Supplementary Material

S1. Sensitivity analysis using exposure-adjusted indicators

To evaluate whether the severity ranking was sensitive to differences in urban exposure and socioeconomic scale, an alternative exposure-adjusted specification was constructed. Absolute impact indicators were replaced by relative measures based on corresponding urban exposure conditions: affected population was expressed relative to the resident population, affected area relative to built-up area, and economic loss relative to GDP. This analysis does not represent an alternative definition of the core FSI, but evaluates whether the conclusions are affected by differences in city size, population scale, and economic capacity.

Table S1. Comparison of original and exposure-adjusted impact indicators.

Indicator	Normalization
Affected population	Affected population / population
Affected area	Affected area / built-up area
Economic loss	Economic loss / GDP
Overall ranking	Spearman ρ=0.930

S2. Sensitivity analysis of fatality indicator representation

Fatalities represent an important dimension of flood impacts; however, absolute fatality counts may be affected by differences in population size among cities. To evaluate whether the representation of the fatality indicator influenced the constructed Flood Severity Index (FSI), an alternative specification was tested by replacing absolute fatalities with population-normalized fatalities.

In the original FSI framework, fatalities were represented using the recorded number of deaths associated with each flood event. In the sensitivity analysis, fatalities were converted into a population-normalized indicator:

$$Fatality\ rate = \frac{Fatalities}{Population} \times 100{,}000$$

The revised FSI was recalculated using the alternative fatality specification, while all other indicators, weighting procedures, and model settings remained unchanged. The consistency between the original and alternative FSI results was evaluated at both event and city scales using rank correlation and ranking comparison metrics.

Table S2. Sensitivity analysis of fatality indicator representation

Metric	Result
Event-level correlation	ρ=0.9995
Mean absolute rank change	3.3 positions / 1,157 events
Overlap of top-50 severe events	49/50
City-level Spearman correlation	ρ=0.994
Overlap of top-5 cities	5/5

The results showed that replacing absolute fatalities with population-normalized fatalities produced negligible changes in the final FSI rankings. The event-level correlation between the two specifications was extremely high ($\rho = 0.9995$), and 49 of the 50 most severe events identified by the original FSI were retained under the alternative specification. At the city scale, the ranking consistency was also high ($\rho = 0.994$), with the same five cities occupying the highest positions under both specifications.

These findings indicate that the conclusions of the FSI framework are not sensitive to the representation of the fatality indicator. Therefore, both fatality representations produce highly consistent FSI outcomes, and the absolute fatality formulation was retained in the main analysis because fatalities directly represent human losses.

S3. Nested cross-validation and hyperparameter optimization

A nested cross-validation framework was implemented to obtain unbiased estimates of model performance and avoid information leakage during hyperparameter optimization. The outer loop consisted of 10-fold cross-validation, where each fold was sequentially used as an independent test set for out-of-sample evaluation. Within each outer training partition, a 5-fold GridSearchCV procedure was conducted exclusively for hyperparameter optimization. The optimal parameter combination was selected according to the highest mean R^2 value from the inner validation folds. The model with the selected parameters was then refitted using the complete outer training partition and evaluated on the corresponding held-out test fold. Performance metrics, including R^2, RMSE, and MAE, were calculated for each outer test fold and reported as mean ± standard deviation across the 10 outer folds. A fixed random seed (42) was applied to all data partitions and model training procedures to ensure reproducibility.

Table S3. Hyperparameter search spaces of machine-learning models

Model	Hyperparameters Searched
RF	n_estimators ∈ {100,200}, max_depth ∈ {5,10,None}, min_samples_leaf ∈ {1,2}
GBDT	n_estimators ∈ {100,200}, max_depth ∈ {3,5}, learning_rate ∈ {0.05,0.1}, min_samples_leaf ∈ {1,2}
XGBoost	n_estimators ∈ {100,200}, max_depth ∈ {3,5}, learning_rate ∈ {0.05,0.1}, subsample ∈ {0.8,1.0}
LightGBM	n_estimators ∈ {100,200}, max_depth ∈ {3,5,-1}, learning_rate ∈ {0.05,0.1}, num_leaves ∈ {31,63}
MLP	hidden_layer_sizes ∈ {(50,),(100,),(100,50)}, alpha ∈ {1e-4,1e-3}, learning_rate_init ∈ {0.001,0.01}

S4. Sensitivity analysis of online popularity exclusion

S4.1 Comparison between four-indicator Core FSI and five-indicator Extended FSI

To examine whether online popularity influenced the constructed flood severity index, a sensitivity analysis was conducted by comparing the revised four-indicator Core FSI with the original five-indicator specification that included online popularity. Because online popularity reflects public attention and information dissemination rather than direct disaster consequences, it was excluded from the revised FSI framework.

The relationship between online popularity and the Core FSI was first examined. The correlation analysis showed that online popularity was not significantly associated with the Core FSI (Spearman's $\rho = -0.013$, $p = 0.654$), indicating that it represented a distinct dimension from direct flood impacts.

Despite this conceptual difference, the event-level rankings generated by the two specifications remained highly consistent (Spearman's $\rho = 0.979$). Among the 10 most severe flood events identified by the two approaches, nine events were shared. These results indicate that removing online popularity did not substantially alter the identification of high-severity flood events, while avoiding potential bias caused by differences in information dissemination intensity.

S4.2. Sensitivity analysis of exposure normalization

To evaluate whether differences in city size and socioeconomic scale influenced flood severity assessment, a sensitivity analysis was conducted by comparing the original absolute-indicator specification with the revised exposure-normalized formulation. In the alternative specification, affected area, affected population, and direct economic loss were normalized by built-up area, resident population, and GDP, respectively. The comparison showed a strong consistency between the two approaches, with an event-level Spearman correlation of $\rho = 0.930$.

However, exposure normalization was not equivalent to a simple rescaling operation. Several cities experienced noticeable changes in their rankings, indicating that normalization reduced the influence of city size and altered the relative contribution of different impact dimensions. The attribution analysis further showed that the contribution of affected area decreased after normalization, whereas the relative contribution of fatalities increased.

These results demonstrate that exposure normalization improves comparability among cities with different socioeconomic conditions while preserving the overall spatial pattern of flood severity.

Table S4. Summary of FSI sensitivity analyses.

Sensitivity analysis	Primary statistic	Supporting result	Implication
Four vs five indicators	$\rho = 0.979$	Top-10 overlap = 9/10	Online popularity does not control the broad ranking
Online popularity vs Core FSI	$\rho = -0.013$; $p = 0.654$	No severity gradient across quartiles	Public attention is a separate dimension
Absolute vs exposure-normalized	$\rho = 0.930$	Three cities shifted by >5 ranks	Normalization matters for intercity comparison
Deaths vs death rate	$\rho = 0.9995$	Top-50 overlap = 49/50	Results are robust to mortality definition

S5. Temporal robustness analysis

S5.1 Sensitivity of temporal trends to indicator specification

Because online popularity was excluded from the revised FSI, the temporal robustness of the severity index was evaluated by comparing the four-indicator FSI with the original five-indicator specification.

Table S5. 4-indicator vs 5-indicator FSI Temporal Trends

Metric	4-Indicator FSI	5-Indicator FSI
Annual Median FSI Spearman	—	$\rho = 0.845$ ($p < 0.001$)
Mann-Kendall Trend	Decreasing ($p = 0.042$)	No trend ($p = 0.141$)
Sen Slope	−0.00151 /year	−0.00135 /year

The two specifications showed consistent annual patterns, and highly correlated median FSI sequences. However, the statistical significance of the long-term trend differed, suggesting that the temporal tendency should be interpreted cautiously.

S5.2 Sensitivity to temporal data-source consistency

Table S6. Temporal trend comparison under different observation periods

Metric	Full (2000-2024)	Consistent (2010-2024)
Annual Median FSI	Decreasing ($p = 0.042$)	No trend ($p = 0.767$)
High-Severity %	No trend ($p = 0.870$)	No trend ($p = 0.074$)
Event Count	No trend ($p = 0.072$)	No trend ($p = 0.255$)

The apparent decreasing tendency identified from the full-period analysis was mainly influenced by relatively high-severity events during the early 2000s. When restricted to the period with more consistent data availability, no significant temporal trend was detected.