July 8, 2026

Dear Editor and Reviewers,

We would like to thank the reviewers and the editorial team for their time and for considering our work.

Please find below our point-by-point response to both reviewers' 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 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 and for the constructive comments. 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 line 422 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

a 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 (lines 483-488):

> "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 (lines 289-292):

> "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 (lines 181-182: "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 (lines 178-179): "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 (lines 187-188): "...as indicated by the low positive rate (i.e., the fraction of total samples belonging to the positive class)"
- Clarification of class weighting methodology (lines 350-353): "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 (315-317): "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 (lines 341-348):

> "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 (line 365):

> "**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 (lines 388-389): "The low-density model **seems to rely mostly** on weather features."
- Careful language regarding feature groups (lines 408-409): "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 (lines 430-431): 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 (lines 488-494): "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 (lines 443-445):

> "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 (lines 53-54) "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"

Similarly, we have revised the abstract to contain fewer technical details. In particular, we have merged two previous paragraphs into a single one at the end of the abstract, summarizing the explainability experiments conducted and their main results (lines 17–24).

> "Further experiments focus on explainability to improve trust by understanding the models' behaviours using feature importance analyses like SHAP (SHapley Additive exPlanations), feature group selection, and ablation. Additionally, we present a methodology to evaluate model behaviour across different stages of a rainfall event. These experiments show how ML models can effectively combine various data sources to provide more effective predictions than using these sources in isolation. Moreover, while the best-performing model exhibits moderate performance at the onset of rain, it still consistently improves upon traditional baselines. As the event evolves, its prediction skill rises significantly and is maintained even after precipitation has stopped. This highlights the strong potential of even relatively simple ML pipelines to deliver timely, localised anticipation of weather-related impacts."

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.

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 (lines 483-488):

> "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 updated the results from Figure 8 and changed the baseline definition by removing the part that mentions that observation warnings were ignored (in Section 5.2, we previously said that "The only difference is that for SMC Warnings, we ignore Observation warnings, as these are reactive alerts issued after a phenomenon has already been observed")
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.

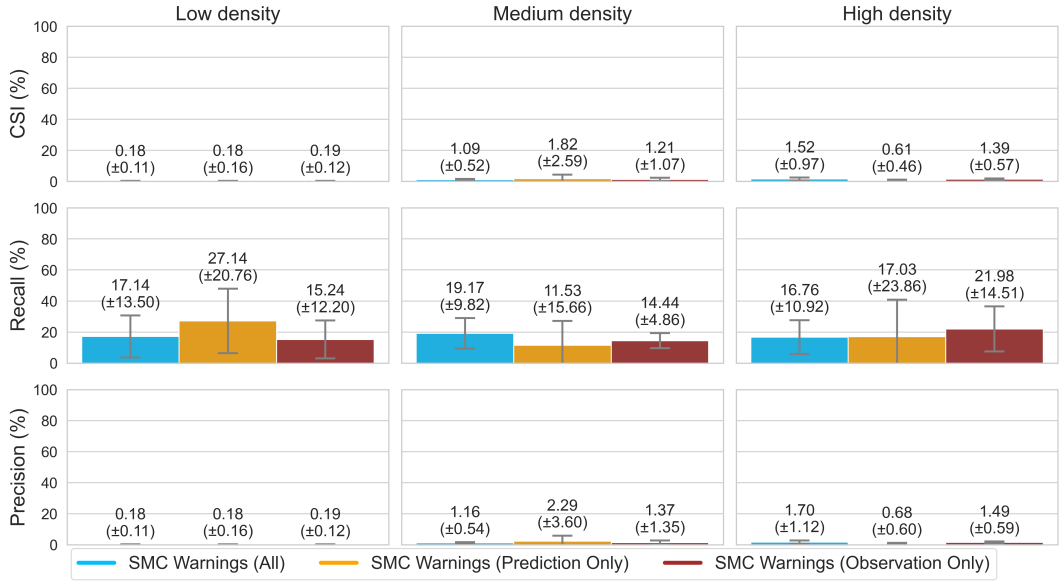

Figure 1: Performance comparison between considering the different types of SMC warnings as baseline.

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 (lines 174-178):

> "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 (lines 517-522):

> "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 (lines 289-292):

> "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."