Review of Manuscript Egusphere-2026-1275

"Hydrologically Enhanced Machine Learning Framework for Urban Flood Inundation Mapping Using Multi-Sensor Remote Sensing Data: A Case Study of Mumbai, India"

General Comments

This is an interesting case study applying hydrologic-topographic inputs to a data-driven ML model for flood inundation mapping. However, the reviewer is not entirely convinced of the novelty of the study, since, contrary to the authors' claims, similar hydrologic-topographic conditioning approaches have already been explored in recent literature. While the overall methodological framework is mostly sound, the apparent lack of independent validation (or even a sensitivity analysis) in the flood identification process makes it difficult to accept the reported results at face value. In addition, although the study is framed around the 2019 monsoon event, the authors do not discuss whether the flooding was driven predominantly by a particular rainfall-runoff process (e.g., fluvial or pluvial) or whether coastal processes (e.g., storm surge, tidal backwater) also contributed, despite the study area's direct exposure to the coast; compound flooding is not considered anywhere in the manuscript. Given these gaps, the reviewer recommends a substantial revision addressing validation, sensitivity analysis, and the treatment of coastal/compound flood processes before this study can be considered for publication.

Specific Comments

- **Lines 95–96 and Table 1:** The authors state that hydrodynamic models are data- and computation-intensive, reducing their scalability to operational scale. This is not entirely accurate given recent advances in hydrodynamic modeling through optimization and modularity (e.g., Neal et al., 2018; Zhao et al., 2021; Garcia et al., 2023) and the availability of simplified-physics models (e.g., LISFLOOD-FP, SFINCS) that are computationally efficient enough for operational or near-operational use. This claim should be revised or qualified.
- **Lines 105–109:** The authors should revise statements claiming a lack of recent research on physics-based (e.g., hydrographic-topographic) conditioning within data-driven/machine learning flood studies, since such work already exists (e.g., Lee et al., 2024; Yang et al., 2024). This bears directly on the claimed novelty of the study.
- **Lines 190–197:** How is the proposed Relative Elevation Model (REM) different from, or superior to, the existing Height Above Nearest Drainage (HAND) method? The formulation described (elevation minus nearest river elevation) is conceptually very close to HAND, and the manuscript should clarify the distinction and justify why a new term/predictor is warranted rather than adopting the established HAND terminology.
- **Lines 200–208:** Does the River Network Index (RNI) apply only to major identified river channels and their main tributaries? What about minor tributaries, man-made channels (e.g., ditches, canals, storm drains), and other water bodies (e.g., lakes, ponds, estuaries) that are common across the Mumbai Metropolitan Region and can materially affect local flooding? In addition, the RNI equation as given (precipitation divided by elevation above the regional minimum) does not appear to include an explicit distance-to-river term, despite the text describing RNI as being computed "via a Euclidean distance transform" as "a measure of hydraulic proximity and connectivity." Please reconcile the equation with the stated definition and clarify how proximity to the drainage network is actually incorporated.

- **Lines 215–227:** No actual validation appears to have been conducted for the flood identification/labeling method — i.e., identified flooded pixels do not appear to be compared against independent, known/observed flood records (e.g., surveyed high-water marks, stream gauge records, official disaster management flood reports, or news-reported waterlogging locations). Given that the consensus labels are themselves used as ground truth for training and evaluating the ML models, this is a significant gap. The reviewer also recommends a sensitivity analysis on the listed flood identification criteria/thresholds (SAR backscatter ratio/difference, NDWI, and REM) to demonstrate that the reported accuracy and AUC are not artifacts of the specific threshold choices.
- **Figure 4 / Lines 320–323:** The text states that hydrologic-topographic variables, "especially REM and RNI," are primarily responsible for the model's discrimination gains. However, the feature importance plot shows RNI ranked sixth, below several SAR backscatter statistics (S1_VV_std, S1_VV, S1_VH) and slope, with only REM clearly standing out as the top feature. The claim regarding RNI's importance should be tempered or the apparent contradiction with Figure 4 explained.
- **Table 4 (lines 306–309):** The ablation baseline ("Without REM & RNI," accuracy 0.69, AUC 0.77) is not attributed to a specific classifier. Given that Table 3 reports per-model metrics for Random Forest, XGBoost, and the ensemble separately, please clarify which classifier this ablation baseline corresponds to (presumably XGBoost, per the surrounding text), and consider reporting ablation results for all three models for consistency.
- **Figure 5 (around line 362):** How is flood risk (low/medium/high, per the risk zone map) classified? The reviewer assumes it is derived by thresholding the flood probability map, but this is not stated. If so, the thresholds used and the classification basis should be added to the figure caption or methods section.
- **General – coastal/compound flooding:** The manuscript does not discuss whether the 2019 monsoon flooding in the study area was driven primarily by pluvial, fluvial, or coastal (e.g., storm surge, tidal backwater) processes, despite the Mumbai Metropolitan Region's direct coastal exposure. Compound flooding is not addressed. Given that REM and RNI are framed as capturing hydraulic connectivity, the omission of tidal/coastal influence on drainage backwater effects is a notable gap that should be discussed, or explicitly acknowledged as a limitation.
- **Lines 382–386 and Conclusion (Section 6):** The authors acknowledge the single-event, consensus-label-only nature of the validation as a limitation, but this is the same fundamental concern raised above (Lines 215–227). Given this limitation, the operational-readiness language used elsewhere in the manuscript (e.g., abstract, conclusion) should be tempered until independent validation is demonstrated, even partially, using available observational records.