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.

Specific comments

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.

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.

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.

Technical corrections

1. Replace “112 (the European equivalent of 911)” with “112, the European emergency number.”
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.”
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.”
4. Use consistent numerical formatting, e.g., “32,000 km²” rather than “32.000 km².”