June 22, 2026

Dear Reviewer #1,

We would like to thank you for your time and effort, and for your feedback on our initial manuscript.

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 #1

General comments:
This manuscript presents a machine-learning framework for predicting flood-related 112 emergency calls at municipal and hourly resolution in Catalonia. The topic is timely and relevant to Natural Hazards and Earth System Sciences, especially in the context of impact-based forecasting and operational early-warning systems. The use of emergency-call data as high-resolution impact proxies is promising, and the comparison with operational warning-based baselines gives the study practical relevance.
The manuscript has a potentially publishable core, but I recommend major revision. The main issues concern the operational validity of the prediction framework, the robustness and comparability of the evaluation design, and the clarity and reliability of the interpretation. These issues should be addressed before the conclusions about machine-learning-based impact anticipation can be fully supported.

Response:
We sincerely thank Reviewer #1 for their positive assessment, the constructive comments, and for recognizing the relevance of our study to the journal. We believe the revised manuscript has greatly benefited from the discussions about operational use and evaluation validity, and we hope the responses presented below satisfactorily address the reviewer's concerns."

Comment 1. Operational validity and target definition:
The manuscript is framed as an impact-based early-warning study, but some modelling choices make the prediction task closer to detecting or updating already emerging impacts. In particular, the model uses flood-related emergency calls from the previous 1–3 hours as predictors, which may strongly benefit cases where impacts have already started. The authors should distinguish more clearly between predicting the first occurrence of impacts and predicting the continuation of ongoing impacts. Additional experiments without previous-call features, or separate results for "before first call" and "after first call" situations, would make the operational contribution clearer. The daily rain filter also requires clarification: if it uses full-day information retrospectively, the evaluation may not represent a real-time forecasting setting. The definition of the positive class, positive rate, class weighing, and evaluation metrics should also be stated more explicitly.

Response:
We thank the reviewer for raising these critical points about the operational validity of our approach. 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 some ablation experiments in Section 6.2.3, 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 occurrence of impacts 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 "before the first call" 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..."

Regarding the real-time use of our methods, the retrospective daily rain filter was primarily employed to accelerate training and mitigate extreme class imbalance in our experiments, which at the same time allowed us to provide clearer performance metrics by focusing in relevant periods rather than in thousands of uninformative hours. However, in an operational, real-time setting, the model can still be deployed hourly without issue, as even with the filter applied, most of the dataset still consists of samples that did not register neither rain nor impacts. Another option could be to only run the model operationally if rain is forecasted exceeding a specific probability, although this particular decision falls outside of the scope of the study. We thank the reviewer for this comment, as we overlooked how this filtering could cause confusion about operational use, and have added some insights from this discussion to Section 4.4:

> "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."

Lastly, we agree that when presenting concepts such as positive rate or class weighting, we did not explicitly state that these were obtained on the filtered dataset. We recognise that these omissions hindered transparency, and we have corrected it with the following changes:

- Table 1 shows target-related statistics from the dataset after applying the daily rain filter. We now explicitly mention this both in the table's caption ("The presented statistic were obtained from the dataset after applying the daily rain filter detailed in Sect. 4.4.") and when introducing the table in the manuscript ("The values presented here reflect the data after applying the daily rainfall filter presented in Sect. 4.4, which significantly reduced the number of negative samples.")

- Clarification of the positive and negative classes: "Instances meeting these respective thresholds constitute the positive class, while all other samples form the negative class, which signifies no-impact."
- Definition of the positive rate: "...as indicated by the low positive rate (i.e., the fraction of total samples belonging to the positive class)"
- Clarification of class weighting methodology: "To mitigate this issue, class weighting was applied to the loss function during model training, imposing a higher penalty for misclassifying the impacts (which are the minority class). Specifically, positive instances were assigned a weight inversely proportional to their overall frequency within the training set."
- We now define the recall and precision metrics more explicitly: "Additionally, we provide the recall (the fraction of total observed positive cases that were correctly predicted; Eq. (3)) and the precision (the fraction of positive predictions that corresponded to actual impacts; Eq. (4)) for a more comprehensive assessment."

Comment 2. Evaluation design and comparability across temporal configurations:
The walk-forward evaluation is reasonable in principle, but the current implementation makes comparisons across temporal configurations difficult to interpret. Although the authors control the number of positive samples, the validation and test periods differ across configurations. Thus, differences in performance may arise not only from model robustness or training history, but also from differences in the test data themselves, such as negative-sample counts, positive rates, rainfall regimes, event severity, seasonality, or spatial distribution of impacts. Because validation sets also differ, hyperparameter tuning and threshold selection are performed under different validation distributions. The results shown in Figure 7 also do not indicate a clear trend.

Response:
We acknowledge the reviewer's concerns regarding the temporal configurations employed. The irregular and sparse nature of impacts poses many challenges when evaluating these systems, and when choosing a particular temporal cross-validation scheme, one must also carefully consider their related trade-offs. In this regard, we experimented with multiple alternatives before deciding on our approach. For instance, fixing temporal periods (e.g., specific months) resulted in highly inconsistent numbers of both positive and negative samples per set, leading to high variance in performance metrics. On the other hand, fixing the total number of samples (positive and negative) led to sets with drastically different positive-to-negative ratios, and varying temporal periods. Moreover, for both previous cases, we could not ensure that the validation and test sets would contain any positive samples, which was especially common in medium- and low-density municipalities.
In the end, we chose to fix the number of positive instances (impacts) per set. When comparing the different schemes, this configuration yielded the most stable results with the lowest standard deviation in model metrics, providing a more reliable setting for comparing performance and interpretability across configurations. While we agree with the reviewer's assessment that this leads to different test periods and background distributions, we have found that stabilizing the positive class is the most effective way to derive meaningful insights from such an imbalanced dataset. To justify this important decision in our modeling strategy, we have added a paragraph at the end of Section 5.3:

> "While we acknowledge that this can lead to different test periods, background distributions, or varying numbers of negative samples, after extensive experimentation with multiple approaches, we have found that stabilizing the positive class was the most effective way to derive meaningful insights from such an imbalanced dataset. Other strategies like fixing temporal periods (e.g., specific months) resulted in highly inconsistent counts

of both positive and negative samples per set, leading to high variance in performance metrics. Conversely, fixing the total number of samples led to sets with drastically different positive-to-negative ratios, and varying temporal periods. Furthermore, under both previous approaches, we could not ensure that the validation and test sets contained any positive samples, an issue especially common in medium- and low-density municipalities."

Additionally, we have added a note to address the lack of clear trends in Figure 7:

> "**Although no clear trend is observed**, we note how the standard deviation intervals of individual configurations..."

Comment 3. Interpretation, uncertainty, and presentation:
The SHAP, ablation, and rainfall-stage analyses are useful, but their interpretation should be more cautious. The low-density model has very weak and unstable performance, so feature-importance or sensitivity analyses for this group should not be used to support strong conclusions; similar caution is needed for the medium-density model. The authors should distinguish robust findings from the high-density model from exploratory findings in lower density settings. In addition, SHAP and ablation results should not be interpreted causally, especially given the likely correlation among rainfall, radar, warning, and call-history predictors.
The rainfall-stage analysis is interesting, but the claimed "warm-up" phase of the first hours needs clearer reasons. More generally, the manuscript would benefit from restructuring: methodological details in the Introduction should be reduced, the short Related Work section could be integrated or better aligned with the research gap, and several redundant or unclear sentences should be revised.

Response:
We appreciate these insightful notes regarding the interpretation of our explainability analyses and the overall structure of the manuscript. We fully agree that the performance of the medium- and, especially, low-density models are less stable. Consequently, we have revised the text to clearly distinguish between the robust findings from the high-density model and the exploratory insights from the lower-density approaches, with the goal to temper our conclusions and avoiding over-interpretation. Moreover, we will clarify that the interpretation of the SHAP and ablation results represents feature contributions within the model, and not necessarily causal relationships. Specific revisions include:

- Slight revision regarding SHAP feature importance: "The low-density model **seems to rely mostly** on weather features."
- Careful language regarding feature groups: "On their own, the number of calls in the past hours has the best prediction capabilities of all feature groups (when considering the mean CSI)."
- Correction in the ablation study: We replaced "For the low-density group, **the best model results when static indicators are ignored, with a good balance between recall and precision. In any case, standard deviations make it hard to confidently make assumptions.**" with "For the low-density group, **the best model results are obtained when static indicators are ignored. However, standard deviations make it hard to confidently draw conclusions for these two population groups.**"
- Tempered the conclusions on interpretability: "Pairing this with explainable AI techniques like SHAP allows for the interpretability of model outputs. **In low- and medium-density areas, rainfall-related features appear to have a greater average contribution to the predictions than other variables, largely ignoring static indicators about population. In contrast, the high-density model achieve the most accurate predictions by modulating information on emergency calls in previous hours with weather data**

and population counts. These findings highlight how interpretability can vary significantly by demographic region. However, since the reliability of SHAP values depends on the model's prediction capabilities, the interpretations from medium- and specially the low-density models must be understood as exploratory insights rather than robust conclusions."

Regarding the rainfall-stage analysis, we have expanded our discussion on the "warm-up" phase to provide a more grounded rationale for this behavior:

> "This phenomenon mimics a "warm-up" effect. Because the model relies heavily on calls received in recent hours, it experiences a "temporal lag" at the start of an event. It is not until it starts receiving the first impacts (or that the event consolidates and the hazard becomes more apparent), that the model has enough context to make predictions more effectively."

Additionally, we agree that the introduction contain some technical details that are not relevant at that point in the manuscript. We have modified the following parts:

- Removed specific details about Catalonia:
  - "$32,000$ km^2, 8.01 million inhabitants (Statistical Institute of Catalonia, 2024),"
  - "receiving $\sim$700 mm of annual rainfall with irregular interannual distribution"
- Removed the footnote specifying the data gap in our dataset ("Excluding the period between 14 March to 31 December 2020, due to a lack of data.")
- Replaced "We address this problem **by dividing municipalities into three groups depending on their population density and training individual models for each one. This allows** for a better assessment of performance" with "We address this problem by **stratifying models based on population densities, allowing** for a better assessment of performance"

Lastly, the related work section (Section 2) has been significantly revised to better align with the research gap targeted in the paper. We added several new references and focused the discussion towards the specific goals of this study.

Technical correction 1.
Replace "112 (the European equivalent of 911)" with "112, the European emergency number."

Response:
Corrected as suggested.

Technical correction 2.
Remove redundancies such as "expanded to the pan-European level ... to apply it at a pan-European scale," and "Static data refers to data that does not vary over time, and instead is used to represent population dynamics through static indicators."

Response:
We have revised the manuscript and corrected several redundant or awkward phrases, including the ones suggested.

Technical correction 3.
Correct typos and grammar issues, including "provnen," "7 input feature," "maximas,", "we are able of covering" , "Also is important to note", and "To successfully model impact, it is essential to provide the model with relevant and diverse information from which it can create relationships between input data and the target impact data."

Response:
We have revised the paper, and corrected several typos, incorrect terms and inconsistent verb tenses.

Technical correction 4.
Use consistent numerical formatting, e.g., "32,000 km^2" rather than "32.000 km^2."

Response: Indeed, the paper presented a few numerical formatting inconsistencies, as highlighted by the reviewer. We have revised the text and corrected them accordingly.
