June 22, 2026

Dear Reviewer #2,

We would like to thank you for your taking the time to review our manuscripts and providing valuable feedback.

Please find below our point-by-point response to your comments on our work titled "**Using machine learning for the prediction of flood-related 112 calls**".

Sincerely,
The authors

Point-by-Point Response to Reviewer #2

General Comments:
This manuscript presents a timely and practically relevant study on the use of machine learning to predict flood-related 112 emergency calls at municipal and hourly resolution in Catalonia. The topic fits well within the scope of impact-based forecasting and early-warning systems, and the use of emergency-call records as impact proxies is an interesting contribution. The manuscript is generally well structured, and the comparison with operational warning products makes the work relevant for both scientific and applied audiences.
However, I recommend major revision before publication. The study has a promising core, but several methodological and interpretive issues need to be clarified before the conclusions about operational impact prediction can be fully supported.
Recommendation
Overall, the manuscript addresses an important and promising topic, and the dataset is valuable. I recommend major revision, mainly to clarify the operational forecasting setting, improve baseline comparability, strengthen the evaluation, and moderate the interpretation of results across population-density groups.

Response:
We are grateful to Reviewer #2 for highlighting the relevance of our study to the journal and for their constructive feedback. With the changes described below, we hope our results are now clearer, and the different comparisons throughout come out as fairer.

Major comment 1. Clarify the operational forecasting setting, especially the use of previous emergency calls as predictors:
The manuscript presents the task as predicting whether flood-related impacts will occur in the next hour, for example in the Abstract and in Sect. 4.1, where the target is defined as the occurrence of one or more flood-related 112 calls in the following hour. However, Sect. 4.3.4 states that the model includes the number of calls received 1, 2, and 3 hours earlier as input features. This means that the model may perform particularly well when impacts have already started, rather than when the first impact is still to occur. This distinction is important for the claimed contribution to early warning. I suggest that the authors separate model performance for "first-impact" situations, where no calls have occurred in the preceding hours, and "continuation" situations, where calls have already been received. An additional experiment excluding previous-call features would also help clarify how much predictive skill comes from hydrometeorological and vulnerability-exposure information alone.

Response:
We appreciate the reviewer's valuable insights regarding the operational validity of our methodology. Regarding the use of emergency call data from the previous 1 to 3 hours as predictors, the reviewer's intuition that these features contribute significantly to the models' performance is correct. To quantify this, we conducted an ablation experiments in Section 6.2.3 (which we believe contains the experiment that the reviewer proposes), where the models were trained excluding previous-call data, and observed a substantial reduction in CSI for the high-density model of -11.61% (we have explicitly added this percentage in the text for clarity), allowing us to identify these features as the most influential. We chose to maintain them in the final configuration because this data would be available during an operational deployment, potentially benefiting situations that require tracking unfolding impacts.
To address the distinction between predicting the "first-impact" versus their continuation, we conducted the study in Section 6.2.4 (Model performance depending on rain stage). By evaluating the

performance of our high-density model (which obtained the best and most stable results along with consistent explanations) across different rainfall stages, we believe the reviewer's operational concerns are addressed. Specifically, we identify a performance drop at the start of a rainfall event, which effectively serves as a proxy for the "first-impact" scenario, compared to subsequent stages. While this is not exactly the same as predicting the first occurrence of an impact, it targets an equivalent operational question: assessing model performance when a potentially impactful event is starting. Furthermore, we found that using a physical magnitude (i.e., rainfall) instead of the appearance of the first impacts results in a more consistent definition of what an "event" is, as in general it is uncommon to receive emergency calls on consecutive hours outside of highly populated cities.
Additionally, in this same experiment, we also show the performance of a model trained exclusively with the number of calls from the previous 1 to 3 hours. We observe how this baseline is consistently outperformed by the "full" model, which also has lower variance (except in the "start" stage, where large SD intervals for both approaches make it hard to definitely assert which one performs better there).
From all of this, we can conclude that recent impacts are indeed the most effective predictors in isolation (particularly for the best-performing, high-density model), and that the model heavily relies on them in many instances. But, although predicting the onset of impacts is arguably a harder task than predicting their continuation (since usually no calls have been received recently), the ML approach shows comparably better results than traditional baselines in this situation. Furthermore, we find that the model can effectively leverage other types of features simultaneously to provide better predictions across all stages of a rainfall event. We have extended our discussion (Section 7) to incorporate some of the discussed valuable ideas regarding operationality:

> "Additionally, we examined the behaviour of the different approaches during different phases of a rainfall event. We observe that all approaches struggle during the first hour of rainfall, showing reduced performance compared to that of following hours. This is expected, as predicting the onset of impacts is arguably more challenging than predicting their continuation. However, the high-density model still obtains better results than traditional hazard-based methods in this situation, and once the event has started, the performance raises and is maintained even after the rainfall stops, showing promising results for tracking unfolding impacts. Pairing this with explainable AI techniques like SHAP..."

Major comment 2. Improve the fairness and transparency of the baseline comparison:
Sect. 5.2 compares the machine-learning model with AEMET warnings, SMC warnings, and the FF-EWS radar-based warning product. This is useful, but the comparison may not be fully balanced. In Sect. 4.3.2, the manuscript states that all SMC warnings are used as model input, including observation warnings, whereas Sect. 5.2 says that observation warnings are excluded from the SMC warning baseline because they are reactive. If reactive observation warnings are available to the machine-learning model but not to the baseline, the model may benefit from information that is closer to real-time event detection than prospective forecasting. The authors should clarify exactly which warning products are available at prediction time and provide a sensitivity experiment using only prospectively available warning information. This would strengthen the claim that the machine-learning model outperforms operational warning approaches.

Response:
The reviewer provides a valuable observation with this comment. We compare our approach against these systems because they are the ones used operationally by emergency centers and civil protection agencies. For instance, the official AEMET warnings are part of the Meteoalarm program, following standards used across European weather agencies. The FF-EWS baseline is a more advanced product used by civil protection authorities in Catalonia, our target region.

Agreeing with the concern that the model may have access to more information than the official SMC warnings baseline, we have recomputed this same baseline to include all warnings (both prediction and observation) to be consistent with the data provided to the model. Initially, we only considered prediction warnings as the baseline because they have a longer forecast horizon than observation warnings, which makes them more relevant for emergency responders. However, since our ML models have a short prediction horizon, considering observation warnings as baselines is indeed more appropriate. We have changed the baseline definition by removing the part that mentions that observation warnings were ignored, and updated the results from Figure 8.
We attach Figure 1 below, which shows a performance comparison between the different types of SMC warnings (with "prediction only" warnings being the previous baseline, and "SMC Warnings (all)" the new one). Ultimately, the new results are not significantly different, as performance is relatively poor in all cases and standard deviations are large. In the study, the FF-EWS baseline continues to be the most promising hazard-based approach, so the conclusions remain unchanged.

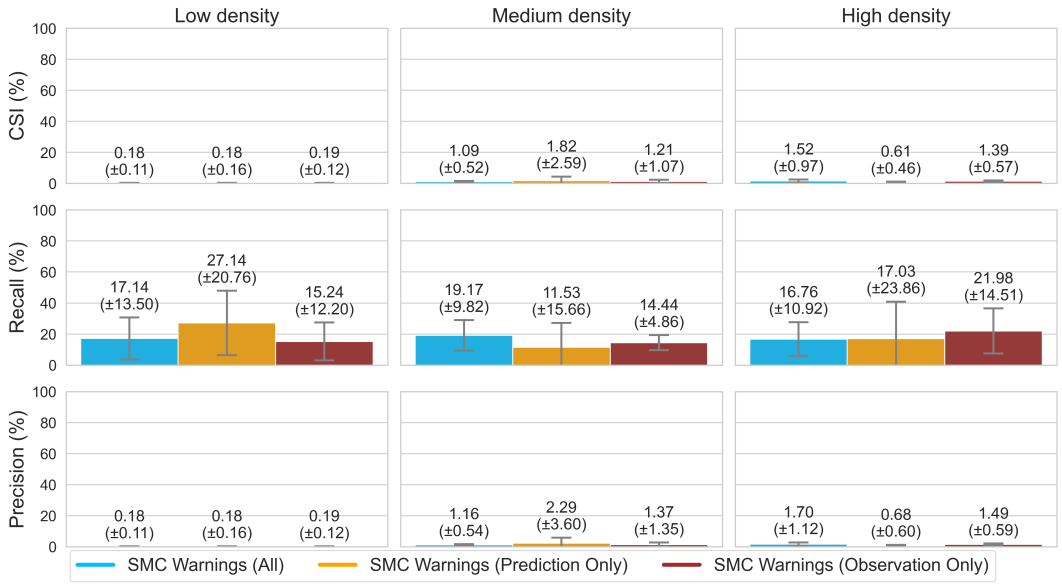

Figure 1: Performance comparison between considering the different types of SMC warnings as baseline.

Major comment 3. Justify the target thresholds and discuss their implications for comparing population-density groups:
In Sect. 4.1 and Table 1, the positive class is defined using different thresholds: 1 call per hour for low- and medium-density municipalities, and 3 calls per hour for high-density municipalities. While this may be reasonable because call volumes differ greatly across population-density groups, it also means that the three models are not predicting exactly the same level of impact. The high-density model is evaluated on a more severe impact definition than the other two models. This complicates interpretation of Table 2 and Fig. 8, where performance is compared across low-, medium-, and high-density groups. The authors should explain how these thresholds were selected, whether alternative thresholds were tested, and how sensitive the conclusions are to this modelling choice. It would also be useful to discuss whether the threshold should represent a fixed operational burden, a population-normalised impact rate, or a severity-based definition.

Response:

We thank the reviewer for this comment. The definition of different thresholds tries to adjust for major demographic differences and provide a more homogeneous impact estimator, since higher populations naturally generate a higher volume of emergency calls. This observation emerged during our data analysis, but was also explicitly raised by civil protection personnel during various meetings. Ultimately, we argue that adjusting these thresholds precisely makes the results more comparable between population groups. The specific impact thresholds were determined through experimentation. For the medium and low density groups, we maintained a threshold of 1 call per hour, as the total volume of calls in these areas was considerably lower than that of high density municipalities, and a higher threshold would have reduced the number of impacts significantly. For high-density areas, however, we found that using thresholds varying from 1 to 5 calls resulted in similar performance. Ultimately, we decided on 3 calls because the resulting positive class rate was comparable to the medium-density group. We have added an explanation of this decision in Section 4.1:

> "We use a threshold of 1 call per hour for the medium- and low-density groups, as selecting a higher threshold would excessively reduce the already limited number of impacts. For high-density municipalities, we experimented with different thresholds and found that values varying between 1 and 5 yielded similar results. Ultimately, we opted for 3 calls, as it resulted in a number of impacts comparable to that of the medium-density group."

Ultimately, these impact thresholds are parameters that must be adapted to the requirements of the operational tasks, but also to the constraints of the data itself. It is possible that other impact indicators could require approaches different from using fixed thresholds.

Major comment 4. Strengthen the evaluation by addressing event dependence and operational usefulness:
Sect. 5.3 describes a walk-forward evaluation with validation and test sets constructed to contain the same number of positive samples, and Fig. 6 illustrates this design. This is a thoughtful approach to temporal imbalance, but the evaluation is still based on municipality-hour samples. Flood impacts are likely clustered in time and space, so many positive samples may belong to the same rainfall event. As a result, the effective number of independent test cases may be smaller than suggested by the sample count. I recommend adding an event-based evaluation, for example by grouping impacts into rainfall or flood episodes and reporting whether the model detects event onset, peak impact periods, and affected areas. In addition, the practical alert burden should be reported, such as the number of alerts generated per event, per municipality, or per day. Metrics such as CSI, recall, and precision are useful, but they do not fully show how emergency managers would experience the model in operation.

Response:
We thank the reviewer for this insightful comment. We acknowledge that the temporal and spatial clustering of flood impacts is indeed an important characteristic of this data, and we agree that evaluating from an event-based perspective provides valuable operational information.
However, implementing an event-based evaluation falls outside the scope of the current study. This is because the episodes we analyse cover a wide range of typologies, ranging from large-scale, prolonged events to highly localised convective storms. Defining discrete, independent events across such a diverse spatio-temporal scale introduces significant subjectivity and complexity. We believe this challenge would be better addressed in future work dedicated specifically to the real-time operational deployment of our (or other) algorithm(s). Regarding the practical alert burden, we completely agree that translating these hourly predictions into user-oriented warnings is critical for emergency managers. Nevertheless, developing and evaluating such a warning system falls beyond the objectives of this study.
To ensure that this operational perspective is reflected in the manuscript, we have added a new paragraph to Section 7.1 discussing the opportunity of future work focused on operational applicability,

which should include event-based evaluations and case studies:

> "In addition to refining the model's resolution, the operational evaluation of these systems must be expanded. While the municipality-hour evaluation framework employed in this study provides statistically robust metrics for comparing model performance, it does not fully translate into how emergency managers would perceive the activations (e.g., the number of distinct alerts generated per event). Future work focusing on operational deployment should address this limitation in detail, integrating strategies such as event-based evaluation and adequate training protocols to better reflect the real-world management of impactful events."

While we acknowledge that the municipality-hour approach can introduce noise by generating multiple predictions for zones already warned, it allows us to compute statistically robust metrics across a highly imbalanced dataset. Since our main focus is assessing the prediction capabilities of the models for general emergency management, the sample-level evaluations are adequate for capturing both the start and continuation of impacts, hour by hour, as well as comparing them to operational baselines. We sincerely hope this justification satisfies the reviewer's concerns.

Minor and technical comments:
The manuscript would benefit from careful language editing. Examples include inconsistent numerical formatting such as "32.000 km2", minor grammatical issues, and some awkward or redundant phrasing. I also suggest replacing "112, the European equivalent of 911" with "112, the European emergency number." The description of the daily rain filter should also be clearer, especially regarding whether it uses information from the whole day retrospectively or whether it could be implemented in real time.

Response:
We have revised the manuscript and corrected several typos, numerical formatting inconsistencies, redundant phrases, and unclear terms in general, including those suggested in this comment.
Regarding the definition of the daily rain filter, we have made a slight modification when introducing it ("To mitigate this, we apply a **retrospective** daily filter to include in the dataset only days with significant **observed** rainfall.") to explicitly mention that it is a retrospective approach. To address the considerations when implementing our models in real time, we have added a paragraph at the end of Section 4.4 that directly targets the operational use of the models:

> "Regarding the operational validity of applying this filter, the models developed in this study could still be deployed every hour. Even after removing a significant number of days without rainfall, samples where neither rainfall nor impacts occurred still constitute the majority of the dataset, allowing models to train on situations without risk. The objective of the filter is just to mitigate class imbalance by focusing the problem towards days with actual impact potential."
