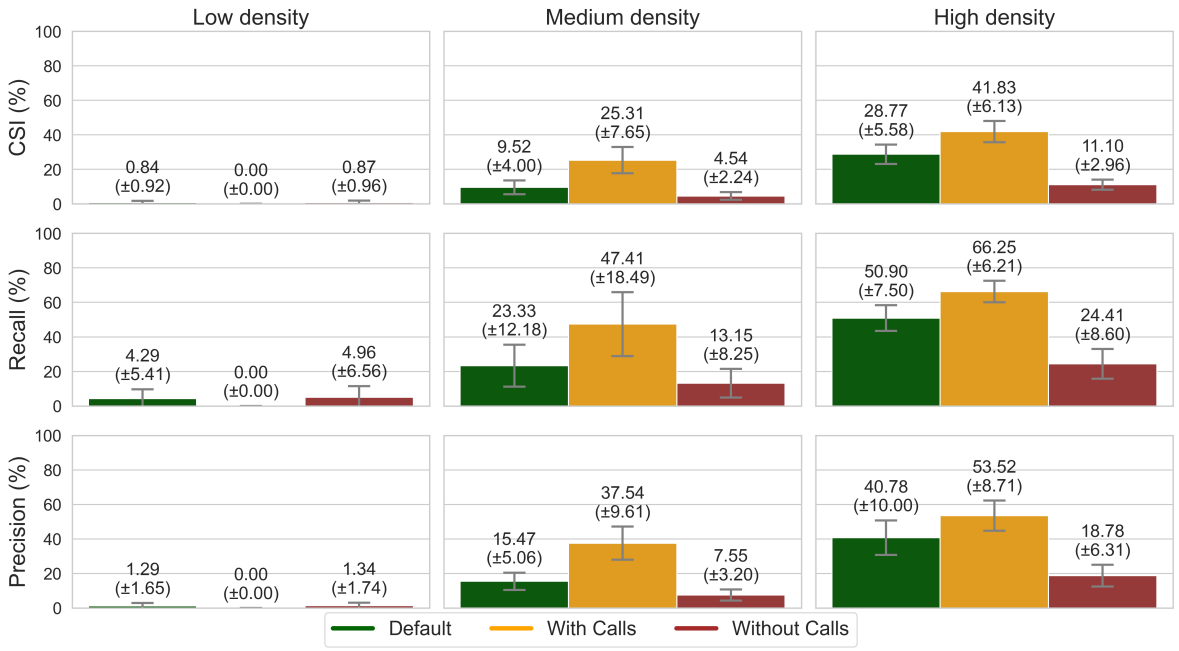

Figure A1: Performance of the ML approaches across population density groups for all data ("Default"), samples with recent emergency calls ("With Calls"), and samples without recent emergency calls ("Without Calls"). Metrics reflect the mean and standard deviation computed across test sets for all temporal configurations.

> where generally there are no preceding impacts recorded, performance is worse.
> We perform the same study with the model trained without previous call features (from Sect. 6.2.3) and identify a similar pattern of superior performance in situations with continuing impacts (CSI of 27.00% ± 6.05% in cases with recent calls versus 8.93% ± 3.20% without) for the high density model. This indicates that while access to recent call history drives significant performance improvements, situations with lingering impacts are inherently easier to predict (coinciding with persistent rainfall that can saturate the soil and drainage systems), even when the model lacks explicit data on these ongoing impacts.
> Moving to medium-density populations, we still find that situations with ongoing impacts yield better results. However, the same cannot be said for the low-density group, which achieves a CSI of 0.00%±0.00%. This discrepancy can be explained by examining the counts of positive cases found in Table A1. In the high-density group, positive samples are more frequent in situations where calls were already received. In contrast, for the medium- and low-density groups, positive samples occur predominantly in cases without recent impacts, indicating that lingering impact events are less common in lower-density regions. Although such events are less frequent in the medium-density group, the model still performs well when they occur. For low-density municipalities, however, these situations are extremely rare, averaging only 2.5 ± 1.96 positive samples with recent impacts across the evaluated time periods. "

This new appendix is referenced in Sect. 6.2.4, as a way of continuing understanding how the "Calls in Previous Hours" features contribute to performance (Lines 458-459):

> ...For further discussion on the influence of incorporating recent emergency calls as predictors, please refer to Appendix A.

Table A1: Sample counts and positive cases for scenarios with and without recent emergency calls across population density groups (mean ± standard deviation across all temporal configurations).

Density Group	Category	Total Samples	Positive Samples
Low	With	58.5 ± 4.1	2.5 ± 2.0
	Without	$622,828.8 \pm 217,341.7$	18.5 ± 2.0
Medium	With	180.6 ± 15.6	22.0 ± 7.9
	Without	$654,454.9 \pm 226,606.0$	50.0 ± 7.9
High	With	$1,192.8 \pm 197.9$	70.5 ± 7.1
	Without	$589,566.7 \pm 260,200.7$	40.5 ± 7.1

Also, we have made a small modification to the abstract to mention the use of recent emergency calls as predictors (Lines 6-7):

> "...Specifically, we develop a model that combines rainfall-related weather data, information on recent emergencies, and static vulnerability-exposure layers to predict..."

Comment 2. Retrospective rainfall filtering
The daily rainfall filter is based on full-day observed rainfall. Consequently, the reported results cannot be directly interpreted as performance under continuous operational deployment, particularly the model's false-alarm behaviour during excluded days.
I suggest evaluating at least the high-density model over a continuous, unfiltered test period. The model may still be trained using filtered data. If this is not feasible, claims concerning operational readiness should be moderated and presented as requiring prospective validation.

Response:
Thank you for raising this point. We agree that our performance metrics do not fully reflect how the models would behave in a continuous operational deployment. Due to computational and time constraints associated with reprocessing weather radar products, we limited our evaluation to days where rainfall was observed, as these represent the most critical events. However, we acknowledge that not testing the models on dry or low-rainfall days leaves potential false alarms unassessed. To account for this, we have revised our discussion of operational readiness in the manuscript to explicitly highlight this limitation.
Specifically, we have made the following revisions:

- **Section 4.4 (Lines 293-295)**: "...Therefore, please note that the metrics presented in the following sections reflect performance strictly under these conditions [the daily rain filter] rather than over a continuous operational deployment."
- **Section 7.1 (Lines 531-532)**: "...Additionally, evaluating performance across a continuous time period is necessary to better emulate operational deployment, as false alarms could potentially arise from the excluded days with little to no rainfall."

Comment 3. Uncertainty under temporally clustered impacts
The current bootstrap resamples individual municipality-hour observations, although many observations may belong to the same rainfall episode. This can underestimate uncertainty.

A block bootstrap by date or rainfall episode would be more appropriate. Otherwise, the limitation of the current bootstrap should be stated explicitly, and the phrase "statistically robust metrics" should be removed. The interpretation of Figure 7 should also be revised because within-configuration bootstrap variability and variability across temporal configurations represent different sources of variation.
One additional qualification is needed when comparing population-density groups. Because different call thresholds define the positive class, differences in performance should not be attributed solely to population density. A more precise statement would be:
Under the group-specific impact thresholds adopted in this study, predictive performance was highest in the high-density group.

Response:
We agree that our initial bootstrap strategy was limited due to temporal dependencies among individual observations belonging to the same rainfall event. To mitigate this temporal correlation, we have adopted the suggested block bootstrap by date. Specifically, we decided on 1-day blocks, as the majority of rainfall events in the region typically occur within a single day rather than extending across multiple days. We have changed the description of the method accordingly (Lines 340-345):

> "Specifically, we follow a 1-day block bootstrap strategy on the test set to generate 250 random samples with replacement, compute performance metrics for each, and take the standard deviation as a measure of uncertainty. To maintain a consistent positive-to-negative class balance across resamples, we fixed the number of sampled "positive days" (i.e., days containing at least one positive instance). We opted for a block bootstrap to account for the temporal correlation between observations belonging to the same rainfall event, using 1-day blocks since the majority of rainfall events in the region last no longer than one day."

We have updated Figure 7 following this strategy. As the reviewer correctly suspected, our previous approach was indeed underestimating uncertainty. The revised standard deviation intervals across the different temporal configurations are now wider and more consistent with the overall variability over all cross-validation windows. This update confirms that the previously narrower within-configuration intervals were an "artifact" of the municipality-hour resampling. With event-level dependencies accounted for, we believe that the sampling uncertainty is now more realistically estimated.
Additionally, we also agree with the reviewer that bootstrap variability and cross-validation variability reflect distinct sources of variation, with the first capturing event-level dynamics within a given evaluation period, while the latter captures performance fluctuations across different temporal configurations. Below is the revised text regarding the interpretation of Figure 7, which incorporates all the discussed points (Lines 369-378):

> "In Fig. 7, we show the performance of the high-density model across the different temporal configurations (test set results), along with a bootstrap variance estimate for each test set. Additionally, we display the overall mean (dotted line) and standard deviation (grey strip) across all configurations. Although no clear trend is observed, we note how the standard deviation intervals of individual configurations (indicated by the error bars) closely align with the overall performance variability across all three metrics. This suggests that once intra-event correlation is accounted for through block bootstrapping, uncertainty within specific time windows matches the overall variability across all temporal configurations. It should be noted, however, that these two types of intervals reflect different sources of uncertainty, with bootstrapping capturing intra-event dynamics, while the overall standard deviation captures performance fluctuations across temporal configurations."

Lastly, we have revised the population-density group result comparison to address the difference in impact thresholds (Lines 363-366):

"It is worth mentioning that these results are obtained using the group-specific impact thresholds from Sect. 4.1, so population density is not the only factor influencing performance. However, because the low- and medium-density groups both share the same impact threshold of 1 call, the superior results of the medium-density model confirm the underlying effect of population density."