General Comment

The paper presents a global, sub-national impact-based forecasting (IBF) framework for tropical cyclones using open data and a two-stage XGBoost architecture. While the comparison against historical and windspeed-exposed baselines shows that the model provides balanced performance and can be applied to operational applications, with four case studies analysed, several methodological, structural, and reporting concerns limit the paper's contribution. It's recommended a major revision before the paper is accepted.

Major Comments

1. It would be better to combine the Introduction and Related Work sections. In the Introduction, many statements lack citations, whereas in the Related Work section, some citations could fill these gaps. See the comments on citations, mainly in the Introduction and Related Work sections, below. In addition, some content is duplicated across both sections, such as lines 45–46 and lines 68–69.

   a. There is an incorrect citation on line 16: the cited paper is mainly about anticipatory action rather than cyclone damage. The authors should cite more appropriate papers.

   b. For line 18, the following should be cited, as Parks et al. (2023) cited this paper for a sentence similar to line 18: R. M. Adams, C. M. Evans, M. C. Mathews, A. Wolkin, L. Peek, "Mortality from forces of nature among older adults by race/ethnicity and gender," *J. Appl. Gerontol.* 40, 1517–1526 (2021).

   c. From lines 23–29, add citations for the following claims: "AA can reduce losses and improve preparedness," "limiting both their effectiveness and comparability," "Impact-based forecasting provides a pathway to address these limitations by linking hazard information to expected consequences," and "challenges that motivate the need for a consistent global baseline."

   d. On line 32, "many existing models" should be supported with examples and citations. e. On line 57, add citations for "which impose rigid assumptions and limit transferability."

2. The WorldPop layer uses only the year 2020 for model training and prediction; however, the target data from EM-DAT span 2000–2022. The differences in population distribution over these 23 years are neither considered in the study nor discussed in the limitations.

3. No hyperparameter search is reported for the XGBoost model itself. The parameters (max_depth=4, learning_rate=0.01, n_estimators=100) are fixed without justification or tuning. This is a significant omission given the paper's claim of optimizing model performance. Although the authors claim that the

XGBoost model outperforms the wind-exposed model in impact detection, its precision is 0.3 while the wind-exposed model achieves 0.57. Regarding recall, the historical baseline achieves 0.95 while XGBoost achieves 0.68. One could argue that combining the historical baseline and the wind-exposed model would allow both the detection of where impacts occur and the identification of severe impacts.

Specific Comments

1. **Abstract** a. The abstract adequately summarizes the work, but the claim of being a "first-generation global benchmark" is strong and not sufficiently justified against existing global IBF systems (e.g., CLIMADA-based frameworks already operate globally).

   b. Briefly describe the improvement of the approach in line 8, for example, by providing at least one quantitative metric illustrating the improvement.

   c. In line 12, briefly describe the limitations.

   d. Provide a clear description of the decision-relevant metrics and operational thresholds.

   e. It is not clear why the machine learning approach reveals a fundamental trade-off between coverage and conservative severity detection. Add one sentence to explain this.

2. **Data Description and Standardization**

   a. Lines 90–94 should be in the Introduction, not in the Methods section.

   b. From lines 96 to 98, spell out the full names of these abbreviations and add relevant citations the first time they appear.

   c. On line 155, it is unclear which version of the WorldPop data was used in the study; furthermore, the GitHub repository is not available in the Code Availability section for verification.

3. **Model Training and Evaluation**

   a. On line 230, which administrative unit is referred to—Admin 1 (province/state) or Admin 2 (city)? The difference in geographic resolution can influence the absolute number of affected people.

   b. The confusion matrices (Figure 5) show that the two-stage XGBoost produces 1,238 false high-impact predictions versus 309 true high-impact detections. This false alarm rate is operationally problematic and is not adequately discussed.

4. **Results** The geographic sensitivity analysis (Table C1) shows that the Australian Region has F1=0.00 at the 15% threshold under both the global and geo-constrained specifications. This means the model completely fails for an entire cyclone basin. Furthermore, line 460 claims that the global model "captures patterns that generalize across basins," but Table C1 shows highly variable performance (F1 at 15% ranging from 0.00 to 0.38). This variability undermines the generalization claim. This limitation and the spatial differences across the globe deserve discussion and emphasis in Section 5.
5. **Discussion** The discussion of hybrid strategies is conceptually sound but vague. No specific protocol for combining machine learning and exposure-based approaches is proposed or tested.
6. **Conclusion** The conclusions are reasonable but overly optimistic given the results. Phrases such as "demonstrates the feasibility of transferable, sub-national predictions" gloss over the fact that the model fails entirely in some basins and achieves F1<0.30 for high-impact detection globally.
7. **Appendix**

   a. Appendix E (SHAP analysis) reveals that prev_events_5years behaves counterintuitively in Stage 1 (higher values → lower predicted impact probability). This is not adequately explained. If regions with more historical events are predicted to have a lower impact probability, this suggests the model may be learning a spurious inverse relationship, possibly due to class imbalance or collinearity.

   b. Appendix F lists 26 forecast events but provides no statistics on how representative these events are of the broader 780-event dataset in terms of intensity, basin, or impact level.

Minor Comments

1. Line 6: for XGBoost, spell out the full name first, then give the abbreviation.
2. For the citations, it is better to use parentheses, as in line 58: "between hazard, exposure, and vulnerability from historical data (Mandal et al., 2022)." Please check the other citations as well.
3. Figure 1: the caption is missing a word, such as "prediction" or "simulation," before "pipeline."
4. Line 162: EM-DAT does not provide cyclone tracks; "events" may be a better term.
5. Lines 409–410: "we compute SHAP values for both stages (Fig. E1 in the Appendix." → Missing closing parenthesis.

6. Lines 453–454: "specifications.At the 0% threshold" → Missing space after the period.
7. Line 346: "We consider the detection of any impact (0%) and high impact (15%), and explore the error structure magnitude-based and by impact category." This sentence is unclear and grammatically incomplete.
8. Line 733: "root-mean absolute error (RMSE)" → RMSE is root-mean-squared error.