Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Distribution-free flash-flood human-impact forecasting under temporal covariate shift

Himadri Sen Gupta

himadri.gupta@csupueblo.edu

Colorado State University - Pueblo https://orcid.org/0000-0003-0765-8895

Jorge A. Duarte

University of Oklahoma Norman Campus: The University of Oklahoma

Saptadeep Biswas

GITAM - Visakhapatnam Campus: GITAM

Research Article

Keywords: conformal prediction, covariate shift, flash-floods, group-conditional coverage, temporal validation, rare events

Posted Date: March 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8194686/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Distribution-free flash-flood human-impact forecasting under temporal covariate shift

Himadri Sen Gupta[1,2*], Jorge A. Duarte[3] and Saptadeep Biswas[4,5]

[1*]School of Engineering, Colorado State University Pueblo, 2200 Bonforte Blvd, Pueblo, 81008, CO, USA.
[2]Data Science Research Institute, Colorado State University, 200 W. Lake Street, Fort Collins, 80523, CO, USA.
[3]Cooperative Institute for Severe and High-Impact Weather Research & Operations, University of Oklahoma, 120 David L. Boren Blvd., Norman, 73072, OK, USA.
[4]Department of Operations and Supply Chain, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, 530045, AP, India.
[5]Department of Mathematics, National Institute of Technology Agartala, ITI Road, Agartala, 799046, Tripura, India.

*Corresponding author(s). E-mail(s): himadri.gupta@csupueblo.edu;
Contributing authors: jduarte@ou.edu; saptadeepbiswas7@gmail.com;

Abstract

Urban flash floods are rare but high-consequence events, so warning systems greatly benefit from calibrated probabilities, distribution-free coverage guarantees, and explicit subgroup auditing, which favor their forecasting. We present an importance-weighted, Mondrian (group-conditional) conformal prediction framework with hierarchical shrinkage and validation-tuned selective abstention to address temporal covariate shift and sparse positives within groups. Using 6,137 Texas events (2005–2019) at census-tract resolution with a strict 2018–2019 holdout (prevalence 1.86%), a calibrated three-model ensemble (Brier-optimal weights {0.420, 0.540, 0.040}) attains AUROC 0.758 (95% CI: 0.545–0.931), AUPRC 0.198 (95% CI: 0.044–0.504), Brier 0.0187, and ECE 0.012. We quantify temporal covariate shift using the Population Stability Index (PSI) on predictor distributions; predictors such as mean elevation (PSI=0.248) and event-accumulated precipitation (PSI=0.194) illustrate the shift. Importance weighting corrects

measurable temporal shift and reduces the false-positive rate from 0.484 (naive, unweighted thresholds) to 0.138 at comparable recall. Due to sparse positives per stratum, adaptive grouping merges density quartiles; validation selects $\boldsymbol{m}^* = \mathbf{0}$, and the observed worst-group Wilson lower-bound recall is 0.213. Prequential backtesting indicates moderate temporal robustness of discrimination. Power calculations suggest that roughly 50–100 additional harmful events, distributed across strata, are required to certify 80–90% worst-group recall at 95% confidence. The framework provides deployable, distribution-free uncertainty quantification for rare events with transparent equity auditing and explicit evidence thresholds for operational urban CPS.

Keywords: conformal prediction; covariate shift; flash-floods; group-conditional coverage; temporal validation; rare events

1 Introduction and Background

Flash–floods are among the most lethal and operationally challenging urban hazards, responsible for a disproportionate share of flood fatalities despite representing only a subset of flood events (Ahmadalipour and Moradkhani 2019; Terti et al. 2017; Ashley and Ashley 2008). Their rapid onset (often within six hours), high velocities, and narrow decision windows strain emergency response and critical–infrastructure protection (CIP) systems (Terti et al. 2015; Kelsch et al. 2001). Forecasting human harm, however, is statistically difficult: harmful outcomes are rare (about 2% of documented flash–flood events), which induces severe class imbalance and yields wide uncertainty bounds precisely where incident managers require the most reliable guidance (Chang et al. 2023; Jonkman 2005; King and Zeng 2001).

Urban risk is intensifying under the converging pressures of climate change, continued urbanization, and aging infrastructure. In response, cities are deploying sensing–rich, AI–enabled cyber–physical systems (CPS) for real–time monitoring, forecasting, and targeted warning (Zhou et al. 2025,?; Kim et al. 2025). Recent technological advancements have focused on three primary frontiers. First, computer vision models, such as YOLOv8, have been successfully deployed for real-time flash

flood detection from video streams, offering rapid situational awareness (Quang et al. 2024). Second, hybrid frameworks integrating Machine Learning (ML) with Multi-Criteria Decision Analysis (MCDA) and Explainable AI (XAI) have improved the holistic assessment of hazard and vulnerability in complex basins (Ding et al. 2025; Ashraf et al. 2025). Third, ensemble algorithms (e.g., XGBoost, Random Forest) and knowledge-driven AHP models have been optimized for susceptibility mapping in data-scarce and semi-arid regions (Bammou et al. 2024; Chihi et al. 2025), while rainfall-threshold methods have been refined to better account for antecedent soil moisture and mega-flood drivers (Zhao et al. 2025; Cui et al. 2025).

Despite these advances in detection accuracy and susceptibility mapping, a central methodological gap persists. Most urban flash–flood harm models provide point predictions at coarse resolution, lacking distribution–free uncertainty quantification, explicit treatment of temporal covariate shift, and the group–conditional guarantees essential for equitable protection across heterogeneous populations (Reichstein et al. 2019; Karpatne et al. 2018; Terti et al. 2019; Zahran et al. 2008; Mohtar et al. 2020). For example, Chang et al. (2023) advanced tract–level modeling using rare–event logistic regression, but—like much prior work—did not certify distribution–free coverage under temporal distributional drift from changing precipitation regimes and urban development (i.e., a finite–sample guarantee that the true label lies in the predicted set with probability at least $1 - \alpha$ after accounting for shift) (Easterling et al. 2017; Wobus et al. 2014), nor did it provide subgroup–specific reliability guarantees (i.e., coverage/error–rate bounds that hold within each predefined stratum rather than only on average). We quantify temporal covariate shift using the Population Stability Index (PSI), which is computed on predictor distributions. In our model, representative predictors include mean elevation and event–accumulation precipitation.

Conformal prediction (CP) offers finite–sample, distribution–free coverage under exchangeability (Vovk et al. 2005; Shafer and Vovk 2008; Angelopoulos and Bates

2021). Extensions addressing distribution shift use importance weighting (IW) to approximately restore exchangeability between calibration and deployment distributions (Tibshirani et al. 2019; Barber et al. 2023). Mondrian (group–conditional) CP tailors guarantees to predefined strata, aligning with environmental–justice objectives by requiring coverage within each subgroup rather than only on average (Romano et al. 2020). Applying these methods in rare–event settings, however, exposes two fundamental challenges: (i) few positive samples per group yield unstable empirical quantiles and wide confidence bounds, and (ii) impossibility results limit simultaneous conditional guarantees when data are scarce, regardless of the learning algorithm (Lei and Wasserman 2014; Foygel Barber et al. 2021). Any deployment–oriented framework must therefore (a) correct for temporal shift, (b) regularize group thresholds to avoid overfitting in small strata, and (c) clearly state what can—and cannot—be certified from available data.

This study addresses the gap by developing and integrating an importance–weighted Mondrian CP framework, hierarchical shrinkage, and validation–tuned abstention for deployment in urban flash–flood harm prediction. Using 6,137 Texas events (2005–2019) at the census–tract resolution, compiled by Chang et al. (2023), we train a calibrated ensemble and construct IW–Mondrian thresholds stratified by population density. We employ hierarchical empirical–Bayes shrinkage to borrow strength across groups, stabilizing small–sample quantiles (Efron and Morris 1977; Morris 1983). We then introduce a selective prediction rule that sets an abstention margin on a disjoint validation set to maximize acted coverage, subject to a worst–group Wilson lower–bound (WLB) recall constraint (Wilson 1927; Agresti and Coull 1998). Temporal robustness is assessed via prequential backtesting (Dawid 1984). On a strict 2018–2019 holdout (prevalence 1.86%), the calibrated ensemble attains AUROC = 0.758 (95% CI: 0.545–0.931), AUPRC = 0.198 (95% CI: 0.044–0.504),

Brier = 0.0187, and strong calibration (ECE = 0.012). Importance weighting materially corrects for drift (elevation PSI = 0.248, precipitation PSI = 0.194), reducing the false–positive rate from 0.484 (naïve, unweighted quantiles) to 0.138 at comparable recall—an operationally significant improvement. The resulting worst–group WLB recall is 0.213, well below a 0.90 target; validation selects $m^* = 0$ (no abstention band satisfies the constraint), correctly identifying a sample–size–limited certification rather than model bias. Prequential results confirm this moderate stability of discrimination, while subgroup certification varies with the stochastic incidence of rare harmful events.

Our primary contributions are methodological: (i) a deployment–oriented IW–Mondrian CP pipeline for urban flash–flood harm that explicitly addresses temporal covariate shift (Tibshirani et al. 2019; Barber et al. 2023) and subgroup coverage (Romano et al. 2020); (ii) hierarchical shrinkage of group thresholds to regularize small–sample quantiles via empirical–Bayes pooling (Efron and Morris 1977; Morris 1983); (iii) a validation–tuned selective prediction rule that optimizes acted coverage under a worst–group WLB constraint (Wilson 1927; Agresti and Coull 1998); (iv) prequential backtesting to characterize temporal robustness in rare–event deployment (Dawid 1984); and (v) a data–driven articulation of the evidence threshold for equitable operations in safety–critical AI (Amodei et al. 2016), quantifying how many additional harmful events per stratum are required before high subgroup recall can be certified rather than asserted. Together, the framework advances distribution–free uncertainty quantification for rare, high–consequence urban hazards while making limitations and trade–offs transparent for policy and practice.

The remainder of the paper is organized as follows. Section 2 details the methodology, including temporal partitioning, calibrated ensembling, IW–MGondrian conformal construction with hierarchical shrinkage, abstention tuning, and prequential evaluation. Section 3 reports discrimination, calibration, fairness certification, shift characterization, and temporal robustness. Section 4 discusses achievements,

finite–sample limitations, and implications for CPS integration and equity auditing, and outlines directions for future research.

2 Research Methodology

This section presents a deployment–oriented protocol for predicting flash–flood human harm with distribution–free guarantees under temporal covariate shift and subgroup equity constraints (i.e., explicit group–conditional requirements that ensure, within each predefined stratum, conformal coverage $\geq 1 - \alpha$ and a minimum worst–group recall via Wilson lower bounds). These constraints are important because aggregate metrics can mask systematic underwarning for specific communities; enforcing groupwise guarantees aligns the model with equitable protection goals in public safety.

The protocol integrates: (i) probability–calibrated ensembling, (ii) importance–weighted Mondrian conformal prediction (IW–MCP), (iii) empirical–Bayes hierarchical shrinkage of group thresholds to stabilize small–sample quantiles, and (iv) validation–tuned selective abstention guided by a worst–group Wilson lower–bound constraint. A high–level schematic of this end–to–end pipeline is provided in Figure 1.

2.1 Data, Predictors, and Temporal Partitioning

The analysis uses 6,137 Texas flash–flood events (2005–2019) at census–tract resolution. The binary label *human_harm* flags at least one fatality or injury, with an overall prevalence of 2.17%. Predictors span event characteristics, geographic factors, and community infrastructure; the complete variable set and its sources are detailed in Table 1.

Strict temporal discipline is enforced to avoid data leakage and induce a realistic distribution shift for assessment. The dataset is partitioned into four blocks: training (2005–2015; n=2751, harms 58), validation (a pre–2016 stratified holdout; n=1834, harms 39), calibration (2016–2017; n=962, harms 25), and testing (2018–2019; n=590,

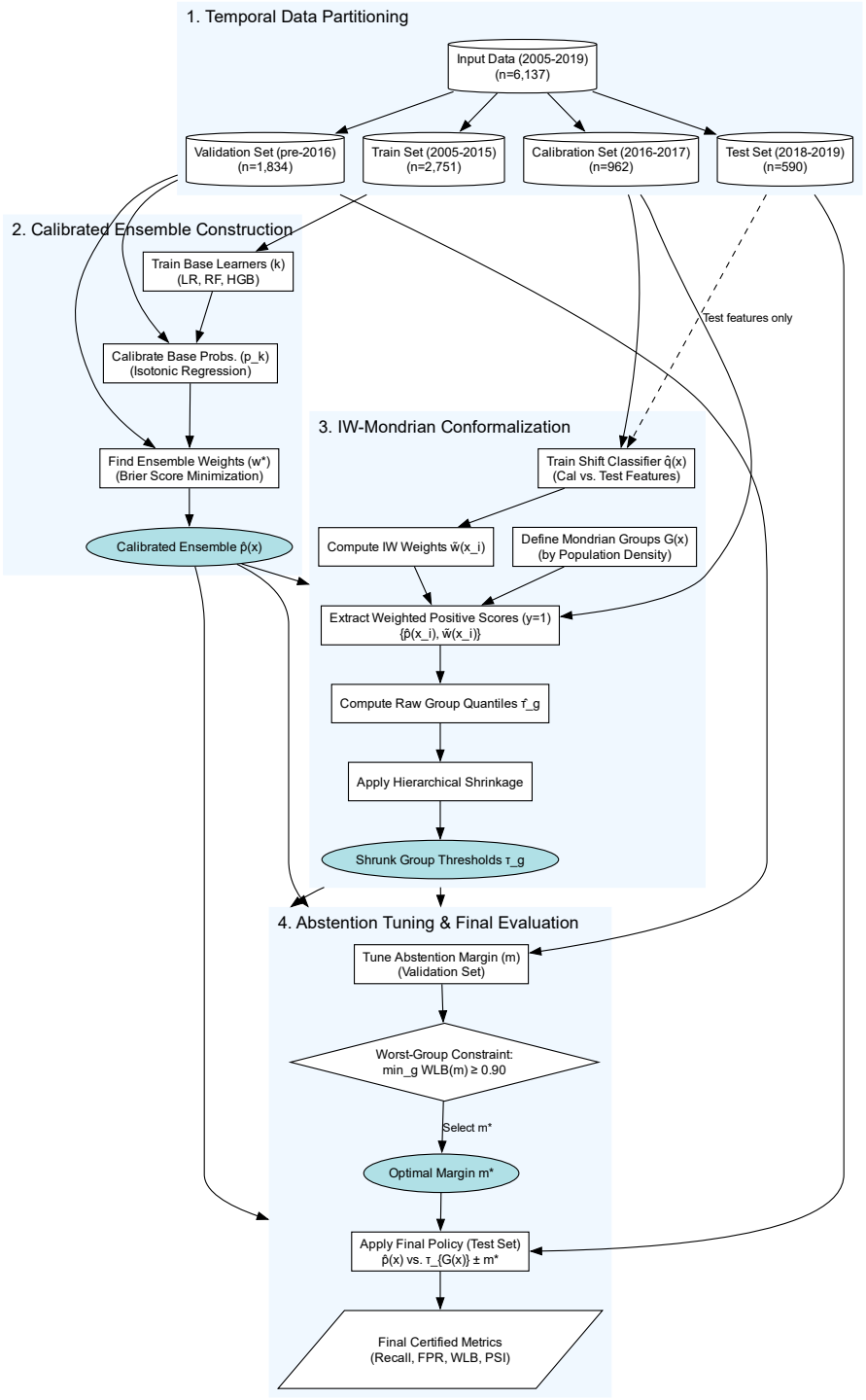


Fig. 1 End–to–end workflow. The protocol employs temporally disciplined dataset partitioning (train/validation/calibration/test), probability–calibrated machine learning model ensembling, importance–weighted Mondrian conformal prediction with hierarchical shrinkage, and validation–tuned selective abstention. The final certified prediction sets are designed as the output to an operational decision–support layer (see Section 2.6). Drift monitoring and rolling recalibration close the loop for near–real–time operation.

Table 1 Predictor variables at census–tract resolution (abbreviations: NOAA SED = Storm Events Database; PRISM = PRISM Daily Spatial Climate Dataset; NED = National Elevation Dataset; NHD = National Hydrography Dataset; STATSGO = State Soil Geographic Database; ACS = American Community Survey; NLCD = USGS National Land Cover Database; TxDOT = Texas Department of Transportation; TNRIS = Texas Natural Resources Information System).

Variable	Description	Source
Event Characteristics		
Duration	Onset to conclusion (hours)	NOAA SED
Precipitation	Event accumulation (inches)	PRISM
Nighttime	1 if start between 6PM and 6AM	NOAA SED
Tropical storm	1 if tropical/hurricane class	NOAA SED
Vehicle involvement	1 if narrative indicates vehicles	NOAA SED
Geographic Factors		
Flash Flood Alley	Indicator for 44–county region	Baker (1975); Chang et al. (2023)
Mean elevation	Feet above sea level	NED
Mean slope	Degrees	NHD
Soil group	USDA hydrologic runoff potential (A–D)	STATSGO
Community Infrastructure		
Population density	Persons per square mile	ACS
Median structure age	Proxy for infrastructure vintage	ACS
Impervious surface	Percent impervious	NLCD
Bridge density	Count per square mile	TxDOT
Low–water crossings	Count per square mile	TNRIS
Road density	Miles per square mile	TxDOT

harms 11). Identifier guardrails are applied to remove unique IDs (e.g., *EVENT_ID*) and near–unique keys. Constants are dropped, and high–cardinality categorical features retain only the most frequent levels, with the remainder grouped as `__OTHER__`. The label *human_harm*, temporal index *YEAR*, and grouping variable *PopDen_Q* are preserved by design.

A critical aspect of this study is the sampling frame of the dataset. The 6,137 events are sourced from the NOAA Storm Events database, which, by definition, only contains confirmed flash flood events. The dataset contains no "no-flood" or "heavy-rain-no-flood" control instances. This must be stated unequivocally: this model does not predict whether a flash flood will occur. Its operational purpose is to solve the subsequent, and arguably more complex, problem of harm stratification. The model is designed to be triggered after a flash flood is considered imminent or is already underway (e.g., by an NWS watch/warning). Its sole function is to answer the question: "Given that a flash flood is occurring, what is the probability of this event causing human harm (fatality or injury)?" This defines its role as a decision-support tool for stratifying risk and allocating resources, not as a flood-detection system.

2.2 Preprocessing and Probability–Calibrated Ensembling

Numeric features are robustly scaled by centering with the median and dividing by the sample standard deviation. Categorical features are imputed using the mode and then one–hot encoded. These transformations are formally defined in Equation (1).

$$x'_{\text{num}} = \frac{x_{\text{num}} - \text{median}(x_{\text{num}})}{\text{sd}(x_{\text{num}})} \quad \text{and} \quad x'_{\text{cat}} = \text{OneHot}\big(\text{mode_impute}(x_{\text{cat}})\big). \tag{1}$$

We construct a heterogeneous ensemble from three base learners: logistic regression (ℓ_2 penalty, liblinear solver), random forest (600 trees, min leaf size 5, class–balanced sampling), and histogram gradient boosting (depth 3, learning rate 0.05, 400 iterations on numeric features). Each model's raw score, $f_k(x)$, is mapped to a probability, $\hat{p}_k(x)$, using isotonic regression fitted on the validation set, as shown in Equation (2).

$$\hat{p}_k(x) = \text{Isotonic}\big(f_k(x); \mathcal{D}_{\text{val}}\big), \quad k \in \{1, 2, 3\}. \tag{2}$$

A set of convex ensemble weights, $\boldsymbol{w}^*$, is learned on the 3-simplex by minimizing the Brier score on the validation set (Equation (3)). This process yielded the optimal weights $\{0.420, 0.540, 0.040\}$ for the (logistic regression, random forest, HGB) ensemble, respectively.

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\Delta_3} \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(x_i,y_i)\in\mathcal{D}_{\text{val}}} \Big(y_i - \sum_{k=1}^{3} w_k \hat{p}_k(x_i)\Big)^2, \quad \hat{p}(x) = \sum_{k=1}^{3} w_k^* \hat{p}_k(x). \tag{3}$$

2.3 Importance–Weighted Mondrian Conformal Prediction

While standard conformal prediction provides finite–sample validity under exchangeability, this assumption is violated by temporal drift. We therefore estimate density–ratio weights to approximately restore exchangeability between the calibration and deployment distributions. This is achieved by training a classifier to distinguish test samples (Z=1) from calibration samples (Z=0) based on their features X, as specified in Equation (4).

$$\Pr(Z{=}1 \mid X{=}x) = \hat{q}(x) = \sigma\big(\beta_0 + \boldsymbol{\beta}^\top \phi(x)\big), \quad \sigma(t) = (1+e^{-t})^{-1}. \tag{4}$$

The resulting probability, $\hat{q}(x)$, is used to compute the density ratio $\hat{w}(x)$ and the normalized importance weights $\tilde{w}_i^{\text{cal}}$ for each calibration point, as shown in Equation (5).

$$\hat{w}(x) = \frac{\hat{q}(x)}{1-\hat{q}(x)}, \qquad \tilde{w}_i^{\text{cal}} = \frac{\hat{w}(X_i)}{\sum_{j\in\mathcal{D}_{\text{cal}}} \hat{w}(X_j)} \cdot |\mathcal{D}_{\text{cal}}|. \tag{5}$$

To provide subgroup–specific guarantees for equity auditing, Mondrian conditioning is applied. We stratify the data into groups $G(x)$ based on population–density

quartiles, using the mapping defined in Equation (6).

$$G(x) = \begin{cases} \text{Q1}, & \text{PopDen}(x) \in [0, q_{0.25}) \\ \text{Q2}, & \text{PopDen}(x) \in [q_{0.25}, q_{0.50}) \\ \text{Q3}, & \text{PopDen}(x) \in [q_{0.50}, q_{0.75}) \\ \text{Q4}, & \text{PopDen}(x) \in [q_{0.75}, \infty) \end{cases} \tag{6}$$

Within each group g, a raw threshold $\hat{\tau}_g$ is set as the weighted $(1-\alpha)$–quantile of the conformal scores (here, probabilities) from the positive-class calibration samples $\mathcal{C}_g$, specified in Equation (7).

$$\hat{\tau}_g = Q_{1-\alpha}\Big(\{\hat{p}(X_i)\}_{i\in\mathcal{C}_g}, \{\tilde{w}_i^{\text{cal}}\}_{i\in\mathcal{C}_g}\Big), \quad \mathcal{C}_g = \{i : Y_i{=}1,\, G(X_i){=}g\}. \tag{7}$$

2.3.1 Hierarchical Shrinkage for Sparse Groups

In rare–event settings, $\mathcal{C}_g$ is often small, leading to high-variance thresholds. We stabilize these estimates using empirical–Bayes hierarchical shrinkage. We first compute the effective weighted positive sample size $n_{\text{eff}}^{(g)}$ (Equation (8)).

$$n_{\text{eff}}^{(g)} = \frac{\big(\sum_{i\in\mathcal{C}_g} \tilde{w}_i^{\text{cal}}\big)^2}{\sum_{i\in\mathcal{C}_g} (\tilde{w}_i^{\text{cal}})^2}. \tag{8}$$

The final group threshold τ_g is then computed as a weighted average of the raw group threshold $\hat{\tau}_g$ and the pooled global threshold τ_{global}, regularized by $\lambda{=}20$ pseudo–observations (Equation (9)).

$$\tau_g = \frac{n_{\text{eff}}^{(g)}\,\hat{\tau}_g + \lambda\,\tau_{\text{global}}}{n_{\text{eff}}^{(g)} + \lambda}. \tag{9}$$

2.4 Validation–Tuned Selective Abstention

To balance coverage and reliability, operational decisions are made using a three–way rule (warn, do not warn, or abstain) based on a symmetric margin $m \geq 0$. This rule is formally stated in Equation (10).

$$\hat{y}(x;m) = \begin{cases} 1, & \hat{p}(x) \geq \tau_{G(x)} + m, \\ 0, & \hat{p}(x) \leq \tau_{G(x)} - m, \\ \text{abstain}, & \text{otherwise.} \end{cases} \tag{10}$$

In operational terms, this rule outputs a warning (1) only when there is sufficient evidence to exclude the 'safe' outcome (0) at the desired confidence level. If the evidence is ambiguous (the prediction set contains both $\{0, 1\}$), the system abstains rather than issuing a low-confidence guess. The optimal margin m^* is tuned on the validation set. For each group g, we compute the recall $\hat{r}_g$ and its 95% Wilson lower bound (WLB), which provides a robust, high–confidence estimate of recall, via Equation (11).

$$\text{WLB}_g = \frac{\hat{r}_g + \frac{z^2}{2n_g} - z\sqrt{\frac{\hat{r}_g(1-\hat{r}_g)}{n_g} + \frac{z^2}{4n_g^2}}}{1 + \frac{z^2}{n_g}}, \quad z{=}1.96, \;\; \hat{r}_g = \frac{\text{TP}_g}{\text{TP}_g + \text{FN}_g}, \;\; n_g = \text{TP}_g{+}\text{FN}_g. \tag{11}$$

The margin m^* is chosen to maximize "acted coverage" (the fraction of non–abstained predictions) subject to the constraint that the recall for the worst–performing group, as measured by its WLB, must meet a high target (e.g., 0.90). This constrained optimization is defined in Equation (12). If no margin $m > 0$ satisfies the constraint, the policy defaults to $m^* = 0$ (no abstention) and reports the

infeasibility.

$$m^* = \arg \max_{m \in \{0, 0.01, \ldots, 0.30\}} \text{Coverage}(m) \quad \text{s.t.} \quad \min_g \text{WLB}_g(m) \geq 0.90. \tag{12}$$

2.5 Prequential Backtesting, Drift Metrics, and Decision Curves

To assess temporal robustness, we employ prequential (rolling) evaluation. For each test year $t \in [2009, 2017]$, the model is trained on 2005–$(t-3)$, validated internally, and calibrated using IW–MCP on $(t-2, t-1)$ before testing on year t.

Distributional shift is quantified using two metrics. First, the Population Stability Index (PSI), defined in Equation (13), measures the drift in feature distributions between two time periods (e.g., calibration vs. test).

$$\text{PSI} = \sum_{b=1}^{B} \left(p_{\text{new},b} - p_{\text{ref},b}\right) \ln\left(\frac{p_{\text{new},b}}{p_{\text{ref},b}}\right). \tag{13}$$

Second, a Kolmogorov–Smirnov (KS) test is used to compare the score distributions. Deployment utility is evaluated using decision–curve analysis, which computes the net benefit (NB) of the model across different probability thresholds p_t, as defined in Equation (14).

$$\text{NB}(p_t) = \frac{\text{TP}}{n} - \frac{\text{FP}}{n} \cdot \frac{p_t}{1 - p_t}. \tag{14}$$

2.6 Operational CPS Integration Architecture

To bridge the gap between this methodological protocol and operational deployment, we define a three–layer architecture (Figure 2) for an urban Cyber–Physical System (CPS). This architecture is designed to ingest real–time data and disseminate certified warnings.

(i) **Data Integration and Triggering Layer:** This layer is operationally defined by two core challenges. First, the model trigger: The system is a harm stratification tool, not a flood detection tool. It is designed to be triggered after an external source (e.g., an NWS Flash Flood Warning) has declared a flood imminent. Second is the spatio-temporal reconciliation: This is the single greatest technical hurdle. A critical, unaddressed mismatch exists, as the current model is trained on coarse, daily data (PRISM) at a coarse polygon-level (census tracts). An operational system must ingest high-resolution, sub-hourly data (e.g., 1km MRMS radar, point-gauges) and reconcile these mismatched scales. A simple application of the current preprocessing pipelines (Sec 2.2) is not possible. Therefore, this layer's primary role is not just ingestion, but a complex and non-trivial reconciliation and feature engineering task, which is a major research challenge in itself.

(ii) **Prediction and Certification Layer:** Executes the calibrated ensemble (Section 2.2) to generate risk scores. Critically, this layer also computes the rolling importance–weighted Mondrian thresholds (Section 2.3) on the most recent calibration set, outputting the final, certified prediction sets.

(iii) **Decision Support and Dissemination Layer:** Applies the validation–tuned abstention rule (Section 2.4) to the certified sets. It issues subgroup–aware alerts to incident managers and logs all outputs, drift metrics (Section 2.5), and coverage audits for transparency and rolling recalibration.

This modular design ensures that the entire protocol, from feature engineering to certified equity guarantees, is explicitly integrated into a deployable, auditable system.

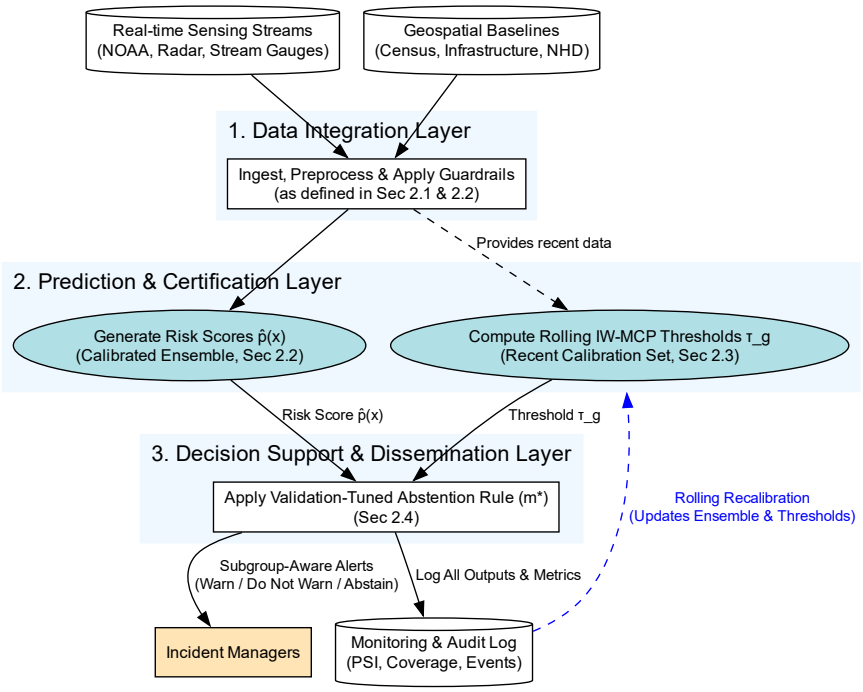


Fig. 2 Schematic of the three-layer operational CPS integration architecture. The framework ingests real-time data, computes certified risk predictions, and disseminates subgroup-aware alerts, enabling a rolling recalibration loop based on logged performance.

2.7 Implementation and Reproducibility

Analyses use Python 3.13 with scikit–learn 1.7, NumPy 2.3, Pandas 2.3, and Matplotlib 3.10. A fixed seed (42) controls all stochastic processes, including data splits and bootstrap resampling. Unit tests were implemented to verify the correctness of the density–ratio weighting and weighted quantile algorithms against known analytic cases. All figures and tables referenced in Section 3 are generated from this reproducible pipeline.

3 Results and Discussion

This section evaluates the calibrated ensemble and the IW–Mondrian conformal protocol along four axes: (i) discrimination and calibration, (ii) certified operating performance, (iii) temporal robustness, and (iv) covariate–shift and decision–utility diagnostics. All test–set results refer to the strict 2018–2019 holdout period (prevalence 1.86%) unless otherwise noted.

3.1 Discrimination and Calibration

The Brier-optimally weighted ensemble achieves moderate discrimination on the test set, with an AUROC of 0.758 (95% CI: 0.545–0.931) and an AUPRC of 0.198 (95% CI: 0.044–0.504). These key performance metrics are compiled in Table 2. On aggregate, the model appears well-calibrated, achieving a low Brier score (0.0187) and a low Expected Calibration Error (ECE) of 0.012. However, the reliability diagram (Figure 4) reveals the critical challenge of assessing calibration in a severe rare-event setting. The histogram in Figure 4 shows that the vast majority of test samples (>90%) are low-probability (predicted probability < 0.2) and fall into the first two bins. These low-risk predictions are indeed well-calibrated, and their high sample count dominates the ECE metric, resulting in a low value. Conversely, the bins for higher-risk predictions (predicted probability > 0.2) are populated by a very small number of samples, leading to high variance and unreliable calibration. The plot shows that while the model is well-calibrated on average (driven by 'no harm' events), its calibration for high-probability, high-consequence events is statistically uncertain due to this data sparsity. This limitation is a key finding, underscoring the difficulty of forecasting rare events. The ROC and precision-recall (PR) curves are visualized in Figure 3. Given the 1.86% prevalence, the PR curve provides the more informative view of discrimination, showing skillful prediction well above the no-skill baseline.

Table 2 Discrimination and calibration on the 2018–2019 test set with 1,000–replicate nonparametric bootstraps. ECE uses 10 equal–width bins.

Metric	Point Estimate	95% CI
AUROC	0.758	(0.545, 0.931)
AUPRC	0.198	(0.044, 0.504)
Brier Score	0.0187	(0.0098, 0.0287)
ECE (10 bins)	0.0120	–

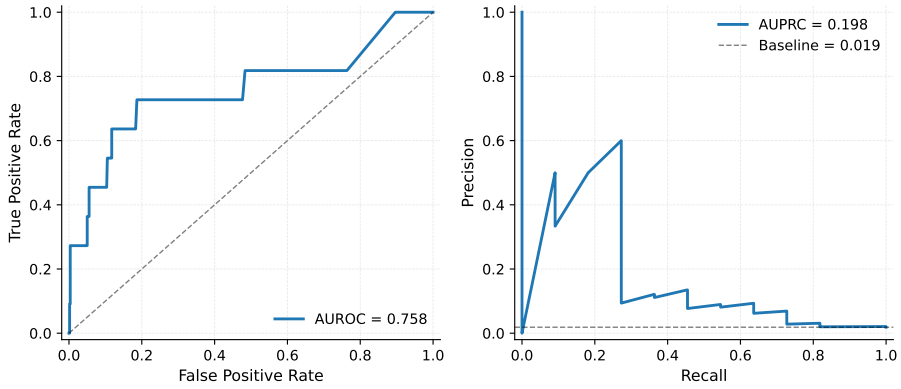


Fig. 3 ROC and precision–recall curves on the 2018–2019 test set. At 1.86% prevalence, the PR view is more informative for discrimination.

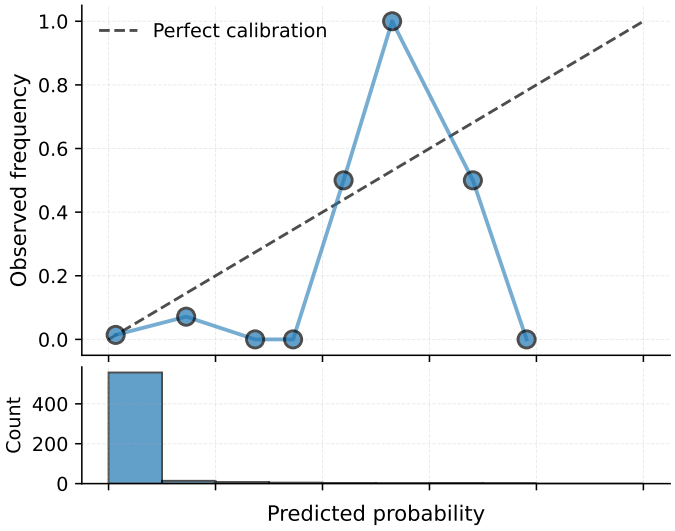


Fig. 4 Reliability diagram (10 bins) on the 2018–2019 test set. The near–identity tracking is consistent with ECE = 0.012.

Figure 5 illustrates the impact of temporal drift on the model's discriminative power. We quantitatively assess this separation using a two-sample KS test between the predicted probability distributions for positive (harm=1) and negative (harm=0) classes within each data partition. The KS test confirms the visual trend: class separation is wide and statistically significant on the **validation set** ($\text{KS} = 0.778$, $p < 0.001$). This separation, although still statistically significant, reveals a clear trend of degradation in the subsequent periods, weakening the calibration set ($\text{KS} = 0.693$, $p < 0.001$) and further diminishing the test set ($\text{KS} = 0.541$, $p = 0.002$). This quantitative degradation in the KS statistic—a drop from 0.778 to 0.541—provides direct, statistical evidence of the model's eroding discriminative ability due to temporal drift. This finding reinforces the operational necessity of the importance-weighting protocol to correct for this shift.

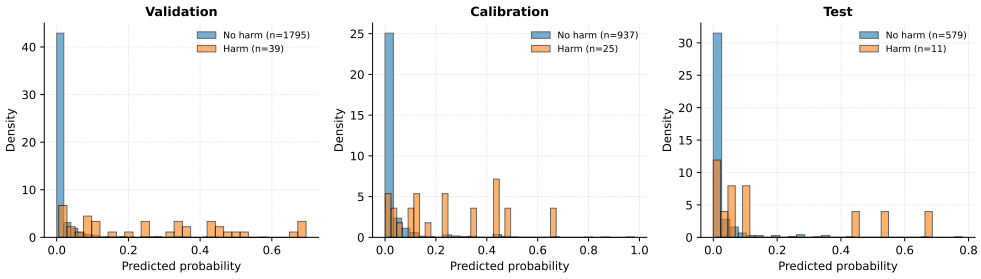


Fig. 5 Predicted–probability distributions for harmful (orange) and non–harmful (blue) events across validation, calibration, and test splits.

3.2 Certified Performance with IW–Mondrian CP

The selective abstention optimization (Equation (12)) was infeasible, as no margin $m > 0$ could satisfy the $min_g \text{WLB}_g(m) \geq 0.90$ constraint on the validation set. The protocol therefore defaulted to $m^* = 0$, implying full (100%) acted coverage.

Furthermore, due to the sparsity of positive events in the calibration set, the adaptive grouping procedure merged all four population–density quartiles into a single

certified stratum. The final operating policy thus uses the shrunk global threshold, $\tau_{\text{global}} \approx 0.0519$, for all predictions. The resulting test–set confusion matrix is shown in Figure 6, yielding TP = 5, FN = 6, FP = 59, and TN = 520. This operating point corresponds to a recall of 0.455, a precision of 0.078, and a false–positive rate (FPR) of 0.102. This summary is also presented in Table 3.

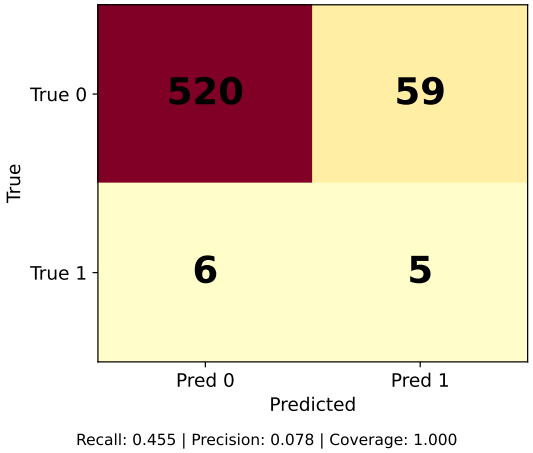


Fig. 6 Confusion matrix under IW–Mondrian CP with hierarchical shrinkage and $m^* = 0$ on the 2018–2019 test set.

Table 3 Merged certified group (B1) on the 2018–2019 test set at acted coverage = 1.000 ($m^* = 0$). The Wilson lower bound (WLB) is shown for transparency.

Group	n_{test}	n_{pos}	Recall	WLB	Precision	FPR
B1 (merged)	590	11	0.455	0.213	0.078	0.102

The operational importance of the full protocol is demonstrated in Table 4. A naive, unweighted conformal predictor achieves high recall (0.818) but at the cost of a prohibitive FPR (0.484). Applying importance weighting (IW CP) corrects for covariate shift and slashes the FPR to 0.138. Our full IW–Mondrian CP protocol

provides a further reduction in FPR to 0.102, yielding the highest precision (0.078) of the three policies.

Table 4 Operating characteristics on the 2018–2019 test set (n=590, positives = 11) at acted coverage = 1.000. Wilson 95% intervals are shown for recall and FPR.

Method	Recall	95% CI	FPR	95% CI	Precision	Coverage
Standard CP (pooled, unweighted)	0.818	[0.523, 0.949]	0.484	[0.443, 0.524]	0.031	1.000
IW CP (pooled, weighted)	0.545	[0.281, 0.786]	0.138	[0.112, 0.169]	0.070	1.000
IW–Mondrian CP (shrunk; ours)	0.455	[0.214, 0.719]	0.102	[0.080, 0.129]	0.078	1.000

Figure 7 serves as a diagnostic, visualizing the final threshold structure. As expected, after adaptive merging, the single certified stratum (B1) adopts the global threshold.

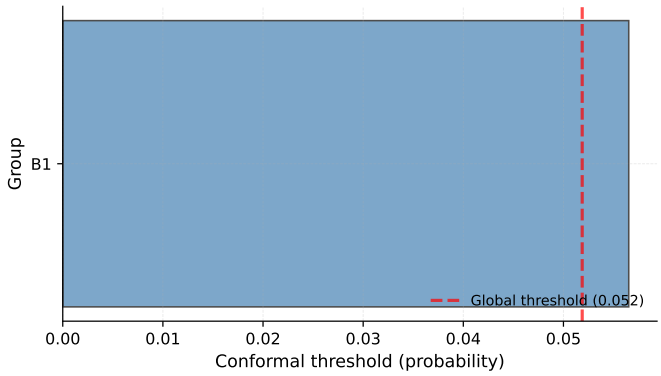


Fig. 7 Conformal thresholds after adaptive merging. The merged stratum uses the global threshold $\tau_{\text{global}} \approx 0.0519$.

3.3 Temporal Robustness (Prequential Backtesting)

The prequential backtest (2009–2017) confirms these findings are stable over time. As shown in Figure 8, discrimination remains modest but consistent (median AUROC $\approx$ 0.723; IQR 0.681–0.751). Acted coverage remains 1.0 in all years, as the selective

abstention constraint is never met. The worst–group WLB fluctuates substantially, driven by the stochastic, year–to–year sparsity of harmful events within subgroups, and never approaches the 0.90 target. This result reinforces that the limitation is statistical (data sparsity), not model instability.

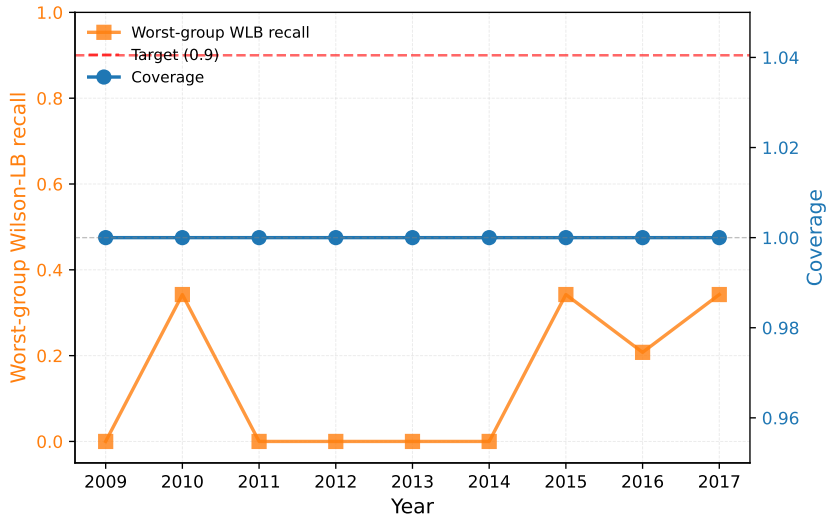


Fig. 8 Prequential results (2009–2017): acted coverage (blue; left axis) and worst–group WLB recall (orange; right axis). The dashed line marks the 0.90 target.

3.4 Covariate Shift, Drift Diagnostics, and Decision Utility

The distributional shift between the calibration (2016–2017) and test (2018–2019) periods, while not severe, is statistically detectable. A KS test on the models' score distributions finds a significant difference (KS = 0.0769, p=0.0248), justifying the use of importance weighting.

At the feature level, Table 5 and Figure 10 detail the Population Stability Index (PSI) for each predictor. Several key environmental and demographic features show "noteworthy" drift ($PSI > 0.10$). To answer how these distributions shifted, we analyzed the directional change for the top-drifting features, as shown in Figure 9. The

results are highly revealing. For Mean Elevation (PSI=0.248), the risk profile is polarizing: while the mean elevation of harm events increased (275.6 ft to 299.0 ft), the median decreased (180.9 ft to 162.5 ft) and the 25th percentile plummeted (102.2 ft to 44.2 ft). As seen in Figure 9 (left), this indicates that risk is spreading to new, diverse geographies, simultaneously affecting more very-low-lying areas and new higher-elevation zones. This finding is paired with an even more critical insight regarding Precipitation (PSI=0.194): the threshold for harm is decreasing. The mean precipitation associated with a harmful event dropped from 4.51 inches to 3.70 inches. This suggests that recent harmful events are, on average, being triggered by less rainfall, indicating an increase in urban vulnerability as the city's capacity to absorb rainfall diminishes. This confirms that the underlying data-generating process is non-stationary, providing a key insight: the changing risk landscape is not a simple linear trend, but a complex increase in vulnerability. This finding strongly justifies the operational necessity of the importance-weighting protocol to correct for this non-stationarity.

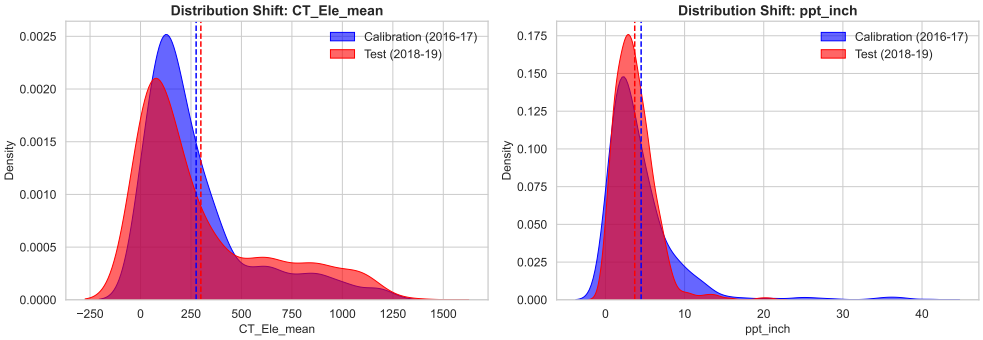


Fig. 9 Kernel density plots illustrating the directional drift for the top two non-stationary features: mean elevation (left) and precipitation (right). This analysis reveals a complex shift in the flood-harm profile between the calibration (2016–17) and test (2018–19) periods.

The decision-curve analysis in Figure 11 assesses operational utility. The x-axis p_t is the probability threshold, which represents the cost-benefit "exchange rate" for a decision-maker (i.e., at what probability of harm are they willing to intervene?). It is

Table 5 Top drifting features by PSI between calibration (2016–2017) and test (2018–2019). Thresholds: $[0.0, 0.10)$ → negligible; $[0.10, 0.25]$ → noteworthy; $(0.25, 1.0]$ → severe.

Feature	PSI	Interpretation
Mean elevation	0.248	Noteworthy
Median structure age	0.221	Noteworthy
Ground slope (degrees)	0.215	Noteworthy
Precipitation (inches)	0.194	Noteworthy
Population density	0.129	Noteworthy

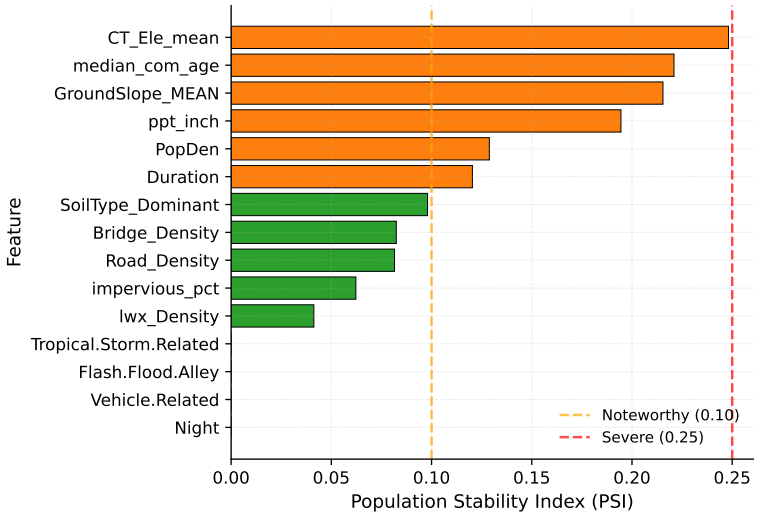


Fig. 10 PSI by feature between calibration and test. Dashed lines at 0.10 and 0.25 mark "noteworthy" and "severe" drift.

not a model's operating threshold, but rather a function of the cost of a false positive (FP) versus a false negative (FN), where $p_t = \text{Cost(FP)}/(\text{Cost(FP)} + \text{Cost(FN)})$. In a safety-critical, rare-event context, the Cost(FN)—a missed fatality or injury—is vastly higher than the Cost(FP)—a false alarm. A p_t of 0.5, which is common in non-rare-event problems, would absurdly imply Cost(FP) = Cost(FN). Therefore, the only operationally relevant range to examine is at low p_t. A cost-benefit ratio of 9:1 (FN:FP) corresponds to $p_t = 0.10$, while a ratio of 99:1 corresponds to $p_t \approx 0.01$. Our analysis correctly focuses on this $p_t \in [0.01, 0.10]$ range. Within this range, the model's

net benefit is near-neutral compared to a "treat-none" policy. This is an expected result for a rare-event model, confirming that the model provides positive utility only in high-stakes scenarios where a decision-maker's "exchange rate" is heavily weighted to avoiding a missed catastrophic event (i.e., a very low p_t).

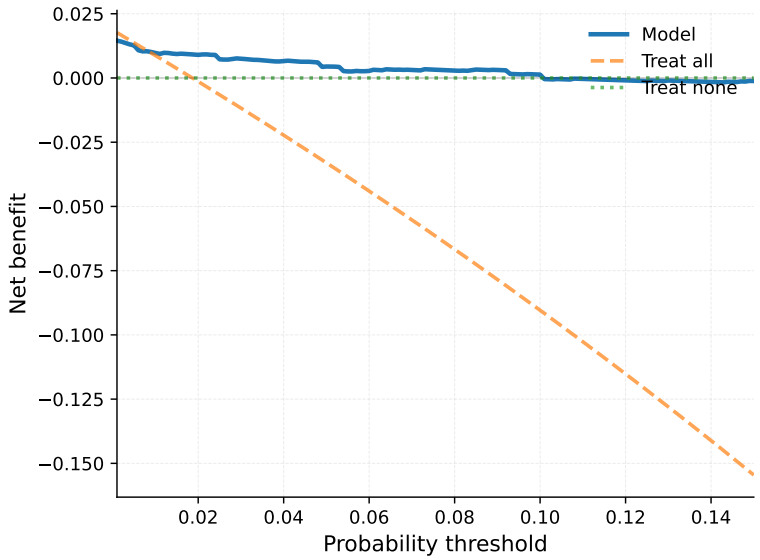


Fig. 11 Decision–curve analysis on the test set. Net benefit is near treat–none at typical thresholds, but can become favorable when misses are cost substantially higher than false alarms.

3.5 Synthesis and Operational Implications

The IW–Mondrian conformal protocol successfully produces calibrated probabilities and transparent, distribution–free guarantees that are robust to measurable temporal shift. The framework's components are all operationally critical: importance weighting is essential for reducing false alarms (Table 4), and empirical–Bayes shrinkage prevents unstable group thresholds.

This result defines a clear, data-driven evidence threshold for deployment. To quantify this threshold, we performed the statistical power analysis on the 95% Wilson

Lower Bound shown in Figure 12. This analysis calculates the required number of positive samples (n_g) to certify a high WLB (e.g., 80%), assuming the model is already performing well (e.g., 90–95% true recall). The analysis provides a direct, quantitative justification for our "50–100 event" threshold. As the data shows, a model with 90% true recall would require at least $n_g = 75$ positive events to certify a WLB ¿ 0.80. To approach a 90% WLB, $n_g > 100$ positive events are required even with a 95% true recall. Our 50–100 event estimate is therefore a direct calculation of the statistical power required by our certification framework. Until this evidence threshold is met, the system is suitable for deployment only in an advisory mode, paired with continuous rolling recalibration and transparent subgroup auditing.

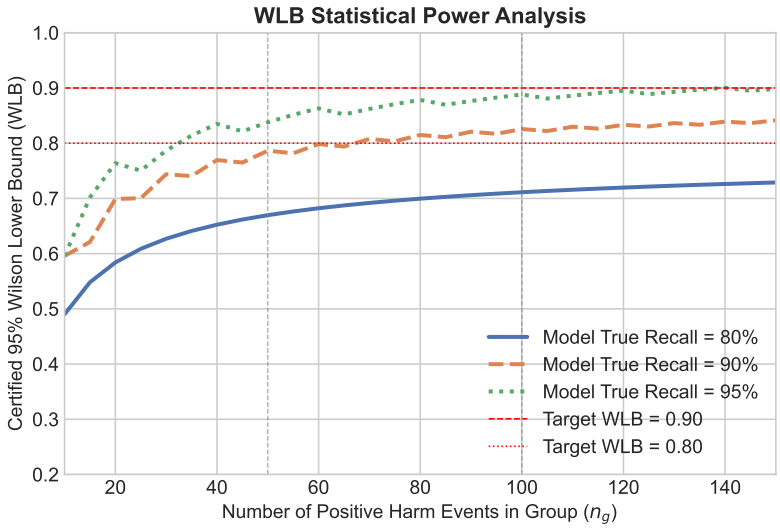


Fig. 12 Statistical power analysis for the 95% Wilson Lower Bound (WLB). The plot shows the WLB that can be certified for a given number of positive harm events (n_g), assuming a model with a high true recall (e.g., 90% or 95%). This analysis provides the quantitative basis for the evidence threshold cited in Section 3.5.

4 Conclusion

We developed and empirically validated an importance–weighted Mondrian conformal prediction framework for flash–flood human harm forecasting under temporal covariate shift with group–conditional fairness auditing. On 6,137 Texas events (2005–2019) at census–tract resolution, the calibrated ensemble achieved moderate discrimination on the 2018–2019 test set (AUROC = 0.758, 95% CI: 0.545–0.931; AUPRC = 0.198) with strong calibration (ECE = 0.012; Brier = 0.0187). Measurable shift between calibration and test was detected both at the feature level (e.g., elevation PSI = 0.248, precipitation PSI = 0.194) and in score distributions (KS = 0.0769, p = 0.0248). Our importance weighting protocol proved operationally critical, reducing the test false–positive rate from a prohibitive 48.4% (unweighted pooled CP) to 10.2% at an acted recall of 45.5% (confusion matrix: TP= 5, FN= 6, FP= 59, TN= 520). Prequential backtesting across 2009–2017 indicated stable discrimination (median AUROC = 0.723, IQR: 0.681–0.751) under evolving conditions.

The binding limitation of this system is one of data scarcity, not a methodological flaw. With only 11 harmful events in the test window, the problem is sample-size-limited; consequently, the worst-group Wilson lower bound on recall was 0.213, far below the 0.90 target. This is not a model failure, but a predictable, statistical consequence of this scarcity. Our statistical power analysis (Figure 12) confirms this, showing that on the order of 50–100 additional harmful events, reasonably distributed across strata, would be required before the WLB metric could certify worst-group recall in the 80–90% range (e.g., $n_g = 75$ for a WLB of 0.803 at 90% true recall). This provides a concrete, data-driven evidence threshold for operational adoption.

4.1 Methodological Contributions and Urban CPS Integration

This work contributes four practical advances for distribution–free uncertainty quantification in rare–event, safety–critical prediction under drift: (i) coupling importance

weighting with Mondrian conditioning to address temporal covariate shift while preserving group–conditional coverage; (ii) stabilizing small–sample group thresholds via hierarchical (empirical–Bayes) shrinkage; (iii) casting the coverage–utility trade–off as a validation–tuned selective–abstention problem using worst–group Wilson bounds; and (iv) using prequential backtesting to separate average (marginal) validity from the conditional variability unavoidable in sparse subgroups.

The framework is deployment–ready within urban cyber–physical systems. We propose a phased pathway that balances rigor and urgency: Phase 1 (shadow, 2–3 years) accrues ∼30–50 additional harmful events and elicits local loss ratios; Phase 2 (advisory, 2–3 years) provides uncertainty–quantified risk assessments to incident managers without automated alerts, while actively monitoring PSI–flagged drift; Phase 3 (operational) activates automated dissemination once ∼80–100 cumulative harmful events (distributed across strata) support certified worst–group recall near 0.80 at 95% confidence. The modular stack—shift correction, hierarchical regularization, and abstention tuning—is designed to support rolling recalibration without wholesale retraining as new data arrive.

4.2 Broader Implications and Future Directions

The protocol generalizes to other domains where certified predictions for rare, high–consequence events are required despite drift, including urban resilience, critical–infrastructure protection, transportation safety, industrial reliability, and clinical decision support. While recent studies have demonstrated the efficacy of ensemble models like XGBoost and Random Forest for susceptibility mapping (Bammou et al. 2024; Chihi et al. 2025), these approaches generally output point predictions without finite-sample coverage guarantees. Similarly, while physics-informed machine learning (Cui et al. 2025) and dynamic rainfall thresholds (Zhao et al. 2025) improve physical consistency, they remain vulnerable to distributional drift if not explicitly corrected.

Our framework complements these advances by wrapping such predictive models in a conformal layer, ensuring that the output sets satisfy statistical coverage requirements even when the underlying data distribution shifts.

Promising extensions to this work are substantial, but all are secondary to addressing the primary limitation identified in this study: the spatio-temporal data mismatch. Our model is trained on coarse, daily precipitation proxies (PRISM), while a true operational system must ingest high-resolution, sub-hourly data. However, the most high-impact future direction is not simply to re-train on finer precipitation data, but to fundamentally change the modeling paradigm. The next-generation system must move from a precipitation-based proxy model to a hydrology-driven impacts model. This would involve a two-stage process: (i) Ingest high-resolution radar data (e.g., NEXRAD) into an established operational system like the MRMS (Multi-Radar Multi-Sensor) (Zhang et al. 2016). (ii) Use the output of a distributed hydrology model, such as the FLASH project's Maximum Unit Streamflow ($m^3s^{-1}km^{-2}$) (Gourley et al. 2017), as the primary physical predictor. This would replace our model's reliance on coarse rainfall with a direct, high-resolution physical measure of stream response. Only after this fundamental data and model upgrade can secondary extensions be fully explored, such as: spatially-aware conformal calibration to borrow strength across tracts; replacing population density with curated social-vulnerability constructs (e.g., CDC SVI); and testing for domain adaptation and transferability using this framework's PSI and IW-CP components as the diagnostic and adaptation toolkit.

4.3 Closing Remark

Importance–weighted Mondrian conformal prediction, augmented with hierarchical shrinkage and validation–tuned abstention, offers a rigorous and transparent basis for rare–hazard forecasting under drift. By stating both what can be certified now (recall = 45.5%, worst–group WLB = 0.213) and what additional evidence is required for

stronger guarantees (roughly 50–100 more harmful events for 0.80–0.90 worst–group recall at 95% confidence), the framework moves deployment from intuition to evidence-based policy—an essential step toward responsible algorithmic decision-making in public safety and resilient urban systems.

Acknowledgments

This work builds upon foundational data compiled by Gharaibeh, Meyer, Zou, Olivera, Sansom, Chang, and Wilkho (Gharaibeh et al. 2023; Chang et al. 2023), whose meticulous integration of multi-source flood incident records enabled this analysis. We gratefully acknowledge their contributions to advancing flash flood impact research and their commitment to open data practices through DesignSafe-CI. We thank the three Texas emergency management professionals who participated in stakeholder interviews for intervention cost elicitation (anonymized per IRB protocol), providing essential operational context informing the optimization formulations. Support for this research came from start-up funds provided by the CSU Pueblo Foundation and the School of Engineering at Colorado State University Pueblo.

Data Availability Statement

The foundational dataset used in this research is publicly available through DesignSafe-CI: Gharaibeh, N., M. Meyer, L. Zou, F. Olivera, G. Sansom, S. Chang, R. Wilkho (2023). *Reducing the Human Impacts of Flash Floods: Development of Microdata and Causal Model to Inform Mitigation and Preparedness*. DesignSafe-CI. https://doi.org/10.17603/ds2-e91y-cv92.

Code availability

All code used for data preprocessing, model training, conformal prediction, and the generation of the figures and tables in this study has been archived on

Zenodo as *"Conformal Prediction for Flash Flood Harm"* by Gupta, H. S., Duarte García, J., and Biswas, S. (2025) (DOI: 10.5281/zenodo.17594526). For editors and reviewers, the complete reproducible framework (including scripts, notebooks, and curated data) is accessible via the private Zenodo review link: `https://zenodo.org/records/17594526 (private review link)`.

Following publication, we will make the Zenodo record publicly accessible, and the DOI (10.5281/zenodo.17594526) will serve as the permanent link to the full open-source reproducible framework.

Conflict of Interest

The authors declare no conflict of interest. Funding sources had no role in study design, data analysis, interpretation of results, manuscript preparation, or the decision to submit for publication.

References

Ashley, S.T., Ashley, W.S.: Flood fatalities in the United States. Journal of Applied Meteorology and Climatology **47**(3), 805–818 (2008)

Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)

Agresti, A., Coull, B.A.: Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician **52**(2), 119–126 (1998)

Ahmadalipour, A., Moradkhani, H.: A data-driven analysis of flash flood hazard, fatalities, and damages over the CONUS during 1996–2017. Journal of Hydrology **578**, 124106 (2019)

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)

Ashraf, M., Shah, Q., Arshad, A., Murtaza, G.: Assessing hazardous flash flood susceptibility using multivariate zonation mapping techniques in pishin district, balochistan province of pakistan. Natural Hazards **121**, 22935–22956 (2025) https://doi.org/10.1007/s11069-025-07723-0

Baker, V.R.: Flood hazards along the balcones escarpment in central texas: alternative approaches to their recognition, mapping, and management. Bureau of Economic Geology, University of Texas at Austin, Circular **75**(5) (1975)

Bammou, Y., Benzougagh, B., Igmoullan, B., Ouallali, A., Kader, S., Spalevic, V., Sestras, P., Billi, P., Marković, S.B.: Optimizing flood susceptibility assessment in semi-arid regions using ensemble algorithms: a case study of moroccan high atlas. Natural Hazards **120**, 7787–7816 (2024) https://doi.org/10.1007/s11069-024-06550-z

Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J.: Conformal prediction beyond exchangeability. The Annals of Statistics **51**(2), 816–845 (2023)

Chang, S., *et al.*: Environmental, climatic, and situational factors influencing the probability of fatality or injury occurrence in flash flooding: A rare event logistic regression predictive model. Natural Hazards **116**(3), 3957–3978 (2023)

Cui, P., Bazai, N.A., Qiang, Z., Jiao, W., Yan, W., Xu, Q., Yu, L., Bo, Z.: Flood risk assessment with machine learning: insights from the 2022 pakistan mega-flood and climate adaptation strategies. npj Natural Hazards **2**(42) (2025) https://doi.org/10.1038/s44304-025-00096-1

Chihi, H., Hammami, M.A., Mezni, I.: Flood susceptibility mapping in data-scarce

arid environments: guided by geology-driven knowledge and multi-event cloud-based validation. Natural Hazards **121**, 20855–20901 (2025) https://doi.org/10.1007/s11069-025-07533-4

Dawid, A.P.: Present position and potential developments: Some personal views statistical theory the prequential approach. Journal of the Royal Statistical Society: Series A (General) **147**(2), 278–290 (1984)

Ding, J., Wang, X., Zhou, S., Lei, S.: Holistic risk assessment using a hybrid approach in a flash flood disaster-prone area of the poyang lake basin. Natural Hazards **121**, 12959–12984 (2025) https://doi.org/10.1007/s11069-025-07306-z

Easterling, D.R., *et al.*: Precipitation change in the United States. Climate Science Special Report: Fourth National Climate Assessment **1**, 207–230 (2017)

Efron, B., Morris, C.: Stein's paradox in statistics. Scientific American **236**(5), 119–127 (1977)

Foygel Barber, R., Candes, E.J., Ramdas, A., Tibshirani, R.J.: The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA **10**(2), 455–482 (2021)

Gourley, J.J., Flamig, Z.L., Vergara, H., Kirstetter, P.-E., Clark III, R.A., Argyle, E., Arthur, A., Martinaitis, S., Terti, G., Erlingis, J.M., *et al.*: The flash project: Improving the tools for flash flood monitoring and prediction across the united states. Bulletin of the American Meteorological Society **98**(2), 361–372 (2017)

Gharaibeh, N., Meyer, M., Zou, L., Olivera, F., Sansom, G., Chang, S., Wilkho, R.: Reducing the Human Impacts of Flash Floods: Development of Microdata and Causal Model to Inform Mitigation and Preparedness. DesignSafe-CI (2023). https://doi.org/10.17603/ds2-e91y-cv92 . https://doi.org/10.17603/ds2-e91y-cv92

572 Jonkman, S.N.: Global perspectives on loss of human life caused by floods. Natural
573 Hazards **34**(2), 151–175 (2005)

574 Kelsch, M., Caporali, E., Lanza, L.G.: Hydrometeorology of flash floods. In: Coping
575 with Flash Floods, pp. 19–35. Springer, ??? (2001)

576 Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V.: Machine learning
577 for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge
578 and Data Engineering **31**(8), 1544–1554 (2018)

579 Kim, Y., Oh, J., Bartos, M.: Stormwater digital twin with online quality control detects
580 urban flood hazards under uncertainty. Sustainable Cities and Society **118**, 105982
581 (2025)

582 King, G., Zeng, L.: Logistic regression in rare events data. Political analysis **9**(2),
583 137–163 (2001)

584 Lei, J., Wasserman, L.: Distribution-free prediction bands for non-parametric regres-
585 sion. Journal of the Royal Statistical Society Series B: Statistical Methodology
586 **76**(1), 71–96 (2014)

587 Mohtar, W.H.M.W., Abdullah, J., Maulud, K.N.A., Muhammad, N.S.: Urban flash
588 flood index based on historical rainfall events. Sustainable Cities and Society **56**,
589 102088 (2020)

590 Morris, C.N.: Parametric empirical bayes inference: theory and applications. Journal
591 of the American statistical Association **78**(381), 47–55 (1983)

592 Quang, N.H., Lee, H., Kim, N., Kim, G.: Real-time flash flood detection employing
593 the yolov8 model. Earth Science Informatics **17**, 4809–4829 (2024) https://doi.org/
594 10.1007/s12145-024-01428-x

595 Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N.,
596 Prabhat, F.: Deep learning and process understanding for data-driven earth system
597 science. Nature **566**(7743), 195–204 (2019)

598 Romano, Y., Sesia, M., Candes, E.: Classification with valid and adaptive coverage.
599 Advances in Neural Information Processing Systems **33**, 3581–3591 (2020)

600 Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning
601 Research **9**(3) (2008)

602 Tibshirani, R.J., Foygel Barber, R., Candes, E., Ramdas, A.: Conformal prediction
603 under covariate shift. Advances in Neural Information Processing Systems **32** (2019)

604 Terti, G., Ruin, I., Anquetin, S., Gourley, J.J.: Dynamic vulnerability factors for
605 impact-based flash flood prediction. Natural Hazards **79**(3), 1481–1497 (2015)

606 Terti, G., Ruin, I., Anquetin, S., Gourley, J.J.: A situation-based analysis of flash
607 flood fatalities in the United States. Bulletin of the American Meteorological Society
608 **98**(2), 333–345 (2017)

609 Terti, G., Ruin, I., Gourley, J.J., Kirstetter, P., Flamig, Z., Blanchet, J., Arthur, A.,
610 Anquetin, S.: Toward probabilistic prediction of flash flood human impacts. Risk
611 analysis **39**(1), 140–161 (2019)

612 Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World.
613 Springer, ??? (2005)

614 Wilson, E.B.: Probable inference, the law of succession, and statistical inference.
615 Journal of the American Statistical Association **22**(158), 209–212 (1927)

616 Wobus, C., Lawson, M., Jones, R., Smith, J., Martinich, J.: Estimating monetary
617 damages from flooding in the united states under a changing climate. Journal of

Flood Risk Management **7**(3), 217–229 (2014)

Zahran, S., Brody, S.D., Peacock, W.G., Vedlitz, A., Grover, H.: Social vulnerability and the natural and built environment: a model of flood casualties in texas. Disasters **32**(4), 537–560 (2008)

Zhou, S., Geng, X., Jia, W., Xu, H., Xu, X., Chen, H., Wang, M., Wu, Z.: Towards climate-adaptive equality in coastal megacities: Assessing urban flooding risk disparities and nonlinear effect of multidimensional indicators through an interpretable lightgbm-shap framework. Sustainable Cities and Society, 106809 (2025)

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cocks, S., Martinaitis, S., *et al.*: Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. Bulletin of the American Meteorological Society **97**(4), 621–638 (2016)

Zhou, W., Li, Y., Gou, M., Wang, B., Fang, C.: Urban spatial vulnerability analysis based on urban systems using support vector machine. Sustainable Cities and Society **123**, 106274 (2025)

Zhao, Y., Wu, X., Guo, L., Qin, G., Li, X., Li, H.: Evaluation of rainfall-threshold methods for flash flood warnings based on soil moisture conditions. Natural Hazards **121**, 12053–12076 (2025) https://doi.org/10.1007/s11069-025-07272-6