

Dear Editor,

We sincerely thank you and the reviewers for the careful evaluation of our manuscript entitled “Multi-Data Source Based Quantifying Urban Flood Severity in Major Chinese Cities (2000–2024) Using a Hybrid Machine-Learning Weighting Framework” and for providing constructive and insightful comments.

We have carefully considered all comments and have substantially revised the manuscript accordingly. The revised version has been improved in terms of methodological clarity, data description, model interpretation, and discussion of implications. Specifically, we have further clarified the construction of the multi-source flood-event database, refined the Flood Severity Index (FSI) framework, improved the description of the machine-learning-assisted weighting strategy, and added additional robustness analyses, including sensitivity analyses and independent external validation. We have also revised the Discussion and Conclusion sections to ensure that the interpretations are more directly supported by the proposed framework and quantitative findings.

In addition, all figures, tables, and conceptual diagrams have been carefully revised and regenerated to improve accuracy, consistency, and readability. The revised manuscript now provides a clearer description of dataset sources and applications, strengthens the reproducibility of the methodology, and better highlights the contribution of the proposed framework for comparative urban flood severity assessment. A detailed point-by-point response to all reviewer comments is provided in the accompanying response letter. All changes in the revised manuscript have been highlighted in yellow for clarity.

We believe that these revisions have substantially strengthened the manuscript and addressed the concerns raised during the review process. We sincerely appreciate the valuable suggestions from the reviewers and the opportunity to further improve our work. We hope that the revised manuscript will be suitable for publication in *Natural Hazards and Earth System Sciences*.

Thank you very much for your time and consideration.

Yours sincerely,

Yujing Sun

Affiliation: School of Geography and Tourism, Shaanxi Normal University, No. 620, West Chang'an Street, Chang'an District, Xi'an, Shaanxi Province, China

Email: syj1011@snnu.edu.cn

Response to Referee

Manuscript ID: egusphere-2026-2501

Manuscript title: Multi-Data Source Based Quantifying Urban Flood Severity in Major Chinese Cities (2000–2024) Using a Hybrid Machine-Learning Weighting Framework

Dear anonymous reviewers,

We are grateful to all reviewers for their insightful comments and constructive suggestions, which have been invaluable in improving the quality and clarity of our manuscript. We have carefully addressed all comments and revised the manuscript accordingly. In this response letter, reviewers' comments are presented in black font with a gray background, while our responses are provided in black text. All changes made in the revised manuscript are highlighted in yellow, and the corresponding page and line numbers are provided to indicate the locations of the revisions. Please see detailed responses below.

Response to Referee #1

General comments

The review of this paper is relatively brief as I focus on the main methodological flaws.

Answer:

We appreciate the reviewer's focused and constructive assessment of the manuscript. Although the review primarily concentrates on the major methodological issues, these comments have helped us identify several aspects requiring substantial clarification and improvement. In response, we have carefully revised the manuscript, particularly regarding data collection and processing procedures, FSI construction, machine-learning implementation, validation strategies, and the interpretation of temporal patterns. Detailed responses to each specific comment are provided below, together with corresponding changes made in the revised manuscript.

Comment R1.1: The methodology of collecting data is described much too briefly. It would not be possible to replicate this method.

Answer:

We appreciate the reviewer's comment and agree that the original description of the data-collection procedure was insufficient for replication. The revised manuscript now provides a detailed description of the complete workflow used to construct the urban flood-event database. Specifically, we have added the following information:

(1) Data sources and temporal coverage.

The revised manuscript now explicitly describes the two major data sources used in this study: social-media text data and authoritative documentary sources. Sina Weibo data were collected for the period 2010–2024, while flood events before 2010 were identified mainly from official records, statistical documents, government reports, and archived news sources. The temporal differences in data availability are now explicitly discussed.

(2) Social-media retrieval strategy.

We have added details of the Python-based crawling procedure. Flood-related posts were collected by combining the names of the 20 study cities with flood-related keywords, including “urban waterlogging,” “rainstorm flooding,” “road inundation,” and “drainage congestion.” The collected information included post content, publication time, geographic information, and available user attributes.

(3) Text-processing and event-identification procedure.

The revised manuscript clarifies that Chinese word segmentation and text cleaning were performed using Jieba. We filter out ads by using keywords related to floods and exclusion words., weather forecasts without reported impacts, unrelated discussions, duplicate posts, and repeated reposts. Posts describing the same flood process in the same city and during the same period were consolidated into a single candidate event.

(4) Cross-source verification and event confirmation.

Candidate events extracted from social-media information were further verified using authoritative sources, including government announcements, emergency-management reports, Statistical Yearbooks, Xinhua News Agency, People’s Daily, and China Weather Network. Only confirmed flood events with identifiable occurrence time and location were retained in the final database.

(5) Final database construction.

After data cleaning, duplicate removal, cross-source verification, and event consolidation, the final database contained 1157 documented flood events across the 20 study cities. The database records event occurrence time, location, fatalities, affected area, affected population, direct economic losses, and other relevant attributes.

These revisions provide a clearer and more reproducible description of the database construction process and improve the transparency of the overall research framework.

Changes in the manuscript:

The Data Collection and Processing section has been substantially expanded. The revised manuscript now includes detailed descriptions of data sources, temporal coverage, keyword-based retrieval strategies, Python-based crawling, Jieba-based text

processing, duplicate-event consolidation, cross-source verification, and final database construction. The limitations associated with temporal differences in data availability are also explicitly discussed (please see pages 4–6, lines 92–158).

Comment R1.2: Some of the timelines of the flood severity index are extremely constant. For example, cities like Chengdu, Shenzhen and Guangzhou are affected by the same severity of floods almost every year in the proposed index. This is unlikely to be the case in reality (or the index is too sensitive to some input parameters). However, because the methodology of constructing the index is not described exactly this cannot be checked.

Answer:

We thank the reviewer for pointing out the apparently constant severity levels observed in some cities. We agree that the original manuscript did not sufficiently distinguish between the continuous FSI values and the categorized severity levels, which may have caused confusion regarding the temporal sensitivity of the index.

In the original version, the continuous FSI values were converted into five ordinal categories (Level 0–4) using quantile-based classification for visualization and comparative analysis. Because each category covered a relatively broad range of continuous FSI values, events with different continuous severity scores could be assigned to the same category. Therefore, the apparently constant severity levels observed in cities such as Chengdu, Shenzhen, and Guangzhou mainly reflected the reduced resolution introduced by the categorical classification rather than a lack of variation in the underlying FSI values. In response to this comment, we have refined the severity classification by increasing the number of categories from five levels (0–4) to six levels (0–5) (please see pages 9, lines 215–228). The additional category provides a finer representation of differences among flood events, particularly for cities with relatively frequent moderate-severity events. Furthermore, the revised manuscript now explicitly distinguishes the continuous FSI from the categorized severity levels. The continuous FSI values are used for temporal analysis and statistical evaluation, whereas the six-level classification is used only for visualization, communication, and comparison with independently assessed severity grades (please see pages 18–20, lines 403–462).

We have also revised the methodology section to provide a clearer description of the FSI construction process, normalization procedure, weighting strategy, and classification method. The revised results now present the continuous temporal variation of flood severity together with the categorized levels, reducing the possibility of interpreting the classification results as the original FSI signal (please see pages 6–10, lines 159–255).

These revisions demonstrate that the apparently stable severity levels in some cities were primarily related to the discretization procedure rather than the inability of the FSI framework to capture interannual variations in flood severity.

Changes in the manuscript:

The methodology section has been revised to clarify the construction of continuous FSI values and the subsequent classification procedure. The original five-level severity classification (Level 0–4) has been replaced by a six-level classification (Level 0–5) to improve severity discrimination (please see pages 18–19, lines 403–425). We have also clarified that continuous FSI values, rather than categorical levels, are used for temporal trend analysis and statistical evaluation. The presentation of temporal results has been revised accordingly.

Comment R1.3: The timeline of flood events created for the various cities is based on a number of different sources some of which are not available for part of the timeline, which means that there is a probability that the datasets are biased over time. While this is discussed in the discussion, it should be addressed in the methodology.

Answer:

We thank the reviewer for highlighting the potential temporal bias caused by changes in data-source availability. We agree that the original manuscript did not sufficiently explain this issue in the Methodology section and only discussed it later as a limitation.

In the revised manuscript, we have added a detailed description of the temporal differences in data sources during database construction. Specifically, flood events before 2010 were mainly identified from official records, statistical documents, government reports, and archived news sources, whereas Sina Weibo data were additionally incorporated after 2010 to complement official information and improve the identification of localized flood events.

We acknowledge that this difference in information availability may influence the probability of detecting documented flood events over time. Therefore, we clarify that the database represents documented flood events rather than a complete record of all flood occurrences. Social-media-derived information was not treated as independent event records; instead, candidate events were cross-validated with official records and authoritative reports, and duplicate descriptions of the same flood process were merged (please see pages 9, lines 215–228). In addition, we have revised the temporal analysis to avoid overinterpreting changes around 2010. The previous division of the study period into distinct temporal phases has been removed, and sensitivity analyses

were conducted to evaluate whether the temporal patterns remained consistent under different indicator specifications and observation periods.

These revisions clarify the limitations associated with temporal source coverage and ensure that the interpretation of temporal variation is based on documented flood-event severity rather than assuming a constant observation probability throughout the entire study period.

Changes in the manuscript:

The "2.2 Data Sources and Processing" section has been revised to describe the evolution of data sources over time, the role of Sina Weibo data from 2010 onwards, and the procedures used for cross-source verification and event integration. The revised text clarifies that the database covers recorded flood events and that missing "city-year" records are treated as "missing observations" rather than events with "zero severity" (please see pages 4–6, lines 92–158). Additionally, the analysis regarding the temporal dimension has been modified: predefined time-period categorizations were removed, and sensitivity analyses concerning indicator definition and the selection of the observation period were added.

Comment R1.4: It is unclear how the target data is constructed beyond using a quantile-based approach. However, it seems that this is based on the same data as the training data, and thus it seems that the same values are used to construct the target variables and also used as training, and thus the ML algorithms simply re-discover the linear combination of the FSI. This also would be consistent with the extremely high correlation between the various machine learning models.

Answer:

We thank the reviewer for identifying the potential structural dependence between the predictor variables and the machine-learning targets. We agree that the original manuscript did not describe the target-construction procedure clearly enough and overstated the independent predictive role of the machine-learning models.

In the revised framework, the machine-learning targets are not the quantile-based severity classes. Instead, two continuous reference indices (AHP-FSI and EWM-FSI) are first calculated from the normalized impact indicators using expert-based and entropy-based weighting schemes, respectively. These continuous reference indices are then used as regression targets for the five machine-learning models to evaluate their ability to reconstruct the initial weighting-based severity estimates. After obtaining the final continuous FSI through the hybrid weighting procedure, the six-level severity classes are generated only for descriptive visualization and external ordinal comparison. These categorical severity labels are not involved in model

training, hyperparameter optimization, SHAP attribution, or hybrid-weight estimation (please see pages 7, lines 176–186).

We agree that, because the initial FSI is a deterministic weighted combination of the same indicators used as model inputs, the machine-learning models are not learning an independent ground-truth measure of flood severity. Rather, they are reconstructing the initial AHP- and entropy-based reference indices. This structural dependence also helps explain the high correlations among the outputs of different machine-learning models. We have therefore revised the manuscript to remove statements implying that the models independently validate or predict true flood severity (please see pages 8–9, lines 186–214).

The machine-learning component is now interpreted as a data-driven refinement and consistency analysis of the initial weighting schemes. Specifically, the models are used to evaluate whether different algorithms can consistently reproduce the initial AHP- and entropy-based FSI values under out-of-sample conditions and whether they provide stable estimates of the relative importance of different impact indicators. The SHAP-derived indicator contributions are then combined with the initial AHP or entropy weights to obtain the final hybrid weights. Therefore, the machine-learning models are not intended to independently discover the true indicator weights, but rather to provide additional data-driven information for refining the initial expert- and statistical-based weighting schemes.

To reduce in-sample explanation bias, the revised analysis uses nested cross-validation and out-of-fold SHAP values (please see pages 12–13, lines 294–314). The outer 10-fold cross-validation loop is used for out-of-sample reconstruction assessment, while hyperparameters are optimized using five-fold cross-validation within each outer training partition. SHAP values for each event are calculated only from a model that was not trained or tuned using that event. To further address the concern regarding potential circularity between the constructed FSI and model evaluation, we added an independent external validation procedure in the revised manuscript. A subset of 53 events was randomly sampled from the 1,157 documented flood events, covering cities from both northern and southern regions with relatively balanced spatial distribution. For these events, independent six-level severity grades were compiled from authoritative disaster-response classifications and official emergency-management records. These external assessments were completely excluded from FSI construction, machine-learning training, hyperparameter optimization, SHAP attribution, hybrid-weight estimation, and severity-threshold determination. They were used only after the final FSI was obtained to evaluate external consistency through rank correlation, weighted Cohen's kappa, exact agreement, and within-one-level agreement statistics. This additional analysis provides an independent assessment of whether the proposed FSI framework can reproduce externally documented differences in flood severity (please see pages 14, lines 318–336).

Based on the reviewer's comment, we have revised the interpretation of the high cross-model correlations. These correlations are now presented as evidence of reconstruction and attribution stability conditional on the initial AHP- and entropy-based indices, rather than as independent validation of the FSI.

Comment R1.5: The data and methods section mention several datasets that don't seem to be used in the manuscript (although the DEM may be used for visualization in a figure). Other datasets used in the methodology are not mentioned.

Answer:

Thank you for pointing out this inconsistency. We agree that the original description of datasets in the Data and Methods section did not clearly distinguish between datasets used for FSI construction, sensitivity analyses, external validation, and visualization purposes.

We have revised the data description section by adding the previously omitted socioeconomic, urban morphology, and validation datasets. Specifically, annual population data (2000–2024) from the LandScan dataset, prefecture-level GDP data from official statistical yearbooks, and built-up area data from the China Urban Statistical Yearbook have been added and their roles in exposure-normalization sensitivity analyses have been clarified. In addition, precipitation information, including city-wide mean rainfall and maximum accumulated rainfall for the independently validated flood events, has been added, with data sources specified as the China Meteorological Data Service Centre and hydrological yearbooks.

Furthermore, we clarified that DEM data were only used for background visualization and topographic description and were not included as input variables in FSI calculation or the machine-learning framework. These revisions ensure that all datasets mentioned in the manuscript have clearly defined purposes and are consistent with the subsequent analyses.

Changes in the manuscript:

The Data and Methods section has been revised to provide a clearer description of dataset sources and applications. Additional socioeconomic datasets (LandScan population, prefecture-level GDP, and built-up area data) and validation datasets (rainfall information from the China Meteorological Data Service Centre and hydrological yearbooks) have been incorporated. The role of DEM data has also been clarified, indicating that it was only used for background visualization and topographic description and was not included in the FSI calculation or machine-learning framework. The revised manuscript now clearly distinguishes datasets used for FSI construction, sensitivity analysis, validation, and visualization

(please see pages 5, lines 112–131, the “2.2.2 Additional socioeconomic, urban morphology, and validation datasets” section).

Comment R1.6: Hydrological or climatological data seem key in predicting flood severity but they don’t seem to be used in the methods.

Answer:

We thank the reviewer for highlighting the importance of hydrological and climatological information in the analysis of urban flood severity. We agree that rainfall conditions provide essential contextual information for assessing whether the constructed FSI is consistent with the physical intensity of flood-producing events.

The purpose of the core FSI in this study is to quantify the realized consequences of documented flood events rather than to predict flood occurrence or hazard intensity directly. Therefore, the index is constructed from direct impact indicators, including affected population, affected area, fatalities, and direct economic losses. Rainfall variables were not incorporated into the core FSI because they represent hazard-driving conditions rather than disaster consequences. Including both hazard intensity and impact outcomes in the same index would make the interpretation of the FSI less clear and could introduce conceptual overlap between the causes and consequences of flooding.

Nevertheless, in response to the reviewer’s comment, we have strengthened the role of climatological information in two ways. First, city-wide mean rainfall and maximum cumulative rainfall were added to the external evaluation dataset. These variables were used to assess whether events with higher FSI values were generally associated with more intense rainfall conditions. Their relationships with the continuous FSI were evaluated using rank-correlation analysis, providing an external plausibility assessment independent of the indicators used to construct the index (please see pages 5, lines 112–131).

We also expanded the Discussion to clarify the broader role of rainfall and urban environmental conditions in shaping flood impacts. Preliminary results from our related follow-up research indicate that flood frequency and severity generally increase with rainfall, and that precipitation during the main flood season from May to August provides a more stable explanatory signal than annual precipitation. The follow-up analysis also suggests that the estimated association between urban water-body loss and increased flood risk remains relatively stable after controlling for either annual or flood-season rainfall. These findings are discussed only as complementary contextual evidence and are not treated as part of the formal validation or causal analysis of the present study (please see pages 26, lines 557–584).

Accordingly, the revised manuscript now distinguishes clearly among hazard-driving variables, impact indicators, and contextual explanatory factors. We have also clarified that the present FSI framework is an impact-severity assessment framework rather than a hydrological flood-prediction model.

Changes in the manuscript:

The revised Methods section now explains why rainfall variables were not included in the core impact-based FSI. City-wide mean rainfall and maximum cumulative rainfall have been added to the external evaluation dataset and are assessed using rank-correlation analysis (please see pages 8, lines 206–214, the “3.5 Hybrid Weight Construction” section). The Discussion has also been expanded to explain the potential roles of rainfall seasonality and urban water-body change in shaping flood severity, while clearly identifying these findings as preliminary evidence from related follow-up research rather than results generated by the present FSI model (please see pages 24–26, lines 510–584, the “Discussion” section).

Comment R1.7: The discussion and especially conclusions make very general points that are mostly based on existing research rather than the described methodology.

Answer:

Thank you for this important comment. We agree that some statements in the original Discussion and Conclusion sections were overly general and were not sufficiently connected to the specific findings generated by the proposed framework.

In the revised manuscript, we have substantially reorganized and refined these sections to ensure that the interpretations are directly supported by the event-level Flood Severity Index (FSI) and the analyses conducted in this study. Specifically, the Discussion has been revised to focus on the spatial heterogeneity and temporal variability identified from the 20-city FSI dataset, rather than providing general descriptions of urban flooding mechanisms. The revised discussion now interprets the observed north–south differentiation, city-specific severity structures, and management implications based on the severity patterns revealed by the proposed framework. For example, the differences between southern cities with persistent high-severity impacts and northern cities with lower baseline severity but episodic extreme events are discussed according to the results of the FSI-based comparison. In addition, the limitations and future research section has been strengthened to clarify the scope of the current framework. We explicitly state that this study focuses on evaluating documented flood-event severity and identifying spatial–temporal patterns rather than establishing causal relationships. Future directions, including integrating dynamic environmental variables (e.g., urban water-body changes, land-use transitions, and hydroclimatic variations) and coupling with hydrological and

hydrodynamic simulations, are proposed based on the limitations of the current methodology (please see pages 24–26, lines 510–584).

The Conclusion section has also been revised to summarize the quantitative findings derived from the proposed framework, including (1) spatial heterogeneity of documented flood severity between southern and northern cities, (2) differences in severity structures among individual cities, and (3) substantial interannual variability without a consistent long-term trend. These revisions ensure that the Discussion and Conclusion are more closely linked to the methodology and results of this study (please see pages 26–27, lines 585–603).

Changes in the manuscript:

The Discussion and Conclusion sections have been revised accordingly.

The Discussion was reorganized around the findings derived from the event-level FSI framework, including spatial differentiation, city-level severity characteristics, management implications, and methodological limitations. General statements that were not directly supported by the analysis were moderated, and interpretations were revised to emphasize associations rather than causal conclusions (please see pages 24–26, lines 510–584).

The Conclusion was rewritten to summarize the major findings obtained from the proposed framework, including spatial heterogeneity, temporal variability, and the contribution of the FSI-based comparative assessment approach (please see pages 26–27, lines 585–603).

Response to Referee #2

General comments:

In this manuscript, Peng et al. tackle the complex challenge of assessing urban flood risks across diverse geographical and developmental landscapes. The authors attempt to bridge the gap between fragmented historical records and modern datasets by integrating official statistics, news reports and social media into an event-level database. While the study's objective to construct an interpretable, data-driven Flood Severity Index is timely and highly relevant to urban resilience planning, the execution of this framework warrants careful examination. Specifically, a number of methodological uncertainties in the research design and a lack of granular transparency in the methodology undermine the robustness of the study's otherwise compelling spatial and temporal claims.

Answer:

We sincerely thank the reviewer for the balanced assessment of the study's objectives and for identifying the methodological and reporting issues that limited the robustness and reproducibility of the original manuscript. We agree that several components of the initial submission were described at an overly general level and that some spatial and temporal interpretations were stronger than could be supported by the available event-level data.

In response, we have substantially revised the manuscript and strengthened the research framework in several respects. First, the construction of the multi-source flood-event database is now described in detail, including the source coverage for different periods, event-identification criteria, duplicate-event consolidation, missing-value treatment, and the number of records affected by imputation. The social-media processing workflow has also been clarified by specifying the Python-based data collection procedure, Jieba word segmentation, rule-based keyword and exclusion-term screening, duplicate removal, event consolidation, and manual verification against official and authoritative sources.

Second, the FSI framework has been revised to improve its conceptual consistency. The core index now contains four direct-impact indicators representing affected population, affected area, fatalities, and direct economic losses. Online popularity has been removed from the core FSI and is analyzed separately as an auxiliary indicator of public attention. Exposure-related variables are normalized using corresponding city-year population, built-up area, and corresponding city-year GDP data. We also added sensitivity analyses comparing absolute and population-normalized fatalities and comparing the revised four-indicator FSI with the original five-indicator specification.

Third, the machine-learning component has been rewritten to clarify its analytical role and implementation. The models are now explicitly described as reconstructing and refining the initial AHP- and entropy-weighted reference indices rather than independently predicting an observed ground-truth measure of flood severity. Model training uses nested cross-validation, with a 10-fold outer loop for out-of-sample evaluation and a 5-fold inner loop for hyperparameter optimization. Fold-specific preprocessing, complete parameter search spaces, selected model settings, random seeds, and outer-fold R^2 , RMSE, and MAE values are now reported. SHAP values are calculated using an out-of-fold procedure so that each event is explained by a model that did not use that event for training or tuning.

Fourth, the external evaluation procedure has been strengthened. The independently assessed six-level severity labels are used only for post hoc evaluation and are not involved in model training, weight estimation, hyperparameter selection, or threshold determination. The revised manuscript reports rank correlation, weighted Cohen's kappa, exact agreement, within-one-level agreement, and confusion-matrix results.

Finally, the spatial and temporal interpretations have been revised more cautiously. The previous division of the study period into three distinct evolutionary phases has been removed because the temporal pattern was sensitive to indicator specification and observation-period selection. The revised manuscript describes interannual variations in documented flood-event severity and explicitly acknowledges the effects of changing source coverage and event composition. Regional differences are now presented as descriptive spatial patterns rather than as established causal relationships.

These revisions have improved the transparency, reproducibility, and interpretive scope of the study. Detailed responses to each methodological concern and the corresponding manuscript changes are provided in the point-by-point responses below.

In the methodology section, which frequently states what was done without adequately explaining how it was achieved, a thorough revision is necessary for the study to be truly replicable. For instance, the authors claim to resolve missing disaster attributes such as economic losses and affected populations by imputing them “using statistical methods”, yet they fail to specify the actual techniques employed, leaving the foundational dataset vulnerable to undisclosed biases. Similarly, the study heavily relies on a Python-based web crawler and NLP to filter 14 years of Sina Weibo posts, but the manuscript omits the specific algorithms or semantic recognition tools used to isolate valid flood events from the noise. Furthermore, while the authors deploy six different ML algorithms to build their hybrid weighting system, the text lacks critical details regarding hyperparameter tuning or dataset splitting beyond standard

cross-validation. Again, this vagueness prevents replicability and makes it difficult to rigorously verify if the authors successfully implemented the data-driven framework they claim.

Answer:

We thank the reviewer for emphasizing the need for greater methodological transparency and reproducibility. We agree that the original manuscript described several analytical procedures too generally and did not provide sufficient implementation details. In response, the Data and Methods section has been thoroughly revised to clarify the complete workflow for missing-value treatment, social-media text processing, model training, hyperparameter optimization, data partitioning, and SHAP-based attribution.

First, we now explicitly describe the missing-data treatment. Missing attributes of confirmed flood events were imputed using the global median of the corresponding indicator. Median imputation was selected because the disaster-impact variables were strongly right-skewed and contained extreme observations, making the median less sensitive to exceptionally severe events than the arithmetic mean. Imputation was applied only to missing attributes of documented events and was not used to create events for years or cities without event records. Documented zero-fatality observations were retained as observed zero values rather than treated as missing. The revised manuscript also reports the number and proportion of records containing missing attributes (please see pages 6, lines 147–155).

Second, the social-media collection and text-processing workflow has been substantially expanded. Sina Weibo posts from 2010–2024 were collected using a Python-based crawler with combinations of the names of the 20 study cities and flood-related keywords, including “urban waterlogging,” “rainstorm flooding,” “road inundation,” and “drainage congestion.” Text cleaning and Chinese word segmentation were performed using the Jieba library. A rule-based screening procedure based on flood-related keywords and exclusion terms was then used to remove advertisements, weather forecasts without reported flood impacts, unrelated discussions, duplicate posts, and repeated reposts. Posts referring to the same flood process in the same city and during the same period were consolidated into a single candidate event. The remaining candidate events were manually verified against government bulletins, statistical records, and authoritative news reports to confirm their occurrence time, location, and reported consequences (please see pages 4–5, lines 93–131).

Third, the model-development procedure has been rewritten to provide full details on hyperparameter tuning, optimization, preprocessing, and data partitioning. The five regression models were evaluated using nested cross-validation. The outer loop used 10-fold cross-validation to estimate out-of-sample performance, while hyperparameter optimization was conducted exclusively within each outer training partition using

five-fold GridSearchCV. The parameter combination producing the highest mean inner-validation R^2 was selected, after which the model was refitted on the complete outer training partition and evaluated on the held-out outer test fold. Model performance is now reported using the mean and standard deviation of the outer-fold R^2 , RMSE, and MAE (please see pages 11–12, lines 272–293).

Complete hyperparameter search spaces, selected parameter values, model-specific optimization settings, stopping criteria, preprocessing procedures, and random seeds are now reported in the revised manuscript and Supplementary Material. In addition, SHAP values were calculated using an out-of-fold procedure, so that each event was explained by a model that had not used that event for either training or hyperparameter selection (please see pages 13, lines 300–318).

These revisions provide a transparent and reproducible description of the complete data-processing and model-development framework. We have also clarified that the machine-learning models were used to assess the reconstruction and attribution stability of the initial AHP- and EWM-based indices, rather than to independently predict a ground-truth flood-severity label.

The authors identify the period from 2010-2020 as a “transitional phase of adjustment” in urban flood severity and attribute this shift to changes in infrastructure and flood management. However, the underlying flood database undergoes a major methodological change at approximately the same time. Prior to 2010, the database relies primarily on official records and documentary sources, whereas from 2010 onward it additionally incorporates large volumes of social media data from Sina Weibo. This change in data availability and reporting intensity may substantially alter event detection rates, event characterization, and the online-attention component of the FSI itself. Consequently, the apparent temporal transition identified by the authors may partly reflect an observational artifact rather than a genuine shift in flood severity. The manuscript should explicitly evaluate the sensitivity of the results to this data-source discontinuity and demonstrate that the observed temporal phases remain robust when social-media-derived information is excluded or otherwise standardized across the study period.

Answer:

We thank the reviewer for highlighting the potential discontinuity in event detection and reporting associated with the expansion of data sources after 2010. We agree that the original division of the study period into phases of “volatility,” “transitional adjustment,” and “moderation” was stronger than could be supported by the event-level data, particularly because the proposed transition around 2010 coincided with the introduction of Sina Weibo as an additional information source.

To evaluate this issue, we conducted two sensitivity analyses. First, we compared

the annual median FSI derived from the revised four-indicator core specification with that derived from the original five-indicator specification including online popularity. The two annual series were strongly correlated ($\rho=0.845$, $p<0.0001$). However, the four-indicator FSI showed a weak downward trend over 2000–2024 ($p=0.040$; Sen's slope $=-0.0015\text{yr}^{-1}$), whereas the five-indicator specification produced a similar slope (-0.0014 yr^{-1}) but no statistically significant trend ($p=0.141$). Second, we restricted the analysis to 2010–2024, during which the principal official, news, and social-media sources were more consistently available. No significant trend was detected in the annual median FSI during this period ($p=0.767$). The proportion of high-severity events and the number of documented events also showed no statistically significant temporal trends in either the complete or restricted periods. These results indicate that the previously proposed three-stage temporal structure was not robust to indicator specification and observation-period selection. We have therefore removed the division of 2000–2024 into distinct phases (Please see Supplementary Material S5, “Temporal robustness analysis” section). The revised manuscript now describes the findings as interannual variation in the severity distribution of documented flood events rather than as evidence of structural temporal transitions. We have also removed direct causal statements attributing the temporal pattern to sponge-city construction, drainage-system improvement, or changes in emergency-management capacity (please see pages 21–24, lines 462–509).

We further clarify that the absence of a documented event in a given city and year was not treated as a zero-severity observation, because the analysis was conducted at the event level. The revised discussion explicitly acknowledges that temporal patterns may be affected by differences in source coverage, event composition, observation-period length, and the occurrence of individual extreme events.

The FSI is constructed using absolute values of affected area, affected population economic losses, and fatalities. However, the study compares cities that differ substantially in population size, urban extent, and economic activity. Larger cities will naturally tend to report greater numbers of affected people and larger economic losses even when the relative severity of flooding is comparable. The manuscript does not explain whether these indicators were normalized by population, urban area, GDP, or other exposure metrics prior to index construction. Min-max normalization alone does not resolve the issue, as it rescales variables without accounting for differences in underlying exposure. As a result, it is difficult to determine whether the resulting rankings reflect flood severity or simply differences in city size and socioeconomic scale. The authors should justify the use of absolute indicators or evaluate the sensitivity of their results to appropriate normalization procedures.

Answer:

We thank the reviewer for highlighting the influence of differences in population size, urban extent, and economic scale on cross-city comparisons. We agree that min–max normalization alone standardizes the numerical range of the indicators but does not control for differences in underlying exposure.

In response, the principal FSI has been reconstructed using exposure-normalized indicators. Specifically, affected area was divided by the built-up area of the corresponding city, affected population was divided by the annual resident population, and direct economic loss was divided by the annual city GDP (please see pages 5, lines 112–131). The denominators were matched to the city and year in which each flood event occurred. These revised indicators therefore represent the proportion of urban area affected, the proportion of the resident population affected, and the economic loss relative to the size of the local economy (please see pages 6–7, lines 160–176).

Furthermore, an additional sensitivity analysis based on these exposure-adjusted indicators has been provided in the Supplementary Material (Section S1). This analysis evaluates whether the severity ranking and overall spatial patterns are sensitive to differences in urban exposure, population scale, and economic capacity. The results show that the exposure-adjusted specification produces highly consistent rankings with the original FSI ranking (Spearman's $\rho = 0.930$), indicating that the main conclusions are robust and are not primarily driven by variations in city size or socioeconomic scale (Please see Supplementary Material S1, “Sensitivity analysis using exposure-adjusted indicators” section)

Fatalities were retained as an absolute indicator in the principal specification. Unlike affected population and economic loss, fatalities represent irreversible human losses, for which the absolute number has a direct substantive meaning in disaster-severity assessment. Nevertheless, we acknowledge that absolute fatalities may also be affected by differences in population exposure. We therefore constructed an alternative specification in which fatalities were normalized by the corresponding annual city population and repeated the complete FSI calculation. The sensitivity analysis showed that the treatment of fatalities had a negligible influence on the results. The event-level rankings from the absolute-fatality and population-normalized specifications were almost identical, with a Spearman correlation of $\rho=0.9995$. The mean absolute rank change was only 3.3 positions among the 1157 events, and 49 of the top 50 events were common to both specifications. At the city level, the ranking correlation was $\rho=0.994$, and all five highest-ranked cities remained unchanged. The annual temporal patterns were also identical in rank order (Please see Supplementary Material S2, “Sensitivity analysis of fatality indicator representation” section).

These findings demonstrate that the main event-level rankings, city-level spatial patterns, and temporal conclusions are robust to the treatment of fatalities. We therefore retained absolute fatalities in the principal FSI as a direct measure of

irreversible human loss, while reporting the population-normalized fatality specification as a sensitivity analysis.

Finally, the inclusion of online popularity as one of the five core components of the FSI requires substantially stronger theoretical justification. Unlike affected area, affected population, economic losses, and fatalities, online popularity is not a direct measure of flood impacts but rather a measure of public attention and information dissemination. Incorporating online popularity directly into the severity index risks conflating disaster impacts with reporting behaviour. The authors should clearly justify why public attention is treated as a component of flood severity rather than as an auxiliary explanatory variable and should evaluate how sensitive the resulting severity rankings are to the inclusion or exclusion of this indicator.

Answer:

We thank the reviewer for highlighting the conceptual distinction between direct flood impacts and online public attention. We agree that affected area, affected population, direct economic loss, and fatalities represent the realized spatial, demographic, economic, and human consequences of flooding, whereas online popularity primarily reflects public attention, information dissemination, and reporting behaviour.

This distinction is also consistent with the disaster terminology of the United Nations Office for Disaster Risk Reduction (UNDRR). UNDRR defines a disaster as a serious disruption resulting from the interaction of hazardous events with conditions of exposure, vulnerability, and capacity, leading to human, material, economic, and environmental losses and impacts. Based on this definition, the core measurement of disaster severity should primarily reflect realized losses and impacts. Online popularity does not directly represent any of these impact dimensions and may instead be influenced by media exposure, platform use, public interest, and reporting intensity.

In response to this comment, online popularity has been removed from the core Flood Severity Index. The revised principal FSI is constructed using four direct impact indicators: affected-area ratio, affected-population ratio, relative direct economic loss, and population fatalities. These indicators correspond respectively to the spatial, demographic, economic, and human dimensions of realized flood impacts. Online popularity is now treated as an auxiliary indicator of public attention rather than as a direct component of flood-impact severity.

We have recalculated the AHP, EWM, machine-learning attribution, hybrid weights, and final FSI using the four direct impact indicators. In addition, a sensitivity analysis was conducted by comparing the revised four-indicator FSI with the original

five-indicator specification that included online popularity. The FSI ranking is robust to the exclusion of online popularity ($\rho = 0.979$), with 9 of 10 most severe events unchanged. However, events with disproportionately high online attention relative to their physical impact experience large rank increases when online popularity is included—precisely the conflation of disaster impacts with reporting behavior that warrants its removal from the core FSI. Although the inclusion of online popularity noticeably affected the principal rankings and temporal conclusions, the four-indicator specification was retained as the principal FSI because it provides a clearer conceptual representation of direct flood impacts and is more consistent with the UNDRR disaster-impact framework (Please see Supplementary Material S4, “Sensitivity analysis of online popularity exclusion” section).

Specific comments:

The authors state that missing values for crucial impact metrics (e.g., economic losses, affected population) were “imputed using statistical methods”. They need to explicitly identify which methods were used (e.g., mean imputation, multiple imputation, KNN), as imputing extreme variables can severely skew the dataset.

Answer:

We thank the reviewer for pointing out that the missing-value treatment was insufficiently described in the original manuscript. We have now clarified that missing values in affected population and direct economic loss were imputed using median imputation.

Median imputation was selected because these disaster-impact variables exhibited strongly right-skewed and long-tailed distributions, with a small number of extreme flood events producing exceptionally large values. Under such conditions, mean imputation may be disproportionately influenced by extreme observations and may overestimate the typical value of missing records. The median provides a more robust measure of central tendency and therefore reduces the influence of extreme events during imputation. Fatality values of zero were not imputed values; they represented documented events with no reported deaths. No artificial flood events were generated through the imputation procedure. We have added the specific imputation method, its rationale, and the number and proportion of imputed observations for each indicator to the revised manuscript.

Only 15 of the 1157 event records contained missing values, corresponding to approximately 1.3% of the dataset. Because the missing proportion was very low, a global median was used rather than city-specific or period-specific medians, which could be unstable for groups containing relatively few observations. Median imputation was selected because the affected-population and economic-loss variables

exhibited strongly right-skewed and long-tailed distributions. Compared with mean imputation, the median is less sensitive to a small number of extremely severe events and therefore provides a more robust estimate of the central tendency (please see pages 6, lines 147–155).

The imputation procedure was applied only to missing attributes of confirmed flood events and did not generate any artificial events. We have added the specific method, the number and proportion of imputed records, and the rationale for selecting global median imputation to the revised manuscript.

The authors mention applying NLP techniques for semantic recognition and noise reduction. However, it completely omits the specific algorithms, models, or libraries used to achieve this, making the data cleaning process, again, irreproducible.

Answer:

We thank the reviewer for pointing out that the original manuscript did not provide sufficient details on the natural language processing and text-cleaning procedures. We agree that the previous description, particularly the terms “semantic recognition” and “noise reduction,” was too general to ensure methodological reproducibility.

In the revised manuscript, we have clarified that no pretrained language model or supervised text-classification algorithm was used. Instead, social-media posts were processed using a rule-based natural language processing workflow implemented in Python. Flood-related public posts were collected from Sina Weibo for the period 2010–2024 using combinations of the names of the 20 study cities and flood-related search terms, including “urban waterlogging”, “rainstorm flooding”, “road inundation”. The collected fields included post content, publication time, available geographic information, and user-related attributes. The collected text was first cleaned and segmented using the Jieba Chinese word-segmentation library. Posts were then screened using predefined flood-related keywords and exclusion terms. Advertisements, weather forecasts that did not report actual flood impacts, unrelated discussions, duplicate posts, and repeated reposts were removed. Multiple posts referring to the same flood process in the same city and during the same period were consolidated into a single candidate event. The candidate events were subsequently checked manually against official government bulletins and authoritative news reports. This verification step was used to confirm the event occurrence time, location, and reported consequences. Official sources included Statistical Yearbooks, Xinhua News Agency, People’s Daily, China Weather Network, and reports issued by local meteorological, flood-control, and emergency-management departments. These sources were also used to supplement information on affected population, affected area, fatalities, and direct economic losses (please see pages 4–5, lines 92–111).

For the period 2000–2009, when Sina Weibo data were unavailable, flood events were identified primarily from official records and archived news reports. Following text screening, duplicate removal, event consolidation, cross-source verification, and manual quality control, the final database contained 1,157 urban flood-event records.

The revised manuscript now explicitly describes the text-processing library, keyword-based screening procedure, exclusion criteria, event-merging rules, and manual verification process.

While six machine learning models are evaluated, the manuscript lacks critical details regarding hyperparameter tuning, optimization strategies, and train/test splitting beyond a standard 10-fold cross-validation.

Answer:

Following additional analysis, we removed the Support Vector Machine (SVM) model from the revised framework because its performance was consistently lower than that of the other representative models and its inclusion did not provide additional information for indicator contribution analysis. The revised model set retains representative algorithms from ensemble learning, gradient boosting, and neural network architectures, including RF, GBDT, XGBoost, LightGBM, and MLP.

A three-part evaluation framework was used, consisting of nested cross-validation, out-of-fold feature attribution, and external validity assessment. First, all five regression models were trained and evaluated using nested cross-validation. The outer loop comprised 10 folds and was used to estimate out-of-sample performance. In each outer iteration, nine folds formed the outer training partition, while the remaining fold was retained as an unseen test partition. Hyperparameter optimization was conducted exclusively within the outer training partition using five-fold GridSearchCV, with mean validation R^2 as the selection criterion. The selected model was refitted on the complete outer training partition and evaluated on the held-out fold using R^2 , RMSE, and MAE. Final performance was reported as the mean and standard deviation across the ten outer test folds (please see pages 8, lines 186–205).

Second, SHAP attribution was implemented using an out-of-fold procedure. In each outer iteration, the model selected through the inner cross-validation loop was used to calculate predictions and SHAP values for the held-out outer fold. Consequently, each event received one set of SHAP values from a model that had not used that event for training or hyperparameter selection. This procedure was intended to reduce in-sample explanation bias and assess the stability of indicator-attribution patterns when reconstructing the initial AHP- and EWM-based reference indices. The

SHAP values were not interpreted as independently estimated causal effects or objectively true flood-severity weights (please see pages 11–14, lines 272–317).

Third, the final FSI was compared with external six-level severity information available for 53 independently assessed flood events. These external labels were not used in model training, hyperparameter optimization, SHAP attribution, weight refinement, selection of the blending coefficient, or determination of the FSI classification thresholds (please see Supplementary Material S3, “Nested cross-validation and hyperparameter optimization” section). Spearman rank correlation was used to assess the relationship between the continuous FSI and the external ordinal labels. Agreement between the two six-level classifications was additionally evaluated using weighted Cohen’s kappa, exact agreement, and within-one-level agreement (please see pages 9, lines 215–228).

Given the available modeling sample of $N=1157$, no separate fixed holdout set was created. Nested cross-validation allowed every event to serve once as an unseen outer-test observation while preserving the maximum feasible sample size for model development. A fixed random seed of 42 was used for all data partitions, and identical folds were applied across the models and the AHP- and EWM-based branches.

In the main text the FSI is correctly defined as a linear combination of five indicators, however in the flowchart (Fig. 2) the formula is incorrectly typed as a summation of only three variables with incorrect subscripts. The Figure must be corrected to accurately reflect the five variables discussed in the text.

Answer:

Thank you for pointing out this inconsistency. We agree that the formula presented in the original Figure 2 was incorrectly typed and did not accurately correspond to the FSI formulation described in the main text. This was a graphical error in the workflow illustration rather than a methodological inconsistency.

To address this issue, we have completely redrawn Figure 2 and revised the workflow diagram. Specifically, the incorrect mathematical expression has been removed from the figure, and the updated workflow now focuses on clearly illustrating the major procedures of the proposed framework, including multi-source flood-event database construction, indicator extraction, FSI development, hybrid weighting, machine-learning-assisted analysis, and subsequent spatial-temporal assessment. In addition, all indicator definitions, model components, and workflow connections have been carefully checked to ensure consistency with the descriptions in the Methods section. The revised Figure 2 now provides an accurate and clearer representation of the overall framework (Please see the revised Figure 2).

There is a typo in the indexing for the AHP index calculation. The summation index is defined as $i = 1$, but the variables inside the summation use the subscript l . This should be corrected so the subscripts match the index (i.e. using i throughout).

Answer:

Thank you for identifying this notation inconsistency. We agree that the summation index in the AHP index calculation was incorrectly defined due to a typographical error, where the summation used index i while the variables inside the equation were labeled with subscript l .

We have corrected this equation by replacing the inconsistent subscript l with i throughout the formulation to ensure that the summation index and variable notation are fully consistent. We have also carefully checked the related equations and symbols in the revised manuscript to avoid similar notation inconsistencies (Please see the revised formula 1-11).

The manuscript provides conflicting descriptions of the role of ML. Section 3.3 defines continuous and categorical FSI variables for supervised learning, while Section 3.5 states that machine learning is not used to predict predefined labels but only to derive SHAP-based weights. The authors should explicitly clarify the target variable(s) used during model training and explain how SHAP values were derived without introducing dependence on the originally constructed FSI.

Answer:

We thank the reviewer for identifying the inconsistent descriptions of the machine-learning targets and the role of SHAP in the original manuscript. We agree that Sections 3.3 and 3.5 did not clearly distinguish the initial continuous indices, the machine-learning attribution procedure, and the final ordinal severity classification.

In the revised analysis, only two continuous reference targets were used during model fitting: the initial AHP-based FSI and the initial EWM-based FSI. The exposure-normalized direct flood-impact indicators were used as model inputs. The six-level severity classes were not used as supervised-learning targets; they were derived only after the final continuous hybrid FSI had been calculated and were used for descriptive and external ordinal comparison. We also acknowledge that the initial AHP- and EWM-based indices are constructed from the same indicators used as model inputs. Therefore, the resulting SHAP values are necessarily conditional on the initial reference indices and cannot provide independent validation or independently identify objectively true severity weights. We have revised the interpretation of the ML component accordingly. It is now described as an internal data-consistency and attribution-stability assessment, followed by constrained model-assisted refinement of

the AHP/EWM prior weights (Please see the revised methodology “3 Methods” section, Please see pages 6–9, lines 160–228).

To reduce in-sample explanation bias, SHAP values were recalculated using an out-of-fold procedure. Within 10-fold cross-validation, each model was fitted using nine folds, and predictions and SHAP values were generated for the held-out fold. Thus, every event-level SHAP value was obtained from a model that had not been trained on that event. Mean absolute out-of-fold SHAP values were normalized within each model before cross-model comparison and aggregation (Please see pages 11–14, lines 256–317).

The external validity of the final FSI was assessed separately using the six-level external reference information, which was not used in model training, SHAP calculation, weight refinement, blending-parameter selection, or FSI threshold determination. We have therefore separated internal reconstruction consistency from external plausibility assessment throughout the revised manuscript.

Final response

We sincerely thank both anonymous reviewers for their constructive, and valuable comments throughout the review process. The reviewers’ careful evaluations and methodological suggestions have significantly helped us identify several limitations in the original manuscript and improve the scientific rigor, clarity, and reproducibility of this study.

In response to the reviewers’ comments, we have conducted a substantial revision of the manuscript. Major improvements include a more detailed description of the multi-source flood-event database construction, refinement of the Flood Severity Index framework, clarification of the role of machine-learning models, implementation of additional robustness analyses and independent external validation, and a more cautious interpretation of spatial-temporal patterns and research implications. In addition, the Discussion and Conclusion sections have been thoroughly reorganized to ensure that the interpretations are directly supported by the proposed framework and quantitative findings.

Furthermore, all figures, tables, and schematic diagrams have been carefully revised and regenerated to improve accuracy, consistency, and readability. The research framework, workflow diagrams, model interpretation figures, and supplementary analyses have been updated accordingly to better reflect the revised methodology and results.