https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Global Sub-national Impact-based Forecasting for Tropical Cyclones Using Open Data: Combining Machine Learning and Exposure-based Approaches

Federico Moss[1], Yelena Mejova[1], Andreas Kaltenbrunner[4,5,1], Tristan Downing[3], Marc van den Homberg[2], Pauline Ndirangu[3], Leonardo Milano[3], and Kyriaki Kalimeri[1,6]

[1]ISI Foundation, Turin, Italy
[2]510, an Initiative of the Netherlands Red Cross, The Hague, Netherlands
[3]UN OCHA Centre for Humanitarian Data, The Hague, Netherlands
[4]Department of Engineering, Universitat Pompeu Fabra, Barcelona, Spain
[5]Internet Interdisciplinary Institute, Universitat Oberta de Catalunya, Barcelona, Spain
[6]UNICEF, New York, USA

Correspondence: Kyriaki Kalimeri (kyriaki.kalimeri@isi.it)

Abstract. Tropical cyclones (TCs) cause substantial and uneven impacts across regions, driven by differences in exposure and vulnerability. While anticipatory action (AA) systems aim to mitigate these impacts, they are typically based on hazard thresholds rather than predicted consequences, limiting their effectiveness and consistency. Impact-based forecasting offers a promising alternative, but existing approaches are often region-specific or rely on non-transferable data. In this study, we develop a global, sub-national impact-based forecasting framework that predicts affected-population fractions using only openly available data. The model integrates hazard, exposure, and contextual features within a two-stage XGBoost architecture and is evaluated across 780 historical TC events using decision-relevant metrics aligned with operational thresholds. Our results show that machine learning improves the detection and spatial localization of impacts, but does not outperform simpler exposure-based approaches in identifying severe events. This reveals a fundamental trade-off between coverage and conservative severity detection, suggesting that hybrid strategies combining both approaches are better suited for operational use. We position this system as a first-generation global benchmark for impact-based forecasting: it demonstrates the feasibility of transferable, sub-national predictions using open data, while clarifying the limitations that must be addressed for reliable deployment in anticipatory action systems.

1 Introduction

Tropical cyclones (TCs) represent one of the most devastating natural hazards globally, causing substantial economic losses and widespread infrastructure damage (Chaves-Gonzalez et al., 2022). However, their impacts are not distributed evenly. Post-disaster outcomes vary significantly across regions, driven by disparities in access to transportation, financial resources, education, employment opportunities, and timely warnings (Parks et al., 2023). These inequalities shape both exposure and vulnerability, amplifying the consequences of extreme events in already disadvantaged contexts. At the same time, climate

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

change is altering tropical cyclone dynamics, increasing their intensity and shifting their occurrence toward regions that were previously less exposed (Jing et al., 2024), thereby expanding the geography of risk.

Anticipatory action (AA) offers a mechanism to mitigate these impacts before disasters fully unfold. By triggering early interventions based on forecasts, AA can reduce losses and improve preparedness. However, current AA systems are typically based on hazard thresholds (e.g., wind speed or rainfall) rather than expected impacts (see Food and Agriculture Organization of the United Nations (2023) and United Nations Office for the Coordination of Humanitarian Affairs (2023)). These triggers are often defined in an ad hoc manner and lack consistency across countries and contexts, limiting both their effectiveness and comparability. Impact-based forecasting provides a pathway to address these limitations by linking hazard information to expected consequences. Realising this potential operationally requires improvements in impact data, hazard representation, and predictive modeling—challenges that motivate the need for a consistent global baseline.

Previous work has identified key drivers of tropical cyclone impacts, including storm intensity, rainfall, population exposure, infrastructure resilience, and governance conditions (Peduzzi et al., 2012; Cardona et al., 2014; Kim et al., 2019). However, many existing models either focus on single hazard dimensions or incorporate socio-economic vulnerability in a limited or region-specific manner, reducing their applicability across diverse settings. Moreover, predictive models are often developed for individual countries using locally curated data, which constrains their transferability (Kooshki Forooshani et al., 2024). While early global-scale studies have demonstrated the feasibility of mapping exposure to cyclone hazards (e.g., Fang et al. (2014)), there remains a gap between global coverage and sub-national predictive accuracy. Addressing this gap is increasingly important as climate change expands the spatial footprint of tropical cyclone risk (Arachchige et al., 2025).

In response, we develop a global impact-based forecasting framework that predicts affected-population fractions at sub-national (grid) resolution using only openly available data. The model integrates hazard information (track-derived winds and rainfall), topography, coastal exposure, settlement structure, and population distribution into a two-stage XGBoost architecture. By design, the framework prioritizes global coverage and reproducibility, providing a consistent baseline for evaluating impact-based forecasting approaches across regions. We report model performance together with its limitations to assess its suitability for operational use. In doing so, we explicitly test the extent to which globally consistent, open data can support transferable sub-national impact prediction, and where such models fall short relative to simpler operational baselines.

Our main finding is that while machine learning improves early detection and spatial localization of impacts, it does not outperform simpler exposure-based approaches in identifying severe impacts. This distinction is important for operational use, as it implies that different modeling approaches are better suited to different decision tasks. This highlights the value of hybrid strategies that combine both approaches for operational decision-making. Benchmarked against wind-based triggers across 780 historical events, the model improves detection and localization of impacted regions while remaining interpretable through decision-aligned thresholds. We therefore position this system as a first-generation global benchmark that both demonstrates the feasibility of transferable, sub-national impact prediction using open data and clarifies the limits of such approaches for operational decision-making.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

2 Related work

Impact-based forecasting (IBF) for tropical cyclones has developed along two main directions: (i) *open, modular* pipelines that derive hazard and exposure from forecast tracks using transparent assumptions, and (ii) *trained* models that infer impacts directly from gridded predictors. Earlier approaches relied on expert-weighted indices and parametric vulnerability curves, which impose rigid assumptions and limit transferability. Recent work moves toward data-driven IBF, learning relationships between hazard, exposure, and vulnerability from historical data Mandal et al. (2022).

A widely used modular framework is CLIMADA Aznar-Siguan and Bresch (2019); Bresch and Aznar-Siguan (2021), which supports ensemble-based impact estimation and uncertainty quantification. Building on this, Kam et al. (2024) develop a global IBF model for tropical-cyclone-induced displacement, combining forecast-derived hazards with parametric impact functions. Their results highlight how predictive uncertainty shifts from meteorological drivers at longer lead times to vulnerability-related factors near landfall. In contrast, our approach models affected-population fractions directly using a trained XGBoost model on open gridded predictors, without relying on parametric impact functions.

Recent studies also emphasize the role of trigger design in operational performance. Sedhain et al. (2025) compare a statistical IBF model with a wind–damage curve and show that simple exposure-based rules can outperform trained models under strict thresholds. Our work complements this by evaluating performance across thresholds and metrics aligned with operational decisions. Consistent with their findings, we observe that exposure-based rules remain strong indicators of high-impact events, while machine learning improves detection and spatial localization, suggesting a complementary role for both approaches.

Transferability remains a central challenge for IBF. Wagenaar et al. (2021) address sample-selection bias using distribution-matching techniques, improving model performance across regions. Here, we take a different approach by training a single global model on openly available predictors and evaluating its performance across multiple basins and countries.

A parallel line of work focuses on high-resolution regional modeling. Studies such as Lin and Wang (2024) and Meng et al. (2024) demonstrate that regionally calibrated models can capture fine-scale spatial patterns and compound hazard effects. While these approaches achieve high local accuracy, they rely on context-specific data and are difficult to generalize. Our work instead prioritizes global applicability using consistent open datasets.

Finally, Kooshki Forooshani et al. (2024) propose a two-stage XGBoost framework for tropical-cyclone impacts in the Philippines, combining classification and regression with explicit treatment of class imbalance and interpretability. We extend this approach to the global scale and formalize comparisons against exposure-based baselines within a unified evaluation framework.

3 Methods

We present a globally applicable, gridded impact-forecasting pipeline that predicts the proportion of population affected by tropical cyclones (TCs) at 0.1° (∼11km) spatial resolution. The workflow comprises (i) an assembly of harmonized multi-source data layers, (ii) a transformation of administrative impact reports and multi-source features into a gridded pattern, (iii) a supervised learning with Extreme Gradient Boosting (XGBoost), and (iv) rigorous out-of-sample evaluation techniques,

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

including leave-one-out-cross-validation (LOOCV) and Walk-Forward validation. Figure 1 provides an overview of the modelling pipeline, highlighting the distinction between cell-level impact predictions (impact grid) and their aggregation into administrative-level impact maps.

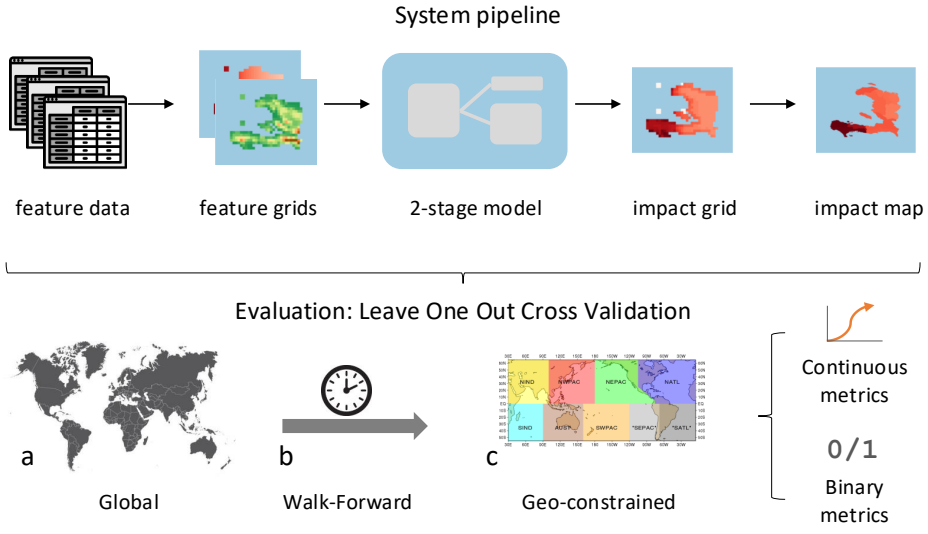


Figure 1. Overview of the gridded tropical cyclone impact pipeline. Input feature layers are harmonized into 0.1° grids and passed to the two-stage XGBoost model. The model first produces an *impact grid*, i.e. cell-level predictions of affected-population fractions, which are then aggregated into an *impact map* for reporting and visualization. Model performance is evaluated using three out-of-sample validation strategies: (a) leave-one-event-out cross-validation, (b) temporal walk-forward validation, and (c) geographically constrained validation, and is assessed using both continuous and binary metrics.

3.1 Data Description and Standardization

Our emphasis on openly available, globally consistent predictors follows recent IBF practice that prioritizes reproducibility and portability across regions (Kam et al., 2024; Kooshki Forooshani et al., 2024). While several studies demonstrate regional gains from bespoke, high-resolution inputs (e.g., a 1-km CNN for Zhejiang, China that integrates detailed environmental layers (Lin and Wang, 2024)), we opt for globally available features to enable cross-country evaluation and transferability without local re-engineering.

We assemble openly accessible layers describing hazards, topography, settlement morphology, demographics, and vulnerability: IBTrACS cyclone tracks/intensity, NASA GPM/IMERG rainfall accumulations, SRTM elevation/slope/ruggedness, coastline metrics from GADM, GHSL Degree-of-Urbanisation (urban/rural/water), WorldPop 2020 population, JRC Global Flood Maps (river flood depth), World Bank/GFDRR rainfall-triggered landslide frequency, and COAST-RP storm-tide levels. All raster-format datasets are ingested at native resolution, reprojected to WGS84, clipped to land, and harmonised onto a 0.1°×0.1° ($\sim$ 11×11 km^2) global grid pattern. Continuous rasters are aggregated using bilinear interpolation, whereas categorical layers are mapped using an area-weighted majority rule. Small gaps are imputed using inverse-distance weighting. For

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

each layer, we record the dataset name, description, native resolution, and source; the full set of predictors and the target are summarized in Table A1. All approaches were evaluated using identical input data, spatial resolution, and validation procedures to ensure a fair comparison.

3.1.1 Climate and Weather

Cyclone track and intensity are taken from the International Best Track Archive for Climate Stewardship (IBTrACS) (Knapp et al., 2010; Gahtan et al., 2024), and include 6-hourly center positions, central pressure, environmental pressure, and maximum sustained windspeed. We interpolate track kinematics between observations and apply the CLIMate ADAptation (CLIMADA's) (Aznar-Siguan and Bresch, 2019; Bresch and Aznar-Siguan, 2021) built-in wind model (Holland, 2008) to estimate event-maximum 1-min sustained windspeed at 10m above ground, at the grid level (`WEA_wind_speed` feature). To reduce artifacts from discretized track geometry and temporal sampling, we apply spatial smoothing to the resulting wind fields before extracting grid-level wind predictors. Accumulated rainfall (`WEA_rainfall_max_24h` feature) is derived from NASA GPM IMERG-Late (Huffman et al., 2023) (30 min, 0.1°) as the maximum 24-hour total within a ±48 h window around landfall (or closest approach for non-landfall events). We use IMERG Late data because it reflects the information that would be available in real-world AA applications: its low latency enables rapid impact estimation before event onset.

3.1.2 Topography and Coastline

Following established risk assessment frameworks Marconi et al. (2016); Lyu and Yin (2023), we derive topographical features from the 90m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model NASA Shuttle Radar Topography Mission (SRTM) (2013); Jarvis et al. (2008). For each grid cell, we calculate the mean elevation (`TOP_mean_elevation_m`), mean slope (`TOP_mean_slope`), and mean ruggedness (`TOP_mean_rug`).

Coastline presence (`TOP_with_coast`) and length (`TOP_coast_length`) are computed from the Global Administrative Areas (GADM) database GADM (2023), which provides high-resolution shapefiles of country boundaries. Missing topographical values, particularly on small islands, are interpolated using Inverse Distance Weighting (IDW) based on nearest neighbors. In specific cases where SRTM data is unavailable at the country level (e.g., Iceland, Svalbard, Jan Mayen, and the Faroe Islands), values are assigned based on median national estimates derived via IDW.

3.1.3 Secondary Hazards and Vulnerability

Primary hazards of tropical cyclones often unfold into different secondary hazards (Peduzzi et al., 2009), triggering landslides, coastal and river flooding, among other hazards. Here, we incorporate a series of hazard-interaction layers that comprise rainfall-triggered landslide frequency (World Bank) and storm-tide level (COAST-RPv2). All rasters are clipped to land, resampled to 0.1°, and gap-filled with IDW.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Landslides Risk. The "Global landslide hazard map" dataset (World Bank, 2022) provides frequency estimates at the raster level for landslides triggered by seismicity and rainfall. To harmonize the data, we aggregate the rainfall-triggered landslides frequency counts by grid cell (`SH_landslide_risk` feature).

Storm Surges. The Global COastal dAtaset of Storm Tide Return Periods (COAST-RP) dataset (Dullaart et al., 2022) provides Storm Tides level (in meters) for various return periods of storm tides (combination of the surges and tides). We estimate the maximum value of storm tide due to tropical cyclones in coastline grid cells (`SH_storm_tides` feature). This feature is 0 in all non-coastline grid cells, while the missing coastal grid cell values are estimated via the IDW approach.

River flooding. A river-flooding variable was not included in this dataset, as no truly global source of information for it exists. For instance, the River Flood Risk Index Baugh et al. (2024) does not provide coverage for certain regions, particularly small islands, as river basins smaller than 500 km^2 are excluded. As a result, 22 of the 72 countries present in the impact dataset lack corresponding hazard information, thereby limiting the geographical applicability of our model.

Socioeconomic Vulnerability. Similarly, socio-economic vulnerabilities were not included as there is not a single globally available feature. We studied the Sub-National Human Development Index (SHDI) (Smits and Permanyer, 2019), but 19 out of 72 countries present in the impact dataset weren't present.

3.1.4 Settlement Morphology and Population

Urbanization. Peduzzi et al. (2012) observed that tropical-cyclone impacts tend to be more severe in rural and remote settings where early-warning coverage, infrastructure robustness, and emergency response are comparatively limited. To capture this settlement-specific vulnerability, we derive three covariates: `URB_urban`, `URB_rural`, and `URB_water`, from the Degree of Urbanization (DoU) 2025 epoch of the Global Human Settlement Layer (GHSL) (Pesaresi et al., 2024). The GHSL assigns each 1km pixel a DoU class: water (DoU=10), rural (DoU=11–13), or suburban/urban (DoU=21–30). We intersect this raster with the 0.1°grid and compute, for every grid cell, the proportion of pixels in each class. These proportions are normalized so that the sum of urban, rural, and water equals 1, and supplied as separate continuous predictors. Incorporating these fractions allows the model to modulate hazard effects by local settlement morphology and infrastructure quality.

Population Estimates. We employ estimates of total population per grid cell (`total_pop`) using the 2020 UN-adjusted WorldPop dataset (WorldPop, 2020). WorldPop provides gridded estimates of resident counts at a 100m resolution, generated by disaggregating official census data and demographic projections for the year 2020. For each grid cell, we summed the counts of all 100 m raster tiles intersecting the cell to determine its population.

3.1.5 Event History and Target Construction

To capture the baseline susceptibility of each location, we introduce the covariate `prev_events_5years`, defined as the number of reported impacting tropical cyclones (TCs) whose reported impacting area intersected a given 0.1°× 0.1°grid cell during the five years preceding the target event. Cyclone tracks are obtained from the Emergency Events Database (EM-DAT)

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

(Delforge et al., 2025). When a complete five-year history is unavailable (e.g., for newly established states), we use the longest retrospective window that EM-DAT provides.

The covariate `prev_events_5years` should be interpreted as a proxy for baseline susceptibility and reporting propensity rather than as a purely physical vulnerability measure. Regions that repeatedly report impacts from past tropical cyclones may reflect persistent exposure patterns (e.g., settlement location, infrastructure fragility, geomorphology), structural vulnerability (e.g., housing quality, access constraints), and institutional reporting capacity. While this feature incorporates information derived from historical impact records, it does not encode information about the target event itself and is computed strictly from prior events within a rolling window. From an operational standpoint, the inclusion of this feature is pragmatic: anticipatory action systems aim to prioritize areas where impacts are most likely to occur and to be experienced by populations, and historical recurrence patterns provide a practical and interpretable signal of such baseline susceptibility. Nonetheless, we recognize that this feature may partially reflect systematic differences in reporting practices across countries, and therefore, it is treated as a combined proxy for vulnerability and impact observability rather than as a direct causal driver. Operationally, the dominant contribution of this feature in both stages (see section 4.2) should be interpreted with caution: regions with stronger EM-DAT reporting histories will systematically receive higher predicted impacts, which may reflect data coverage rather than physical risk. Practitioners deploying this model should be aware that predictions in data-sparse regions may understate risk. A version of the model without this feature is evaluated in Appendix E and may be preferable in contexts where reporting equity is a concern.

EM-DAT defines affected people as those requiring immediate assistance during a period of emergency, including displaced, injured, or otherwise impacted populations[1]. Our target variable is the proportion of the population affected by each TC according to EM-DAT. Our choice of population fraction as a target variable aligns with findings by Fang et al. (2014), who noted that population exposure is a more critical metric for developing nations, where high population density often coincides with lower financial infrastructure risk. Between 2000 and 2022, EM-DAT records 922 TC events in 72 countries. The highest event frequencies occur in the Caribbean, North America, the western Pacific, and eastern Africa; China and the Philippines alone account for 27% of all reported events. Figure 2 shows the number of occurrences and impacts of tropical cyclones at the sub-national level, overlaid with storm tracks. Regions shaded in white indicate areas for which impact data is not present in the EM-DAT dataset.

EM-DAT reports impacts at heterogeneous administrative levels, ADM0 (country), ADM1 (first-level sub-national), and occasionally ADM2 (second-level sub-national). Because ADM2 data is sparse, we first harmonize all records at the ADM1 level. We then downscale these ADM1 impacts to the analysis grid (0.1°resolution) by proportionally allocating the reported figures according to the WorldPop population distribution. This standardization yields a consistent, gridded estimate of the affected population per event, which serves as the target in all subsequent modeling. From the 2020 WorldPop raster, we obtain the resident population in each grid cell $N_{\text{grid}}^{\text{ppl}}$. Summing over the cells that lie within an administrative region yields its total population: $N_{\text{reg}}^{\text{ppl}} = \sum_{\text{grid}\in\text{reg}} N_{\text{grid}}^{\text{ppl}}$ For the same region, EM-DAT reports the total number of people affected by the cyclone, $N_{\text{reg}}^{\text{ppl\_aff}}$. Because the spatial distribution of these affected individuals is unknown, we assume the impact is homogeneous within

[1] https://doc.emdat.be/docs/data-structure-and-content/impact-variables/human/

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

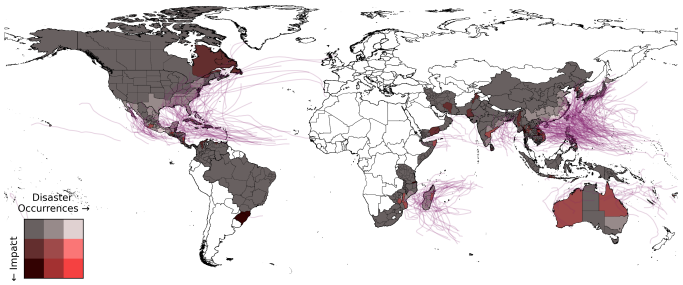


Figure 2. Sub-national impacts based on both event occurrences and affected population. A gray color scale represents the number of occurrences, where dark gray indicates high frequency. A red scale indicates the level of impact, with bright red denoting a high proportion of the population affected, and pure gray indicating no impact at all. Darker tones indicate regions with high values for both occurrences and impact ranges of (i) *No impact*: 0%; (ii) *Low*: 1–15%; (iii) *High*: $> 15\%$. Regions not reported in the EM-DAT dataset are colored white, and EM-DAT TC tracks are represented by purple lines.

the region and allocate it proportionally to the population:

$$N_{\text{grid}}^{\text{ppl\_aff}} = \frac{N_{\text{reg}}^{\text{ppl\_aff}}}{N_{\text{reg}}^{\text{ppl}}} \cdot N_{\text{grid}}^{\text{ppl}}. \tag{1}$$

By construction, $N_{\text{reg}}^{\text{ppl\_aff}} = \sum_{\text{grid}\in\text{reg}} N_{\text{grid}}^{\text{ppl\_aff}}$, where $N_{\text{grid}}^{\text{ppl}}$ is the total number of people by grid and $N_{\text{grid}}^{\text{ppl\_aff}}$ is the total number of people affected in that grid. So the regional totals reported by EM-DAT are exactly preserved while delivering internally consistent impact estimates at the grid resolution required for modeling. Moreover, note that we are not assigning impact to grid cells with no population (Eq. 1 is zero). It is also worth noting that grid cells intersecting multiple ADM1 regions are assigned to the ADM1 unit with which they share the largest area.

3.2 Model architecture

We frame cyclone–impact prediction as a supervised learning task at the 0.1°grid level using the predictors in Table A1. The end-to-end workflow is shown in Figure 3. Predictions at the grid level can be re-aggregated to any administrative boundary for reporting.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

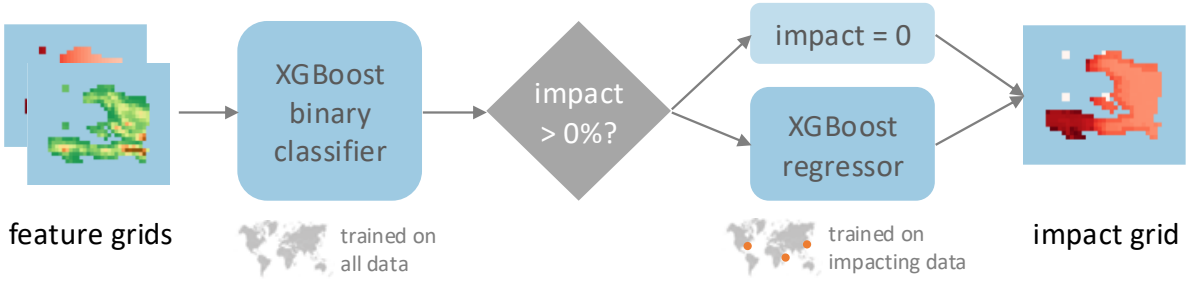


Figure 3. Two-stage XGBoost Model workflow. Stage 1 is a binary impact classifier, while Stage 2 is a regressor.

Two-stage XGBoost Model. We propose a model that incorporates the extensive historical and event information summarized in Table A1. This hybrid formulation is conceptual rather than model-based, and is intended to reflect how different predictive components may be combined in operational decision pipelines. The two-stage (classifier→regressor) design is motivated by prior IBF work that separates occurrence from severity to address extreme class imbalance and operational decision points (Kooshki Forooshani et al., 2024). We train a two-stage model at the grid level:

- **Stage 1 — occurrence (classifier).** An XGBoost classifier (logistic loss, `max_depth`$= 4$, `learning_rate`$= 0.01$, `n_estimators`$= 100$, using Python's `xgboost` library) predicts whether any impact occurs. To address the imbalance, we under-sample negatives to a 3:1 ratio (non-impact:impact). The decision threshold is fixed at 0.5; a cell is labeled as "impact" if the reported impact fraction > 0.

- **Stage 2 — severity (regressor).** For cells predicted/observed as impacted, an XGBoost regressor (squared-error loss, `max_depth`=4, `learning_rate`=0.01, `n_estimators`=100) estimates the affected-population fraction. Training uses under-sampling with a 3:1 ratio of low- to high-impact samples (categories defined in §3.3). We use squared-error loss for stability with sparse, right-skewed targets.

In this way, and unlike parametric impact-function approaches (Kam et al., 2024), we learn data-driven mappings from multi-driver predictors to affected population fractions while retaining transparent, exposure-based baselines for auditability. This hybrid approach combines the strengths of the machine learning model and exposure-based rules, and is implemented as a sequential, decision-level framework rather than a statistical ensemble. The two components are not combined at the prediction level, but instead provide complementary signals corresponding to different decision tasks. This design reflects the distinction between detecting potential impacts and identifying severe, operationally relevant events, and allows each approach to be evaluated in the context where it is most informative. Specifically, the two-stage XGBoost model is used to detect whether impacts are likely to occur and to localize affected regions (0% threshold), while exposure-based thresholds are applied to identify high-impact conditions (15% threshold).

Impact severity is operationalised through a decision threshold on the affected-population fraction. We define high-impact events as those affecting more than 15% of the population within an administrative unit. This threshold is intended to capture

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

events requiring operational attention, consistent with anticipatory action frameworks where intervention triggers are based on predefined impact levels rather than universal physical thresholds Coughlan de Perez et al. (2015).

Accordingly, severity is treated as a binary classification task (above vs below 15%), rather than as a continuous measure of impact magnitude. The 0% threshold captures the detection of any impact, while the 15% threshold focuses on identifying more severe, operationally relevant events. This distinction is important for interpreting model performance, as it reflects the ability to detect high-impact cases rather than to fully characterize the distribution of impacts.

The choice of the 15% threshold is also informed by the data distribution. Affected-population fractions are highly skewed, with most ADM1–event pairs corresponding to no or low impact and only a small fraction exhibiting large impacts. Lower thresholds (e.g., 10%) would include many marginal cases that are difficult to distinguish from noise, while higher thresholds (e.g., 20%) would yield too few positive examples for robust evaluation. The 15% cutoff therefore provides a meaningful separation between moderate and high-impact cases while maintaining sufficient sample size. It should be interpreted as a decision-relevant operational threshold rather than a universal definition of severity.

Historical Baseline. For each country, the affected population at ADM0 is estimated as the country-specific median fraction observed across past tropical-cyclone events (set to zero when no history exists).

Windspeed-based Baselines. The windspeed-exposed model is a wind-threshold *trigger proxy* reflecting common operational exposure rules. It assigns as affected the full population located in grid cells where maximum sustained wind exceeds 33 m/s (Category-1 on the Saffir–Simpson scale (Simpson and Saffir, 1974)). While this assumption is not intended to represent a calibrated impact function, it provides a transparent and conservative baseline for evaluating the spatial extent of wind-based activation. The affected population is estimated in two ways: (i) *windspeed-exposed*: the entire population in windspeed-exceeding cells are treated as affected; and (ii) *windspeed-historical*: exposed population is multiplied by the country-specific historical median impact fraction. Such threshold-based exposure rules are widely used in operational IBF workflows, including global humanitarian trigger systems such as WFP ADAM (World Food Programme, 2016) and related forecast-based financing implementations (Sharma et al., 2020). It is important to note that the windspeed-exposed baseline is not intended as a calibrated impact model, but rather as a transparent "trigger proxy" that reflects operational practice in anticipatory action systems, where wind thresholds are commonly used to delineate exposed areas and activate early response. Accordingly, these baselines represent an intentionally conservative upper-bound assumption in which the entire exposed population is treated as affected once the wind threshold is exceeded.

3.2.1 Model Training and Evaluation

Prediction target and aggregation. Models predict the fraction of population affected at 0.1°grid level. Let g be the grid cell index and r an administrative region (ADM1 by default; ADM0 when only national totals are reported). With predicted fraction $\hat{f}_g \in [0,1]$ and population N_g^{ppl}, the predicted affected population at region level is $\hat{N}_r^{\text{ppl}} = \sum_{g \in r} \hat{f}_g \, N_g^{\text{ppl}}$. If a region's total $\hat{N}_r^{\text{ppl}}$ must be mapped back to cells, we redistribute proportionally to population: $\hat{N}_g^{\text{ppl}} = \frac{N_g^{\text{ppl}}}{N_r^{\text{ppl}}} \hat{N}_r^{\text{ppl}}$, $g \in r$.

Impact categories. For categorical assessment, we use three levels for the fraction of population affected per region–cyclone observation: (i) *No impact*: 0%; (ii) *Low*: 1–15%; (iii) *High*: $> 15\%$. The 0% threshold separates impact from no-impact; the

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

15% threshold flags high-impact regions, following operational guidance National Disaster Risk Reduction and Management Council (2019). Evaluating at two decision thresholds, 0% (any impact) and 15% (high impact), aligns outputs with typical activation/triage choices, and enables a layered view in which predicted impacts support coverage and localization, while exposure serves as a conservative severity flag (Sedhain et al., 2025; Kam et al., 2024).

Training protocol. For the two-stage classifier, the ratio between impacting and non-impacting regions (1st stage) and between low-impacting and high-impacting regions (2nd stage) is tuned to optimize model performance. We selected a subset of 50 events that approximates the overall impact distribution observed in EM-DAT. We iterated through different under-sampling ratios and minimized the Frobenius distance between the sorted categorical confusion matrix and the identity matrix (representing an ideal confusion matrix). This process allowed us to determine the optimal sub-sampling ratios for each stage ($3:1$ ratio for both stages) and apply them to the full training dataset.

3.3 Evaluation with Reported Impact

Generalized prediction performance is assessed with *leave-one-cyclone-out* cross-validation (each cyclone is one fold) and with a *walk-forward* constraint: for a held-out cyclone, models are fit only on earlier events to avoid temporal leakage. We additionally test a basin-constrained variant, training only on events in the same basin, as defined by National Oceanic and Atmospheric Administration, Atlantic Oceanographic and Meteorological Laboratory (n.d.) (NOAA). This probes transfer against local-only training and complements domain-adaptation findings in regional studies (Wagenaar et al., 2021).

We include only events with sub-national reporting (ADM1 preferred) and require that the event affects at least 100 people in total (following EM-DAT's protocol). These constraints make the dataset include 780 events (out of 922 originally in the data). All comparisons are conducted at the ADM1 level.

We report:

- *Distance-based (regression) metrics:* Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) computed on $\hat{N}_r^{\text{ppl}}$ versus observed N_r^{ppl}
- *Categorical metrics:* Quadratic-Weighted Cohen's κ (QWK) for ordinal agreement on the three impact categories
- *Binary Classification Metrics:* Considering 0% for impact/no-impact and 15% for high-impact, we estimate the models' Precision, Recall, F1 score, Specificity, Accuracy, False Positive Rate (FPR), False Negative Rate (FNR), and Critical Success Index (CSI)

Table B1 in the Appendix summarizes all the standard formulations and definitions. This suite of metrics mirrors the decision-centric evaluation frameworks now standard in IBF audits Sedhain et al. (2025). By combining ordinal and magnitude-based perspectives, the evaluation supports both resource prioritization and sizing, while identifying specific error patterns that carry distinct operational consequences.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

3.4 Evaluation with Historical Forecasts

To explore how the model might behave under forecast conditions - not to validate it as operationally ready - we assess it using historical operational forecasts issued before landfall. These forecasts consist of multiple possible track paths, making an ensemble, and are updated as new forecast cycles become available.

Forecasts come from the TIGGE archive (National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce et al., 2008) (Korean Meteorological Services -KMS- tracks), issued every 12 h with valid times every 6 h out to 120 h and up to 25 ensemble members per issuance. We successfully matched 26 EM-DAT reported events (and their associated IBTrACS tracks) across six countries between 2012 and 2020 with their corresponding historical forecast tracks (detailed event list available in Appendix F). For each track, we use the last 72 h before landfall (or closest approach for non-landfall cases). When multiple issuances map to the same lead time for a given storm, we retain the latest issuance, while all available ensemble members are used. Grid–level 10 m winds are generated on a 0.1° grid with CLIMADA's Holland (2008) model, using the same configuration for both forecast and a-posteriori best-track runs. No additional harmonization of wind conventions beyond CLIMADA defaults is applied.

It is important to clarify that the model itself is trained exclusively on observational hazard predictors (IMERG-based rainfall and wind fields derived from observed tracks) together with EM-DAT impact records. In the historical forecast experiment, only the track-based wind hazard is replaced by TIGGE ensemble forecasts, while rainfall predictors remain observation-derived (IMERG Late). Thus, the forecast evaluation should be interpreted as a hybrid experiment that isolates uncertainty propagation from track and wind prediction, rather than as a fully forecast-driven multi-hazard IBF system. Given the small sample of 26 events across six countries and the hybrid nature of this experiment, results should be treated as an indicative proof-of-concept rather than a generalizable operational validation.

For each issuance and ensemble member, the trained two-stage model (as in Section 3.2 and evaluated per Section 3.3) produces grid-level impact predictions. We summarize the ensemble with a fixed order of operations. Central estimate takes the median across members at the grid level and then aggregates that median field to ADM1 (ADM0 when required) using the population-weighted mapping in Eq. (3.2.1). For the uncertainty ribbon, we first aggregate each member's grid-level predictions to ADM1, then compute the interquartile range (IQR) across members; ribbons therefore reflect ensemble spread at the reporting unit, not statistical confidence intervals. This median and IQR summary follows ensemble-aware IBF practice that encourages explicit communication of uncertainty for anticipatory action as suggested by Kam et al. (2024), and provides a single actionable map and spread that aligns with common agency workflows.

A prediction is computed at each lead time using the same decision thresholds as in Section 3.3: (i) impact vs. no-impact at 0% and (ii) high-impact at 15%. Precision and recall are evaluated from the *ensemble-median* ADM1 predictions, and the IQR ribbons quantify member-to-member spread. As a non-forecast benchmark, we run the identical pipeline along the curated observed track for each target storm, using the same two-stage model and aggregation; the 26 target storms are excluded from training in both the hindcast and benchmark runs.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Differences across historical-forecast tracks and a-posteriori best-tracks were analyzed by computing the absolute differences in high-windspeed areas. When contemplating all 26 events, we found a mean absolute difference of $\sim 46\%$ in high–wind grid–times; all winds were harmonized to 1-min sustained values, so discrepancies arise mainly from (i) small track and intensity errors that shift the high-wind ring across grid cells, and (ii) parametric choices in the Holland wind reconstruction (e.g., the radius of maximum winds $R_{\max}$ or the Holland's shape parameter B) interacting with forecast uncertainty. This finding points to the necessity of improving the forecasting models, especially in the high-windspeed conditions, as their errors would propagate to impact prediction models, such as that described in this work.

4 Results

In this section, we assess model performance using the protocol described in Section 3.3. We first compare the proposed model with the baselines using historical tracks, and examine sensitivity to geographical and temporal variation. We then evaluate the models under forecast uncertainty by replacing a posteriori tracks with ensemble forecast tracks. Finally, we present case studies that illustrate the strengths and limitations of the approach.

4.1 Evaluation with Historical TC Tracks

We first perform leave-one-event-out cross-validation (LOOCV) as an upper-bound estimate of predictive performance. Because LOOCV evaluates cross-event generalization without enforcing temporal causality (training may include events occurring after the test event) we complement it with walk-forward validation (Section 4.3), which more closely reflects operational deployment where only past events are available for training.

We consider the detection of any impact (0%) and high impact (15%), and explore the error structure magnitude-based and by impact category. To aid interpretation, we group evaluation metrics according to their operational meaning. MAE and RMSE quantify errors in the continuous affected-population fraction and therefore assess the accuracy of impact sizing. Binary metrics (precision, recall, F1) evaluate triggering performance at decision thresholds (0% for any impact and 15% for severe impact). Finally, the Quadratic Weighted Kappa (QWK) measures ordinal agreement across impact categories, reflecting performance in escalation decisions. Together, these metrics provide a multi-dimensional view of performance aligned with early-warning workflows: detect, escalate, and size. Throughout, we interpret binary metrics as measures of trigger performance, QWK as escalation agreement, and MAE/RMSE as impact-sizing fidelity.

Detecting any impact (0% threshold). At the 0% decision threshold (impact vs. none), the two-stage XGBoost provides the best balance of sensitivity and precision, with Precision $= 0.30$, Recall $= 0.68$, and the highest F1 $= 0.42$ (Table 1). While this represents the strongest performance among the models tested, an F1 of 0.42 indicates substantial room for improvement and highlights that global IBF from open predictors alone remains a difficult problem. For context, more than half of predicted impacts are false positives (Precision = 0.30), and nearly one-third of true impacts are missed (FNR = 0.32). Windspeed baselines are precise but insensitive (windspeed–exposed: 0.57/0.17/0.27; windspeed–historical: 0.56/0.15/0.24 for Precision/Recall/F1), while the historical baseline attains very high recall (0.95) at the expense of precision (0.14), yielding F1$= 0.25$. Specificity/ac-

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Table 1. Binary and aggregate metrics at ADM1. Binary metrics are shown for two thresholds: **0%** (impact vs. none) and **15%** (high impact). Lower MAE and RMSE, and higher QWK are better. Best per column in **bold**.

	0% (impact vs. none)			**15%** (some vs. high impact)			**Aggregate**		
Model	Precision	Recall	F1	Precision	Recall	F1	MAE	RMSE	QWK
Historical	0.14	**0.95**	0.25	0.04	0.05	0.04	4.77	13.93	0.00
Windspeed-exposed	**0.57**	0.17	0.27	**0.22**	0.44	**0.30**	4.30	19.18	**0.31**
Windspeed-historical	0.56	0.15	0.24	0.01	0.01	0.01	**1.57**	9.99	0.19
2-stage XGBoost	0.30	0.68	**0.42**	0.15	**0.50**	0.23	3.67	**9.52**	0.29

curacy confirm these trade-offs (Table B2 in the Appendix). This pattern is consistent with recent operational audits showing that simple exposure-style triggers favor precision at the expense of misses, whereas trained models improve coverage at low thresholds (Sedhain et al., 2025).

Detecting high-impact regions (15% threshold). High-impact observations are rare and therefore statistically challenging to predict. At the 15% threshold, the windspeed-exposed baseline performs best with F1 = 0.30 (Precision = 0.22, Recall = 0.44). The 2-stage XGBoost model reaches F1 = 0.23 (0.15, 0.50), whereas the historical and windspeed-historical baselines are near zero (0.04 and 0.01) (see Table 1). Even the best model has FNR = 0.56 at 15%, underscoring persistent difficulty in capturing severe events (Table B2 in the Appendix). Because high-impact observations are rare and most ADM1-event pairs fall in the no- or low-impact range, a conservative model that rarely predicts severe impact can still achieve low MAE. This explains why windspeed-historical attains the lowest MAE despite performing poorly on high-impact detection. For the ordinal metric, windspeed–exposed obtains the highest QWK$= 0.31$, followed by 2–stage XGBoost (QWK$= 0.29$); windspeed–historical reaches 0.19 and historical 0.00. This discrepancy arises because QWK evaluates ordinal agreement across impact categories, whereas MAE and RMSE penalize errors in continuous magnitude. The windspeed–exposed baseline frequently assigns the correct severity category (e.g., flagging high-impact regions), improving QWK, but systematically overestimates affected fractions by assigning the full exposed population as impacted. These large magnitude errors are strongly penalized by RMSE, explaining the divergence between ordinal and continuous metrics. We further evaluate the contribution of the `prev_events_5years` feature (Appendix Table E1) and find that its removal has limited impact at the 0% threshold, while improving performance for high-impact detection (15%) in terms of F1 and CSI. The difficulty of high-impact detection, where the best model achieves FNR = 0.56 (Table B1), is operationally significant. In humanitarian settings, missing more than half of severe events carries direct consequences for pre-positioned resources. These results reinforce that the current model is suited to a triage and coverage role, and should not be used as the sole basis for activation decisions at severe-impact thresholds. These findings un-

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

derscore the persistent difficulty of severe-impact prediction, consistent with single-event audits where exposure-style triggers can outperform statistical models under strict high-damage thresholds (Sedhain et al., 2025).

Error structure. Fig. 4 shows the distribution of absolute errors across impact categories, both over the full range (top) and with a zoom near zero (bottom). Absolute-error distributions are tight for the *no* and *low* categories (together ∼98% of data) and more dispersed for the *high* category across all models. Because most ADM1–event pairs correspond to no or low affected fractions, the conservative windspeed–historical model achieves the lowest MAE (1.57), despite failing to detect severe cases. The 2-stage XGBoost model is the strongest approach at the 0% threshold, i.e. for detecting whether any impact occurs at all. This is an important result operationally, because it means the model is best suited for early identification and spatial localization of potentially affected regions. While windspeed-based baselines remain stronger for conservative flagging of severe impacts, the XGBoost model provides the best overall balance of recall and precision for broad impact detection.

The error metrics in Table 1 reflect these insights. Because most ADM1–event pairs correspond to no or low affected fractions, the conservative windspeed–historical model achieves the lowest MAE (1.57), despite failing to detect severe cases. This pattern is evident in every impact category (Fig. 4). In contrast, the windspeed-exposed baseline performs worst on RMSE (= 19.18) because it assigns the full exposed population as affected, producing extreme overestimates that are heavily penalized by squared error.

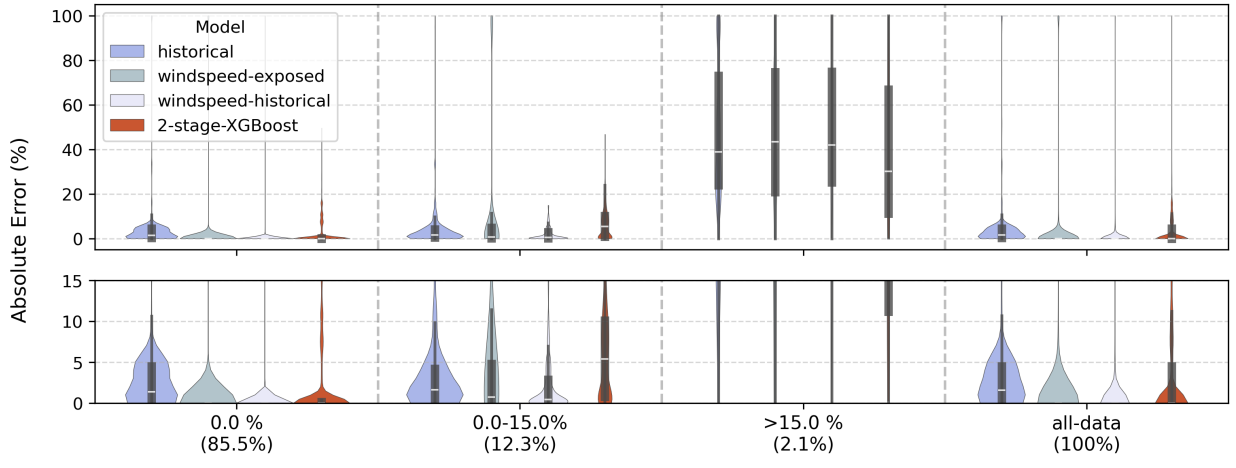


Figure 4. Error distribution: absolute error as percentage of the affected population, stratified by impact category (0%, 1–15%, ≥15%) and model, with an additional all-data panel. Values in parentheses indicate the share of ADM1 observations per category. **Top:** full error range. **Bottom:** a zoomed-in view.

In this context, localization refers to correctly identifying impacted ADM1 units without systematically expanding predicted impacts to neighboring non-affected regions. The model confusion matrices (Fig. 5) show that windspeed baselines correctly

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

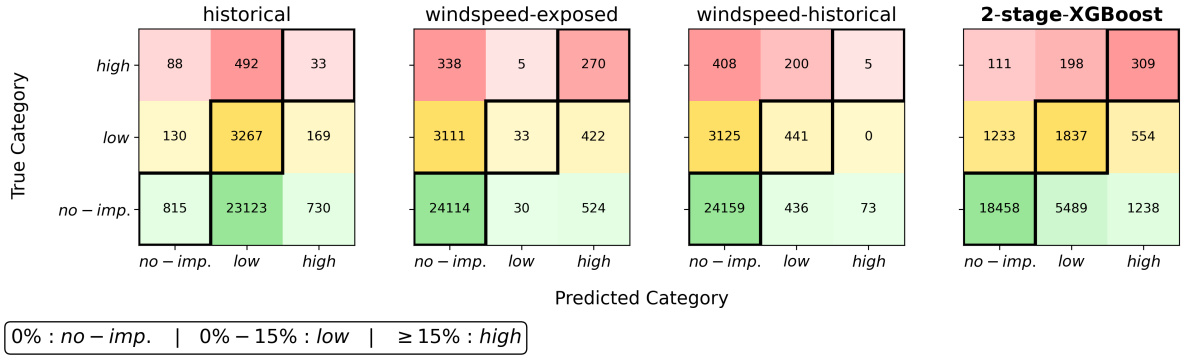


Figure 5. Confusion matrices of events-ADM1-regions for four models (historical, windspeed–exposed, windspeed–historical, 2-stage XGBoost), comparing predicted vs. observed impact categories. Rows show true classes and columns predicted classes (*no-impact* = 0%, *low* = 0–15%, *high* ≥15%). Cells report counts (row-normalized shading; darker = larger share).

identify many *no-impact* regions but tend to underestimate impacting areas; windspeed–historical often predicts *no-impact* for truly *high* cases, while windspeed–exposed alternates between *no* and *high*. The historical model favors the low impact category. This happens because it assigns some impact to all regions of the country, based on the distribution of historical impacts by country-event. In particular, considering the high impact class, the windspeed-historical model under-predicts almost all areas (predicting no-impact) while the windspeed-exposed model tends to predict either no impact or high impact, attaining a better precision for the high class. The last observation is understandable due to the nature of the windspeed-exposed model: it assigns all population as affected in areas of high windspeed, even in highly populated urban areas. The 2-stage XGBoost model is the only approach that clearly separates *no* from *low* impact and increases recovery of *high* cases, albeit at the cost of additional high-class false alarms.

4.2 Feature attribution (SHAP analysis)

To interpret the relationships captured by the two-stage model, we compute SHAP values for both stages (Fig. E1 in the Appendix. We emphasize that SHAP values quantify feature contributions to model predictions and should not be interpreted as evidence of causal impact drivers. In the first-stage classifier (impact vs. no impact), population and recent event history are the most influential predictors, together with hazard variables such as 24-h accumulated rainfall and wind speed. Higher rainfall and stronger winds are associated with increased predicted impact likelihood, whereas low population and low recent-event exposure contribute negatively to impact detection. Terrain-related predictors (elevation, slope, landslide risk) exhibit smaller but consistent contributions, suggesting that geomorphology modulates the translation of hazard forcing into affected-population outcomes.

In the second-stage regressor (impact magnitude conditional on impact), the importance ranking shifts: recent event history and wind speed dominate severity prediction, followed by rainfall extremes. The strong contribution of recent event history

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

likely reflects a combined proxy for baseline susceptibility and systematic reporting capacity, and is therefore best interpreted as a prior risk signal rather than a direct physical driver. Population becomes comparatively less influential once impact occurrence is established, indicating that exposure primarily informs impact detection, whereas hazard intensity and accumulated forcing shape severity conditional on impact. Finally, coastal and storm-tide predictors show limited marginal contribution at the global scale; this may reflect collinearity with wind-based exposure measures and/or the lack of globally consistent coastal vulnerability indicators at comparable resolution. Overall, the SHAP patterns support the rationale of the two-stage design and provide an interpretable decomposition of factors associated with impact occurrence versus impact severity.

4.3 Sensitivity Analysis

Because time is an implicit driver of both exposure and reporting practices, we compare performance when models are trained only on past events (LOOCV-walk-forward), as opposed to the standard leave-one-cyclone-out (LOOCV-global) approach that uses all available events. We compare walk-forward and global training strategies by computing rolling 5-year binary metrics (precision and recall) over time, using bootstrapped confidence intervals on temporally aligned evaluation windows, allowing us to isolate the effect of future information leakage on predictive performance. Figure 6 illustrates the resulting rolling five-year precision and recall estimates for LOOCV-walk-forward and LOOCV-global training across time, for the two already defined impact thresholds. We therefore interpret walk-forward validation as the more realistic estimate of deployable performance, while LOOCV-global provides an upper bound on skill under full-sample information.

At the 0% threshold (detecting any impact), both training approaches demonstrate stable precision and recall over time, characterized by consistently high sensitivity but lower precision. Notably, the model achieves a recall of approximately 0.7 even with limited historical data, and incorporating future data points yields no significant performance gains. Considering the Threshold 15%, where we aim to detect high impact events, we find a discrepancy in performance. In terms of recall, the performance of the model trained on the past data approaches the performance of the LOOCV-global scenario only at around year 2018, at which point the model has 18 years (around 75% of all of training set) of historical data for training available. This suggests that LOOCV-global may overestimate operational performance for severe-impact detection because it leverages future high-impact events during training. Walk-forward validation is therefore the more realistic estimate of deployable performance, highlighting that reliable high-impact prediction requires a sufficiently large historical archive of severe events.

Both approaches show the models achieving low precision throughout the two decades of timeline, except for the increase in performance in the years 2018-2023, where recall also improves. The reasons why the TCs in that time period seem to be “easier” for the models to predict may be attributable to implicit feature correlations or period-specific climate patterns, and we leave for future work the temporal analysis of the model performance. This convergence suggests that a global IBF model requires a critical mass of historical events to stabilize performance, and that operational deployment would benefit from periodic retraining as additional events accumulate.

Similarly, we hypothesize that training data from more geographically proximate regions may improve performance. To test this, we evaluate the model using a geographically constrained approach, in which each leave-one-event-out (LOOCV) fold is trained exclusively on events belonging to the same NOAA basin as the test event. To further characterize this geographic

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

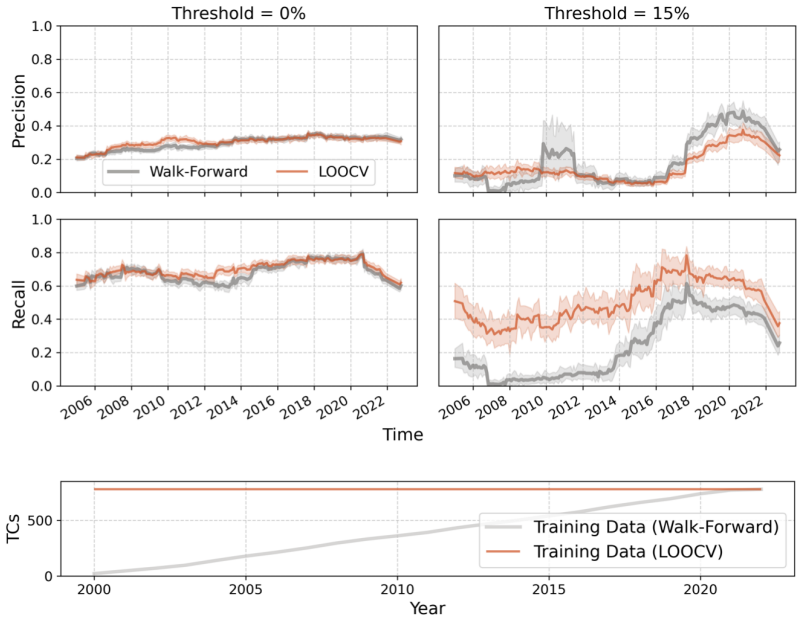


Figure 6. (a) Precision and recall of walk-forward vs. LOOCV training approaches over time, for 0% and 15% thresholds. Shaded ribbons indicating 95% bootstrap confidence intervals across held-out events. (b) Amount of events considered for the training of the two approaches, over time.

sensitivity, Table C1 in the Appendix reports performance stratified by cyclone basin for both the global and geo-constrained specifications.At the 0% threshold, recall under the global specification is consistently high across basins (0.63–0.83), indicating robustness in impact detection, although precision remains moderate (0.28–0.54). At the 15% threshold, performance varies substantially; while the global specification shows relative strength in the South Pacific ($F1 = 0.38$, $recall = 0.72$) and North Atlantic ($F1 = 0.31$, $recall = 0.59$), performance under the geo-constrained approach is weakest in the Australian region. In that basin, high-impact cases are too sparse to support reliable classification without external data. Ultimately, we find that this geographically constrained approach does not improve prediction over the global specification (Table C1). This suggests that the global baseline captures patterns that generalize across basins, effectively leveraging transferable information encoded in the gridded predictors to overcome local data sparsity.

4.4 Evaluation with Historical Forecasts

To assess how the modelling framework behaves under forecast uncertainty, we evaluate the 2-stage XGBoost model using historical tropical-cyclone forecast tracks (TIGGE) instead of a posteriori best tracks. The model itself is trained exclusively on observational hazard predictors (wind fields derived from best tracks and IMERG-based rainfall) together with EM-DAT impact records. In the forecast experiment, only the track-based wind hazard is replaced by ensemble forecast tracks; rainfall predictors remain observation-derived (IMERG Late). Accordingly, the results reflect forecast uncertainty in track and wind

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

fields only, rather than a fully forecast-driven multi-hazard IBF system. This setup should therefore be interpreted as a proof-of-concept demonstration of the ensemble pipeline, rather than as a full evaluation of operational forecast performance, and is based on a subset of 26 events.

Figure 7 illustrates an example of ensemble track spread and member-wise impact variability for TC Kammuri in the Philippines. Insets zoom into two representative grid cells (blue squares): the mini-map marks the selected cell, and paired histograms summarize member-wise variability (left: predicted impact fraction; right: 1-min sustained wind speed for the same members at that cell). The upper example shows low winds and correspondingly low predicted impacts across members; the lower example shows higher winds and a subset of members associated with larger predicted impacts. This figure illustrates how ensemble track dispersion propagates into spatially heterogeneous impact uncertainty at the grid scale. We observe that ensemble-derived impact distributions are generally non-normal, often exhibiting strong skewness in grid cells located farther from the mean forecast track. This behaviour is consistent with recent evaluations of IBF systems that highlight the propagation of forecast uncertainty into impact variability Sedhain et al. (2025).

To generate impact forecasts from ensemble track predictions, we apply the trained model to each ensemble member and pre-landfall forecast. We train the 2-stage XGBoost model on EM-DAT, excluding the 26 events matched to historical forecast

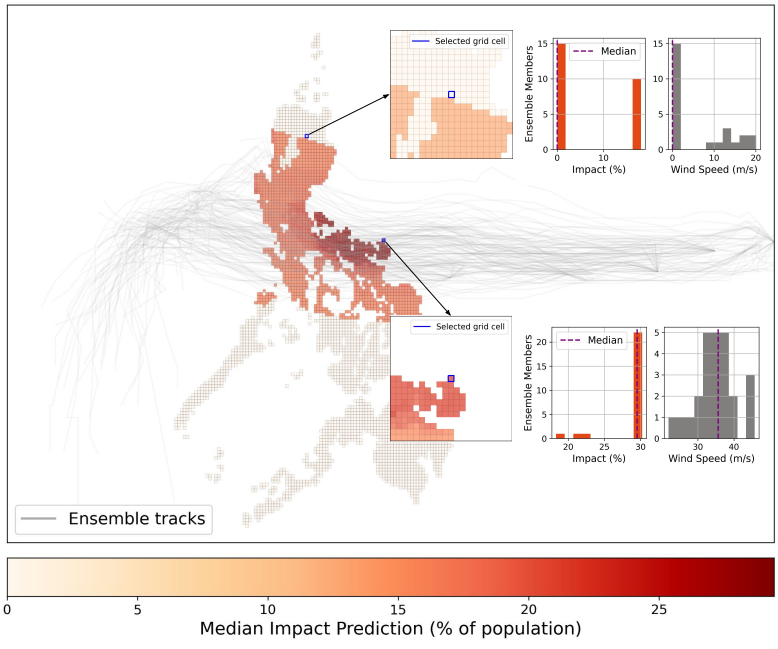


Figure 7. Ensemble-forecast example for the Philippines TC Kammuri. Gray lines show individual ensemble tracks; the background grid is colored by the ensemble median predicted impact. Insets zoom into two representative grid cells (blue squares): each mini-map marks the selected cell, and the paired histograms summarize outcome distributions across ensemble members (one track corresponds to one outcome): on the left, predicted impact (%); on the right, 1-min sustained wind speed (m/s) for the same members at that cell.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

tracks to avoid overlap between training and forecast evaluation. For every ensemble member and each pre-landfall forecast, the model produces grid-level impact predictions. Ensemble output is summarized at grid level using the median predicted impact across members as the central estimate and the interquartile range (IQR) as an uncertainty band; these summaries are then aggregated to ADM1 following Section 3.3. This contrasts with mean ± standard deviation summaries (e.g., Kam et al. (2024)), as member-wise impact distributions are often skewed and non-Gaussian, particularly in grid cells distant from the forecast track envelope. The ensemble-to-impact median/IQR summary aligns with ensemble-aware IBF practice that emphasizes robust central estimates and explicit uncertainty communication.

To contextualize ensemble behavior over time, we evaluate predictions as landfall approaches. Figure 8 reports ADM1-level precision and recall as a function of lead time for the two decision thresholds: 0% (any impact) and 15% (high impact). Curves are computed from ensemble-median impact predictions, with shaded bands indicating the IQR of member-to-member variability propagated to ADM1.

At the 0% threshold, the XGBoost model achieves substantially higher recall than the windspeed baseline across lead times, indicating improved ability to identify affected regions, albeit with lower precision. This results in a more balanced overall performance (F1), consistent with the aggregate results in Table C1.

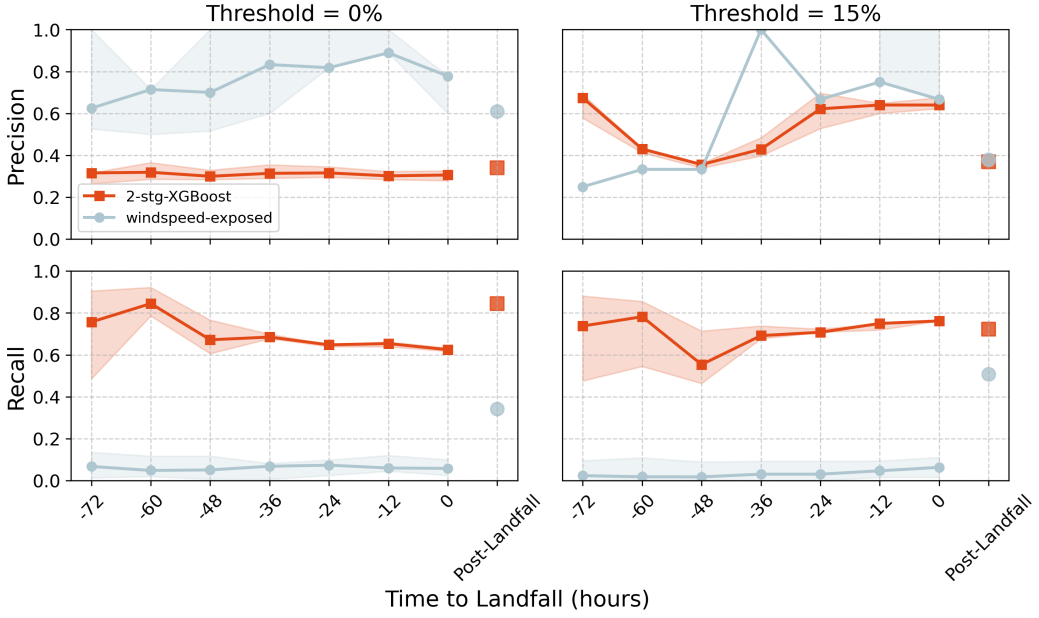


Figure 8. Lead-time prediction using the ensemble median aggregation. Lines show precision and recall for the impact-vs.-no-impact (0%) and high-impact (15%) decisions, and shaded ribbons indicate the IQR of ensemble-member variability propagated to ADM1. The historical track is shown as a reference benchmark ("Post-Landfall").

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

The historical-track ("Post-Landfall") benchmark is shown for reference. Two patterns emerge: median performance remains broadly comparable to the historical-track benchmark up to approximately 72 h before landfall, and the IQR narrows as landfall approaches, indicating decreasing ensemble spread.

These results indicate that the framework can produce stable impact estimates under forecast uncertainty within the limits of this hybrid setup. However, they should not be interpreted as a full assessment of operational forecast performance, as only track-related uncertainty is represented and the sample size is limited.

We note that these results isolate track-related forecast uncertainty; incorporating forecast-based precipitation predictors may alter lead-time performance and remains an important direction for future work. This trade-off between improved detection and increased false alarms is consistent with recent IBF evaluations Sedhain et al. (2025). Given the limited sample size and the hybrid observation–forecast design, these lead-time curves should be treated as illustrative rather than definitive.

4.5 Case studies

We present four case studies that juxtapose the windspeed–exposed trigger with the 2–stage XGBoost model, by providing examples of scenarios in which each performs better, visualized in Figure 9.

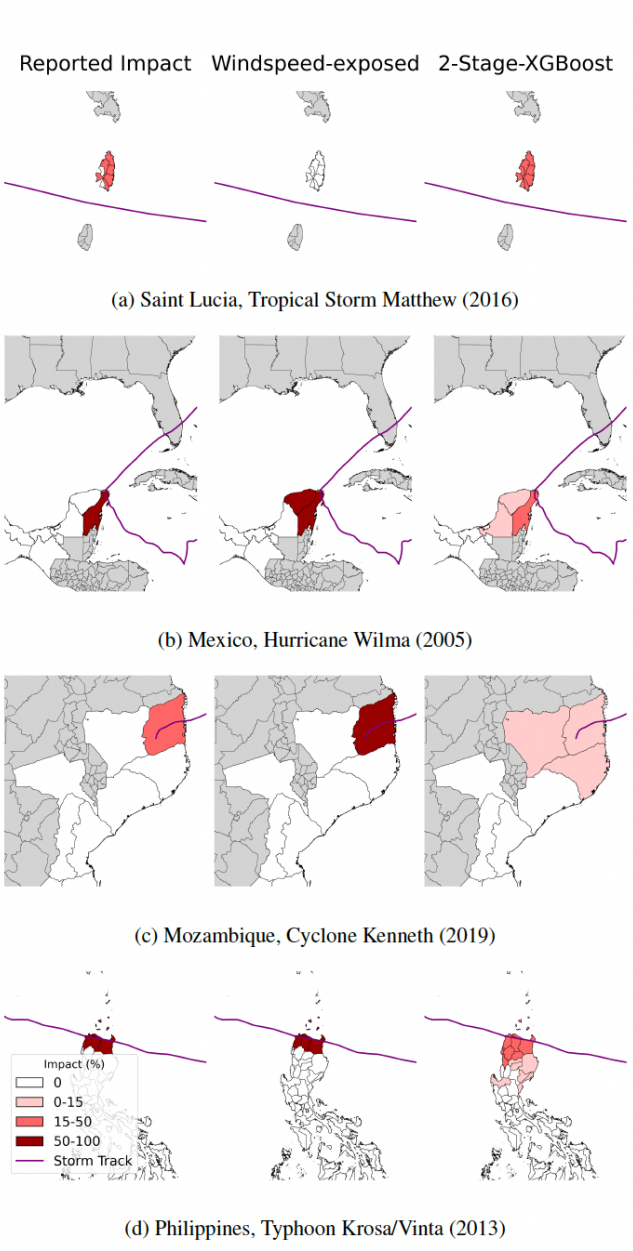
Saint Lucia, Tropical Storm Matthew (2016): 508 mm max. accumulated rainfall with tropical-storm-force winds (two-stage outperforms windspeed-exposed). During Matthew's passage, Saint Lucia experienced heavy rainfall, with national authorities activating preparedness and response measures. Matthew passed between Saint Lucia and Saint Vincent around 1800 UTC 28 September while still at tropical-storm intensity [2]. In this rain-dominated setting, the windspeed-exposed model (threshold 33 m/s) fails to activate, whereas the two-stage model flags the reported ADM1 as impacted (Fig. 9a). This behavior is consistent with evidence that *non-wind drivers* (e.g., rainfall interacting with terrain and settlement form) can dominate affected-population outcomes under sustained winds that remain below wind-impact thresholds (Lin and Wang, 2024; Meng et al., 2024).

Mexico, Hurricane Wilma (2005): prolonged winds with high rainfall (two-stage outperforms windspeed-exposed). Hurricane Wilma stalled over NW Caribbean, exposing the region to very strong winds (up to 82 m/s) over an extended period, together with a 24-h rainfall accumulation of ≈1576 mm recorded at Isla Mujeres, Mexico. Here, a single coastal ADM1 was reported as highly affected. In this setting, the windspeed-exposed model over-assigns high-affected levels, flagging both the reported ADM1 and neighboring ADM1 regions as highly affected. In contrast, the 2-stage XGBoost model correctly assigns high affected levels to the reported ADM1 and lower affected levels to surrounding regions (Fig. 9b). This case highlights the two-stage model's ability to localize high-affected levels under prolonged wind exposure, avoiding spatial over-extension relative to wind-only triggering.

Mozambique, Tropical Cyclone Kenneth (2019): intense winds at landfall with limited rainfall (windspeed-exposed outperforms two-stage). Cyclone Kenneth made landfall in Cabo Delgado as a Category-4 storm, with strong winds (up to 64 m/s). Rainfall associated with Kenneth in northern Mozambique was extreme, with over 500 mm accumulated in several locations following landfall, and local recordings exceeding 570 mm. Reported severe impacts were tightly concentrated in a small

[2]NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT HURRICANE MATTHEW (AL142016)

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026



(a) Saint Lucia, Tropical Storm Matthew (2016)

(b) Mexico, Hurricane Wilma (2005)

(c) Mozambique, Cyclone Kenneth (2019)

(d) Philippines, Typhoon Krosa/Vinta (2013)

Figure 9. Case studies comparing reported impact (left), the windspeed–exposed baseline (center) and the two–stage model (right) for four countries: (a) Saint Lucia, Tropical Storm Matthew (2016); (b) Mexico, Hurricane Wilma (2005); (c) Mozambique, Cyclone Kenneth (2019); (d) Philippines, Typhoon Krosa/Vinta (2013).

number of coastal municipalities, most notably Ibo, Quissanga, and Macomia. In this wind-dominated event, the windspeed-
530 exposed model correctly concentrates high impact on the reported region, exceeding the 15% threshold only where severe

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

damage was observed. By contrast, the two-stage model produces a broader response, assigning low impact levels to neighboring ADM1 regions as well as to reported affected regions (Fig. 9c). This behavior reflects a known limitation of trained, multi-factor models under intense, compact wind cores, and is consistent with evidence that wind-threshold baselines perform well for sharply localized, high-intensity landfalls (Sedhain et al., 2025).

Philippines, Typhoon Krosa / Vinta (2013): wind-and-rain-dominated impacts (windspeed-exposed outperforms two-stage). Typhoon Krosa made landfall over northeastern Luzon with strong winds (up to 41.6 m/s) and high accumulated rainfall (up to 300 mm). Reported impact, including widespread house damage and power outages, was concentrated in a small number of ADM1 regions close to the storm track. In this setting, the windspeed-exposed model more precisely localizes high impacting regions, exceeding the 15% threshold only in the ADM1 regions reported as affected (Fig. 9d). The two-stage model also assigns high impact to the reported ADM1 regions, but spreads lower impact levels to neighboring areas, resulting in weaker precision. This behavior reflects, again, the strength of simple wind-based triggers for sharply localized damage footprints (Sedhain et al., 2025).

Across the case studies, two consistent patterns emerge: (i) Wind-only exposure performs very well when impacts are dominated by peak winds, but performs poorly when non-wind drivers —such as rainfall— play a major role. (ii) The 2-stage XGBoost model prioritizes recall at the any-impact threshold (0%), typically identifying a broader area with low affected levels. As a result, it more often matches the reported spatial extent of affected regions, but can appear weaker under a strict high-impact threshold (15%), where only the most severely affected regions are counted. Operationally, this points to a complementary use of the two approaches: the two-stage model can be used to detect and spatially localize affected regions at low thresholds, while the windspeed-exposed trigger can serve as a conservative indicator of severe impacts when wind intensity is the dominant driver.

5 Discussion

This study evaluates whether open, globally available gridded predictors can support sub-national impact-based forecasting for tropical cyclones, and how a global data-driven model compares with commonly used wind-exposure rules. The discussion focuses on two decision-relevant questions: *(i)* can we detect *any* impact worth acting on, and *(ii)* where are *high-impact* events most likely to occur?

Wind exposure rules remain widely used due to their transparency and simplicity. Our results confirm their strengths for identifying high-impact events (15% threshold), but also their limitations in capturing rain-dominated or compound impacts driven by terrain, settlement patterns, and interacting hazards. The two-stage XGBoost model improves detection of any impact and spatial localization of affected regions, but performs less well for high-impact identification. These findings align with prior work highlighting trade-offs between detection, false alarms, and lead time (Sedhain et al., 2025), and with evidence that hazard-only proxies provide conservative but incomplete representations of impact (e.g., Kam et al. (2024); McDermott (2022)). While global predictors enable transferability, they remain insufficient to achieve both high recall and high precision.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

From an operational perspective, these results highlight complementary strengths: the machine learning model improves coverage and localization, while exposure-based rules provide more reliable signals for high-impact conditions. This supports a sequential hybrid workflow, where data-driven models are used for early detection (0% threshold) and exposure-based thresholds guide escalation decisions (15%). The difficulty in predicting high-impact events can be attributed to several structural factors. High-impact cases are rare, leading to strong class imbalance. In addition, impact magnitude is driven by exposure and vulnerability rather than hazard intensity alone, and available predictors only partially capture these processes. Finally, impact data are subject to reporting noise and inconsistencies across regions. Together, these factors limit the ability of statistical models to learn robust patterns for extreme events and explain the continued competitiveness of exposure-based approaches. When applied to ensemble forecasts, the model enables uncertainty-aware impact prediction. Median–IQR summaries provide stable lead-time behaviour and better represent skewed impact distributions than mean-based approaches, supporting more robust anticipatory decision-making Leutbecher and Palmer (2008); Palmer (2019); Sedhain et al. (2025).

Case studies highlight spatial differences between approaches. The trained model distributes impacts more broadly and aligns better with reported impacts in mixed hazard events, while exposure-based rules concentrate impacts along the wind core. This suggests that contextual predictors improve spatial realism, although operational use still requires rapid verification. Finally, restricting training to individual basins does not yield consistent improvements, suggesting that the global model captures transferable patterns. However, this result should be interpreted cautiously, as differences in training data and evaluation subsets are not fully controlled. Performance limitations are primarily driven by the scarcity of high-impact events and by constraints in the underlying data.

Several limitations affect the current framework. First, EM-DAT impact data are heterogeneous in definition, spatial resolution, and reporting consistency, introducing noise and potential bias. Predictors such as `prev_events_5years` may therefore reflect reporting capacity as well as physical susceptibility. Second, impact severity is approximated through affected-population fractions and a fixed threshold (15%), which simplifies a multidimensional concept of impact. The rarity of high-impact events further limits the ability of statistical models to learn robust patterns. Third, the use of globally available predictors supports transferability but only approximates local vulnerability conditions. The absence of high-resolution socio-economic indicators constrains the model's ability to capture drivers of severe impacts. Finally, while evaluation with hindcast ensembles introduces realistic forecast uncertainty, the model is trained on reanalysis-based hazards. This mismatch may limit performance under real-time operational conditions. Addressing these limitations will require improved global impact datasets, better representation of vulnerability, and training on forecast-derived predictors.

6 Conclusions

This study evaluates whether globally available, open gridded predictors can support sub-national impact-based forecasting for tropical cyclones. Using a two-stage XGBoost model trained on 780 historical events, we benchmark data-driven predictions against commonly used wind-exposure rules within a unified evaluation framework.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Two main findings emerge. First, the machine learning model improves the detection and spatial localization of impacts, particularly at the 0% threshold, making it well suited for early-stage situational awareness and prioritization. Second, for identifying high-impact events (15% threshold), simpler exposure-based rules remain more effective, providing conservative and interpretable signals for severity. These results highlight a fundamental trade-off between detection and severity estimation and suggest that hybrid approaches are better aligned with operational needs. In practice, machine learning models can be used to identify potentially affected areas, while exposure-based thresholds can support escalation decisions for high-impact conditions. Overall, this work establishes a first-generation global benchmark for sub-national impact-based forecasting using open data. While the framework demonstrates the feasibility of transferable predictions across regions, performance remains constrained by limitations in impact data, vulnerability representation, and the scarcity of extreme events.

Future work should focus on training models directly on forecast-derived predictors to better reflect operational conditions, and on integrating improved representations of socio-economic vulnerability. Advances in global impact datasets and uncertainty-aware modeling will be critical for moving toward reliable, operational impact-based forecasting systems.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

References

Arachchige, S. M., Pradhan, B., and Park, H.-J.: A critical review of hurricane risk assessment models and predictive frameworks, Geoscience Frontiers, p. 102012, 2025.

Aznar-Siguan, G. and Bresch, D. N.: CLIMADA v1: A Global Weather and Climate Risk Assessment Platform, Geoscientific Model Development, 12, 3085–3097, https://doi.org/10.5194/gmd-12-3085-2019, 2019.

Baugh, C., Colonese, J., D'Angelo, C., Dottori, F., Neal, J., Prudhomme, C., and Salamon, P.: Global River Flood Hazard Maps, http://data.europa.eu/89h/jrc-floods-floodmapgl_rp50y-tif, 2024.

Bresch, D. N. and Aznar-Siguan, G.: CLIMADA v1.4.1: Towards a Globally Consistent Adaptation Options Appraisal Tool, Geoscientific Model Development, 14, 351–363, https://doi.org/10.5194/gmd-14-351-2021, 2021.

Cardona, O.-D., Ordaz, M. G., Mora, M. G., Salgado-Gálvez, M. A., Bernal, G. A., Zuloaga-Romero, D., Marulanda Fraume, M. C., Yamín, L., and González, D.: Global risk assessment: A fully probabilistic seismic and tropical cyclone wind risk assessment, International Journal of Disaster Risk Reduction, 10, 461–476, https://doi.org/https://doi.org/10.1016/j.ijdrr.2014.05.006, global probabilistic assessment of risk from natural hazards for the Global Assessment Report 2013 (GAR13), 2014.

Chaves-Gonzalez, J., Milano, L., Omtzigt, D.-J., Pfister, D., Poirier, J., Pople, A., Wittig, J., and Zommers, Z.: Anticipatory action: Lessons for the future, Frontiers in Climate, 4, 932 336, 2022.

Coughlan de Perez, E., van den Hurk, B., Van Aalst, M., Jongman, B., Klose, T., and Suarez, P.: Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts, Natural Hazards and Earth System Science, 15, 895–904, 2015.

Delforge, D., Wathelet, V., Below, R., Sofia, C. L., Tonnelier, M., van Loenhout, J. A., and Speybroeck, N.: EM-DAT: the emergency events database, International Journal of Disaster Risk Reduction, p. 105509, 2025.

Dullaart, J., Muis, S., Bloemendaal, N., Chertova, M., Couasnon, A., and Aerts, J. C. J. H.: COAST-RP: A global COastal dAtaset of Storm Tide Return Periods (Version 2), Dataset, https://doi.org/10.4121/13392314.V2, 2022.

Fang, J., Sun, S., Shi, P., and Wang, J.: Assessment and mapping of potential storm surge impacts on global population and economy, International Journal of Disaster Risk Science, 5, 323–331, 2014.

Food and Agriculture Organization of the United Nations: Anticipatory Action Protocol: Typhoon and Tropical Cyclone-Induced Flooding – Viet Nam, http://www.fao.org/3/cc6785en/cc6785en.pdf, brochure, 2023.

GADM: Global Administrative Areas (Version 4.1), https://gadm.org/data, accessed: 2025-10-27, 2023.

Gahtan, J., Knapp, K. R., Schreck, C. J., Diamond, H. J., Kossin, J. P., and Kruk, M. C.: International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4r01, https://doi.org/10.25921/82ty-9e16, [Indicate subset used]. Accessed: 2025-10-26, 2024.

Holland, G.: A revised hurricane pressure–wind model, Monthly Weather Review, 136, 3432–3445, 2008.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., et al.: GPM IMERG Late Run V07 Multi-satellite Precipitation Estimate with Gauge Calibration, https://doi.org/10.5067/GPM/IMERG/V07/06H, accessed: 2025-10-27, 2023.

Jarvis, A., Reuter, H. I., Nelson, A., and Guevara, E.: Hole-filled SRTM for the Globe, Version 4, http://srtm.csi.cgiar.org, accessed: 2025-10-27, 2008.

Jing, R., Heft-Neal, S., Chavas, D. R., Griswold, M., Wang, Z., Clark-Ginsberg, A., Guha-Sapir, D., Bendavid, E., and Wagner, Z.: Global population profile of tropical cyclone exposure from 2002 to 2019, Nature, 626, 549–554, 2024.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Kam, P. M., Ciccone, F., Kropf, C. M., Riedel, L., Fairless, C., and Bresch, D. N.: Impact-based forecasting of tropical cyclone-related human displacement to support anticipatory action, Nature Communications, 15, 8795, 2024.

Kim, J.-M., Son, K., and Kim, Y.-J.: Assessing regional typhoon risk of disaster management by clustering typhoon paths, Environment, Development and Sustainability, 21, 2083–2096, 2019.

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J.: The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone best track data, Bulletin of the American Meteorological Society, 91, 363–376, https://doi.org/10.1175/2009BAMS2755.1, 2010.

Kooshki Forooshani, M., van den Homberg, M., Kalimeri, K., Kaltenbrunner, A., Mejova, Y., Milano, L., Ndirangu, P., Paolotti, D., Teklesadik, A., and Turner, M. L.: Towards a global impact-based forecasting model for tropical cyclones, Natural Hazards and Earth System Sciences, 24, 309–329, 2024.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, Journal of computational physics, 227, 3515–3539, 2008.

Lin, P. and Wang, N.: A data-driven approach for regional-scale fine-resolution disaster impact prediction under tropical cyclones, Natural Hazards, 120, 7461–7479, 2024.

Lyu, H.-M. and Yin, Z.-Y.: Flood susceptibility prediction using tree-based machine learning models in the GBA, Sustainable Cities and Society, 97, 104 744, 2023.

Mandal, P., Maiti, A., Paul, S., Bhattacharya, S., and Paul, S.: Mapping the multi-hazards risk index for coastal block of Sundarban, India using AHP and machine learning algorithms., Tropical Cyclone Research & Review, 11, 2022.

Marconi, M., Gatto, B., Magni, M., and Marincioni, F.: A rapid method for flood susceptibility mapping in two districts of Phatthalung Province (Thailand): present and projected conditions for 2050, Natural Hazards, 81, 329–346, 2016.

McDermott, T. K.: Global exposure to flood risk and poverty, Nature Communications, 13, 3529, 2022.

Meng, C., Xu, W., Su, P., Qin, L., Liao, X., and Zhang, J.: Quantitative assessment of population risk to tropical cyclones using hybrid modeling combining GAM and XGBoost: A case study of Hainan Province, International Journal of Disaster Risk Reduction, 110, 104 650, 2024.

NASA Shuttle Radar Topography Mission (SRTM): Shuttle Radar Topography Mission (SRTM) Global, https://doi.org/10.5069/G9445JDF, survey date: 2000-02-11 to 2000-02-22. Accessed: 2025-10-27, 2013.

National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, Japan Meteorological Agency, Met Office, Ministry of Defence, United Kingdom, China Meteorological Administration, Meteorological Service of Canada, Environment Canada, Korea Meteorological Administration, Meteo-France, European Centre for Medium-Range Weather Forecasts, and Bureau of Meteorology, Australia: THORPEX Interactive Grand Global Ensemble (TIGGE) Model Tropical Cyclone Track Data, https://doi.org/10.5065/D6GH9GSZ, accessed: 6 Oct 2025, 2008.

National Disaster Risk Reduction and Management Council: NDRRMC No. 60 Guidelines Declaration of State of Calamity, 2019.

National Oceanic and Atmospheric Administration, Atlantic Oceanographic and Meteorological Laboratory: Seven Tropical Cyclone Basins, https://www.aoml.noaa.gov/phod/cyclone/seven.php, accessed: 2025-10-26, n.d.

Palmer, T.: The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years, Quarterly Journal of the Royal Meteorological Society, 145, 12–24, 2019.

Parks, R. M., Kontis, V., Anderson, G. B., Baldwin, J. W., Danaei, G., Toumi, R., Dominici, F., Ezzati, M., and Kioumourtzoglou, M.-A.: Short-term excess mortality following tropical cyclones in the United States, Science advances, 9, eadg6633, 2023.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Peduzzi, P., Dao, H., Herold, C., and Mouton, F.: Assessing global exposure and vulnerability towards natural hazards: the Disaster Risk Index, Natural Hazards and Earth System Sciences, 9, 1149–1159, https://doi.org/10.5194/nhess-9-1149-2009, 2009.

Peduzzi, P., Chatenoux, B., Dao, H., De Bono, A., Herold, C., Kossin, J., Mouton, F., and Nordbeck, O.: Global trends in tropical cyclone risk, Nature Climate Change, 2, 289–294, 2012.

Pesaresi, M., Schiavina, M., Melchiorri, M., and et al.: Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data, International Journal of Digital Earth, 17, https://doi.org/10.1080/17538947.2023.2258627, accessed: 2025-10-27, 2024.

Sedhain, S., van den Homberg, M., Teklesadik, A., van Aalst, M., and Kerle, N.: Evaluating impact-based forecasting models for tropical cyclone anticipatory action, International Journal of Disaster Risk Reduction, 129, 105 782, https://doi.org/https://doi.org/10.1016/j.ijdrr.2025.105782, 2025.

Sharma, P., Wang, J., Zhang, M., Woods, C., Kar, B., Bausch, D., Chen, Z., Tiampo, K., Glasscoe, M., Schumann, G., Pierce, M., and Eguchi, R.: DisasterAWARE – A GLOBAL ALERTING PLATFORM FOR FLOOD EVENTS, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, VI-3/W1-2020, 107–113, https://doi.org/10.5194/isprs-annals-VI-3-W1-2020-107-2020, 2020.

Simpson, R. H. and Saffir, H.: The Hurricane Disaster-Potential Scale, Weatherwise, 27, 169–186, 1974.

Smits, J. and Permanyer, I.: The Subnational Human Development Database, Scientific Data, 6, 190 038, https://doi.org/10.1038/sdata.2019.38, 2019.

United Nations Office for the Coordination of Humanitarian Affairs: Fiji Tropical Cyclones – 2023 Anticipatory Action Framework, https://www.unocha.org/publications/report/fiji/fiji-tropical-cyclones-2023-anticipatory-action-framework, 2023.

Wagenaar, D., Hermawan, T., van den Homberg, M. J., Aerts, J. C., Kreibich, H., de Moel, H., and Bouwer, L. M.: Improved transferability of data-driven damage models through sample selection bias correction, Risk analysis, 41, 37–55, 2021.

World Bank: Global Landslide Hazard Map, https://datacatalog.worldbank.org/search/dataset/0037584/global-landslide-hazard-map, accessed: 2025-10-27. License: Creative Commons Attribution-Non Commercial 4.0, 2022.

World Food Programme, U.: Advanced Disaster Analysis and Mapping (ADAM), https://gis.wfp.org/adam/, 2016.

WorldPop: WorldPop 2020 UN-adjusted Population Dataset, https://www.worldpop.org, accessed: 2025-10-27, 2020.

Code availability. The source code for this study is hosted on GitHub at this repository. The repository provides a complete pipeline for (i) generating the global impact database, (ii) model training, and (iii) performance testing. Detailed documentation and step-by-step instructions for reproducing the results are included in the repository's README file.

Data availability. All input datasets used in this study are publicly available. Tropical cyclone track and intensity data were obtained from the International Best Track Archive for Climate Stewardship (IBTrACS). Precipitation data were obtained from NASA GPM IMERG Late Run V07. Topographic data were derived from the SRTM Digital Elevation Model. Administrative boundary and coastline data were obtained from the Global Administrative Areas database (GADM v4.1). Settlement morphology data were obtained from the Global Human Settlement Layer Degree of Urbanisation 2025 epoch. Population estimates were obtained from the WorldPop 2020 UN-adjusted dataset.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Storm tide data were obtained from the COAST-RP v2 dataset. Landslide hazard data were obtained from the World Bank Global Landslide Hazard Map. Disaster impact records were obtained from the EM-DAT International Disaster Database (registration required). Historical ensemble forecast tracks were obtained from the TIGGE archive. The harmonized global impact dataset assembled for this study, combining the above sources at 0.1° resolution for 780 tropical cyclone events across 72 countries, is released openly alongside the model code (see Code Availability).

Author contributions. FM: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing. KK, LM, and MvdH: Conceptualization, Writing – review & editing. TD and PN: Investigation, Writing – review & editing. YM and AK: Methodology, Writing – review & editing. All authors reviewed and approved the final manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. F.M., K.K., and Y.M. acknowledge financial support from the Lagrange Project of the Institute for Scientific Interchange Foundation (ISI Foundation) funded by Fondazione Cassa di Risparmio di Torino (Fondazione CRT). K.K. thanks AECID (Spanish Agency for International Development Cooperation) for their support to data innovation and Frontier Data Technologies through UNICEF's Frontier Data Network. M.vd.H. was supported by the Princess Margriet Fund and the Forecast-based financing project with the Philippine Red Cross, funded by the German Red Cross.

Appendix A: Feature Table

Table A1 provides a comprehensive summary of the variables integrated into the global 0.1° modeling framework, cataloging the physical hazards, socioeconomic predictors, and the target variable alongside their descriptions, native resolutions, and primary data sources.

Appendix B: Validation Metrics

As described in the Section 3.3, we consider distance-based, categorical and binary metrics for model evaluation. While mean absolute error (MAE) and root-mean absolute error (RMSE) are common distance based metrics, we introduce the definitions of the binary-based ones which rely on standard formulations in Table B1. Also, we briefly elaborate on the Quadratic Weighted Kappa (QWK) categorical metric due to its less common use in this context.

The QWK is a measure of agreement between two ordinal ratings that penalizes disagreements based on their severity. Unlike simple accuracy or unweighted kappa, it assigns higher penalties to larger disagreements. This is particularly useful when predicting discrete categories that have an inherent order (e.g., impact severity levels).

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Table A1. Enumeration and description of the globally available features used in the people affected prediction model.

	Feature Label	Description	Type	Resolution	Source
1	WEA_wind_speed	Max. 1-min. sustained windspeed (m/s)	Weather	-	IBTrACS
2	WEA_rainfall_max_24h	Max rainfall within a 24 hour period (mm)	Weather	30min, 0.1°	NASA PPS
3	TOP_mean_slope	Mean slope (degree)	Topography	90m	SRTM
4	TOP_mean_elevation_m	Mean Elevation (m)	Topography	90m	SRTM
5	TOP_with_coast	Boolean: coast or no coast	Topography	90m	SRTM
6	TOP_coast_length	Length of coast (Km)	Topography	-	GADM
7	TOP_mean_rug	Ruggedness of terrain (TRI index)	Topography	90m	SRTM
8	SH_landslides_risk	Rainfall-induced-Landslides Risk index	Secondary Hazards	100m	WBG & GFDRR
9	SH_storm_tides	Storm-Tides Level (m)	Secondary Hazards	-	COAST-RP
10	URB_urban	Proportion of urban areas	Urbanization	1Km	GHSL
11	URB_rural	Proportion of rural areas	Urbanization	1Km	GHSL
12	URB_water	Proportion of areas classified as water	Urbanization	1Km	GHSL
13	total_pop	Total number of people	Demographic	100m	WorldPop
14	prev_events_5years	Number of reported impacting events in the past 5 years	Previous events	-	EM-DAT
15	perc_pop_affected	% of people affected	Target variable	ADM0&1 regions	EM-DAT

It is defined as:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{B1}$$

where:

- O_{ij} is the observed frequency matrix between actual and predicted ratings.
- E_{ij} is the expected frequency matrix assuming ratings are independent (random matrix following original distribution)
- w_{ij} is the quadratic weight matrix, defined as $w_{ij} = \frac{(i-j)^2}{(k-1)^2}$ where i and j are the rating categories and k is the total number of categories.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Table B1. Overview of the binary metrics used for validation.

Metric	Formula
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1 score	$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$
Specificity	$\frac{TN}{TN+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$
False Negative Rate (FNR)	$\frac{FN}{FN+TP}$
Critical Success Index (CSI)	$\frac{TP}{TP+FP+FN}$

The QWK score ranges from -1 (complete disagreement) to 1 (perfect agreement), where 0 indicates random agreement. A higher QWK indicates that the predicted categories align well with the true categories, taking into account the severity of mismatches.

Now, in B2, we report binary-based metrics for all the models considered in our analysis.

Appendix C: Performance by basin

Table C1 compares the performance of the model under two distinct evaluation methods: the baseline LOOCV-global specification and a geographically constrained approach (LOOCV-geo-constrained). In the constrained evaluation, each leave-one-event-out fold is trained exclusively on data from the same cyclone basin as the test event. This side-by-side comparison allows us to assess the trade-offs of relying strictly on region-specific historical patterns versus leveraging broader cross-basin information. Across both evaluation methods, we report mean absolute error (MAE) alongside classification metrics (precision, recall, and F1 score) evaluated at impact thresholds of 0% and 15%.

At the 0% threshold (first stage), performance remains relatively consistent across most basins when restricting training to basin-specific data. Recall is generally strong, indicating the model can still effectively detect impacting areas without cross-basin information. However, a notable exception occurs in data-sparse areas such as the Australian Region. In this basin, the geo-constrained specification yields a higher mean absolute error (MAE) and a sharp decline in both precision and overall F1

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Table B2. Binary classification performance of our four impact models at the Administrative 1 level, evaluated at thresholds 0% and 15%. Metrics include precision, recall, specificity, accuracy, false positive rate (FPR), false negative rate (FNR), F1-score, and critical success index (CSI).

Model	Threshold (%)	Precision	Recall	Specificity	Accuracy	FPR	FNR	F1	CSI
Historical	0	0.14	**0.95**	0.03	0.17	0.97	**0.05**	0.25	0.14
Windspeed-exposed	0	**0.57**	0.17	**0.98**	**0.86**	**0.02**	0.83	0.27	0.15
Windspeed-historical	0	0.56	0.15	**0.98**	**0.86**	**0.02**	0.85	0.24	0.14
2-stg-XGBoost	0	0.30	0.68	0.73	0.73	0.27	0.32	**0.42**	**0.26**
Historical	15	0.04	0.05	0.97	0.95	0.03	0.95	0.04	0.02
Windspeed-exposed	15	**0.22**	0.44	0.97	0.96	0.03	0.56	**0.30**	**0.17**
Windspeed-historical	15	0.06	0.01	**1.00**	**0.98**	**0.00**	0.99	0.01	0.01
2-stg-XGBoost	15	0.15	**0.50**	0.94	0.93	0.063	**0.50**	0.23	0.13

score compared to the global baseline. Outside of this extreme, precision varies across the other basins, but their relatively stable F1 scores suggest a generally balanced classification performance during the impact-detection stage.

Greater heterogeneity emerges in the second stage (15% threshold). The North Atlantic basin, which benefits from a relatively large training sample (5,758 event–ADM1 observations), maintains the strongest performance under the geo-constrained specification, particularly in recall. In contrast, isolating the training data severely limits the model in regions with the smallest sample sizes; both the Australian Region (154 observations) and the South-West Indian basin (515) fail to correctly identify any higher-impact cases (recall of 0.00). Interestingly, while the Western Pacific boasts the largest sample size (21,743), its recall drops significantly in the constrained setup. Meanwhile, the South Pacific and North Indian basins maintain moderate recall despite much smaller training volumes. This pattern suggests that while the second stage—which estimates impact magnitude conditional on occurrence—is highly sensitive to extreme data scarcity, factors beyond mere training volume, such as regional event homogeneity or feature informativeness, are critical to performance.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Table C1. Model performance by cyclone basin for the LOOCV-global and LOOCV-geo-constrained evaluation methods. Metrics include MAE and classification metrics (precision, recall, and F1 score) evaluated at impact thresholds of 0% and 15%.

Threshold (%)	Cyclone basin	LOOCV-global				LOOCV-geo-constrained			
		MAE	F1	Precision	Recall	MAE	F1	Precision	Recall
0	Australian Region	0.86	0.62	0.54	0.72	0.99	0.33	0.20	0.94
	North Atlantic	4.75	0.47	0.34	0.80	4.80	0.48	0.34	0.80
	North Indian	2.58	0.44	0.32	0.70	1.49	0.45	0.34	0.67
	South Pacific	9.06	0.52	0.39	0.81	5.82	0.56	0.52	0.61
	South-West Indian	3.50	0.59	0.46	0.83	1.73	0.60	0.69	0.52
	Western Pacific	3.40	0.38	0.28	0.63	3.13	0.38	0.27	0.65
15	Australian Region	0.86	0.00	–	0.00	0.99	0.00	–	0.00
	North Atlantic	4.75	0.31	0.21	0.59	4.80	0.38	0.26	0.69
	North Indian	2.58	0.28	0.21	0.41	1.49	0.37	0.44	0.32
	South Pacific	9.06	0.38	0.25	0.72	5.82	0.36	0.40	0.33
	South-West Indian	3.50	0.13	0.09	0.22	1.73	0.00	–	0.00
	Western Pacific	3.40	0.19	0.12	0.45	3.13	0.22	0.29	0.18

Appendix D: Performance by aggregation resolution

To assess whether the grid-based formulation of the model (under the assumption of homogeneous impact disaggregation) is appropriate, we compare the previously defined grid-based two-stage XGBoost model with an ADM1-level counterpart. In the ADM1 version, all features are aggregated to Administrative Level 1 units while preserving the same modeling architecture.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Under this specification, the homogeneous impact allocation assumption is applied only when multiple ADM1 regions are assigned a single reported impact value. In cases where only one ADM1 region is affected, the reported impact value remains unchanged, and no disaggregation is performed.

The training set of the ADM1-level model consists of 29427 provinces. Performance metrics for the ADM1 model, alongside those of the grid-based version (for comparison purposes), are reported in Tables D1 and D2.

Table D1. Binary classification performance of our 2-stage-XGBoost models (grid-based and adm1-based) at the Administrative 1 level, evaluated at thresholds 0% and 15%. Metrics include precision, recall, specificity, accuracy, false positive rate (FPR), false negative rate (FNR), F1-score, and critical success index (CSI).

Model	Threshold (%)	Precision	Recall	Specificity	Accuracy	FPR	FNR	F1	CSI
2-stg-XGBoost (grid-based)	0	0.30	0.68	0.73	0.73	0.27	0.32	0.42	0.26
2-stg-XGBoost (adm1-based)	0	0.67	0.14	0.99	0.87	0.01	0.86	0.23	0.13
2-stg-XGBoost (grid-based)	15	0.15	0.50	0.94	0.93	0.06	0.50	0.23	0.13
2-stg-XGBoost (adm1-based)	15	0.34	0.32	0.99	0.97	0.01	0.68	0.33	0.20

Table D2. Categorical and distance-based performance of our 2-stage-XGBoost models (grid-based and adm1-based) at the Administrative 1 level. Metrics include mean absolute error (MAE), root mean squared error (RMSE), and quadratic weighted kappa (QWK) computed on categorized impact levels (defined with the already introduced impact categories).

Model	MAE	RMSE	QWK
2-stg-XGBoost (grid-based)	3.67	9.52	0.29
2-stg-XGBoost (adm1-based)	1.53	8.41	0.28

The comparison between the grid-based and ADM1-based specifications reveals a clear trade-off between precision and recall across both stages. At both thresholds (0% and 15%), the ADM1-based model consistently achieves substantially higher precision but markedly lower recall than the grid-based model. This indicates that the ADM1 specification is more conservative: it is more effective at correctly identifying non-impacting regions (fewer false positives), but at the cost of failing to detect a large share of truly impacted regions (more false negatives).

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

While all metrics are informative, we place greater emphasis on recall, given our focus on correctly identifying regions that are genuinely impacted. From this perspective, the grid-based model performs more favorably in both stages, as it consistently captures a substantially larger proportion of true impacts. The higher spatial resolution of the grid-based features therefore appears to enhance the model's robustness in detecting affected locations.

In terms of distance-based metrics, however, the ADM1 model achieves lower MAE and RMSE. This improvement is largely driven by its strong performance in predicting the majority class (no impact). Because non-impacting cases dominate the dataset, a model that more frequently assigns no impact will naturally achieve better average error metrics. Consequently, while the ADM1 specification performs better under aggregate distance-based measures, this comes at the expense of reduced sensitivity to actual impacted regions.

Appendix E: SHAP Analysis

To better understand the internal decision-making of the two-stage XGBoost framework, we analyze feature contributions using SHAP (SHapley Additive exPlanations) values. SHAP provides a consistent and theoretically grounded measure of each feature's marginal contribution to model predictions, allowing us to interpret both the direction and magnitude of feature effects. Figure E1 presents SHAP summary plots for the two components of our modeling pipeline. The left panel corresponds to the first-stage classifier, which predicts the occurrence of impact (impact vs. no impact), while the right panel shows the second-stage classifier, which estimates impact magnitude conditional on impact occurrence. Together, these plots reveal which predictors drive the probability of impact onset and which variables influence the intensity of impacts once they occur, thereby offering insight into the distinct mechanisms captured at each stage of the model.

As expected, wind speed and rainfall contribute positively to the model's prediction of impact, with higher values increasing the likelihood that an event is classified as impactful. Interestingly, the feature capturing historically impacting events at the country–ADM1 level exhibits different behavior across the two modeling stages.

In the first stage, higher values of this feature contribute negatively to the model output, implying that the predicted probability of impact is greater in historically less affected regions. This suggests that the classifier assigns relatively higher risk to areas with lower historical impact frequency.

In the second stage, however, the feature shows consistently high contribution magnitudes regardless of whether its values are high or low. This indicates that historical exposure remains an influential predictor for estimating impact magnitude once impact occurrence is established, even though its directional effect varies across instances.

Since `prev_events_5years` has implicit information of impacting events that occurred in the country in the previous years, and given the importance we see it has by looking at the SHAP summary plots of Figure E1, we demonstrate that no data leakage is happening at the core of our model. In order to do this, we study performance comparison in the 2-stage-XGBoost models after re-training without the `prev_events_5years` feature. Performance comparison metrics can be seen in the Tables E1 and E2.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

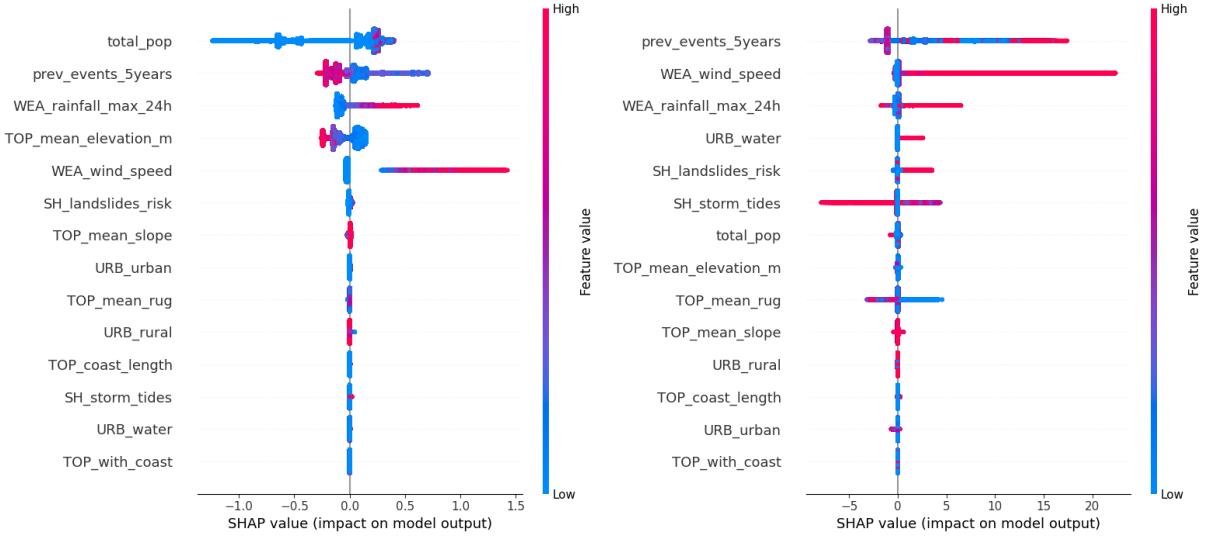


Figure E1. SHAP summary plot for the first-stage classifier on the left (impact vs. no impact), and on the right second-stage classifier (impact magnitude conditional on impact occurrence) of the two-stage XGBoost model.

Table E1 compares the performance of the two-stage XGBoost model with and without the `prev_events_5years` feature. At the 0% threshold, the inclusion of this variable has minimal effect on overall performance, with only small changes in precision and recall and negligible differences in F1 and CSI. At the 15% threshold, however, removing the feature improves precision, specificity, and overall skill (F1 and CSI), although at the cost of reduced recall. This suggests that the `prev_events_5years` variable provides limited additional predictive value and may introduce noise or redundancy in the context of high-impact event detection.

Appendix F: Historical Forecasts Event Set

This section details the 26 tropical cyclone events analyzed within the historical forecast study. These cases (2012–2020) were identified by matching EM-DAT impact records and IBTrACS best-track data with corresponding TIGGE/KMS forecast tracks. Table F1 lists the storm name, country, and date for each matched event.

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Table E1. Binary classification performance of our 2-stage-XGBoost models (with and without the `prev_events_5years` feature) at the Administrative 1 level, evaluated at thresholds 0% and 15%. Metrics include precision, recall, specificity, accuracy, false positive rate (FPR), false negative rate (FNR), F1-score, and critical success index (CSI).

Model	Threshold (%)	Precision	Recall	Specificity	Accuracy	FPR	FNR	F1	CSI
2-stg-XGBoost	0	0.30	0.68	0.73	0.73	0.27	0.32	0.42	0.26
2-stg-XGBoost (no `prev_events_5years`)	0	0.29	0.70	0.72	0.71	0.28	0.30	0.41	0.26
2-stg-XGBoost	15	0.15	0.50	0.94	0.93	0.06	0.50	0.23	0.13
2-stg-XGBoost (no `prev_events_5years`)	15	0.20	0.44	0.96	0.95	0.04	0.56	0.28	0.16

Table E2. Categorical and distance-based performance of our 2-stage-XGBoost models (with and without the `prev_events_5years` feature) at the Administrative 1 level. Metrics include mean absolute error (MAE), root mean squared error (RMSE), and quadratic weighted kappa (QWK) computed on categorized impact levels (defined with the already introduced impact categories).

Model	MAE	RMSE	QWK
2-stg-XGBoost	3.67	9.52	0.29
2-stg-XGBoost (no `prev_events_5years`)	3.78	9.42	0.31

https://doi.org/10.5194/egusphere-2026-1996
Preprint. Discussion started: 15 April 2026

Table F1. Tropical cyclone events matched for the historical forecasts study (2012–2020).

Country	Event Name	Date (YYYY-MM-DD)
Philippines	Typhoon Butchoy (Nepartak)	2016-07-08
	Typhoon Karen (Sarika)	2016-10-16
	Typhoon Helen (Megi)	2016-09-28
	Typhoon “Nina” (Nock-Ten)	2016-12-25
	Typhoon “Lan”/“Paolo”	2017-10-18
	Typhoon “Yutu” (Rosita)	2018-10-30
	TC “Nakri”	2019-11-10
	TC “Kammuri” (Tisoy)	2019-12-02
	Typhoon “Molave” (Quinta)	2020-10-28
Japan	Typhoon “Lan”/“Paolo”	2017-10-22
Taiwan	Typhoon Butchoy (Nepartak)	2016-07-09
	Typhoon Megi	2016-09-27
Vietnam	Typhoon Kai-Tak	2012-08-17
	TS “Nangka” (Nika)	2020-10-13
	Typhoon “Molave” (Quinta)	2020-10-29
China	Typhoon Haikui	2012-08-08
	Typhoon Kai-Tak	2012-08-17
	Typhoon Butchoy (Nepartak)	2016-07-09
	Typhoon Ferdie (Meranti)	2016-09-15
	Typhoon Megi	2016-09-28
	Typhoon “Maria” (Gardo)	2018-07-10
	TS Yagi	2018-08-12
	TC “Hagupit” (Dindo)	2020-08-05
South Korea	TC “Mitag”	2019-10-02
	Typhoon “Maisak” (Julian)	2020-09-03
North Korea	TC “Lingling”	2019-09-06