Dear Editor and Reviewers,

We sincerely appreciate the time and effort that the editor and reviewers have devoted to reviewing our manuscript. We are truly grateful for your insightful comments and constructive suggestions, which have greatly contributed to improving the overall quality, clarity, and rigor of our work. We have carefully addressed each point raised and revised the manuscript accordingly. Below, we provide detailed, point-by-point responses to all reviewer comments, along with explanations of the modifications made to the manuscript.

Thank you once again for your valuable feedback and continued support.

Editor Comment

a) Figure 2 may contain a territory that is disputed according to the United Nations (small map in the upper left corner).

Response: Thank you for pointing this out. We have replaced the map to address this issue. The upper-left inset map in Figure 2 has been changed: it now shows the location of the Yangtze River Delta Urban Agglomeration (YRDUA) within the Euro-Asia continent, without including any disputed territories. The position of the YRDUA is indicated using a red five-pointed star symbol, ensuring both clarity and compliance with the journal's map policies.

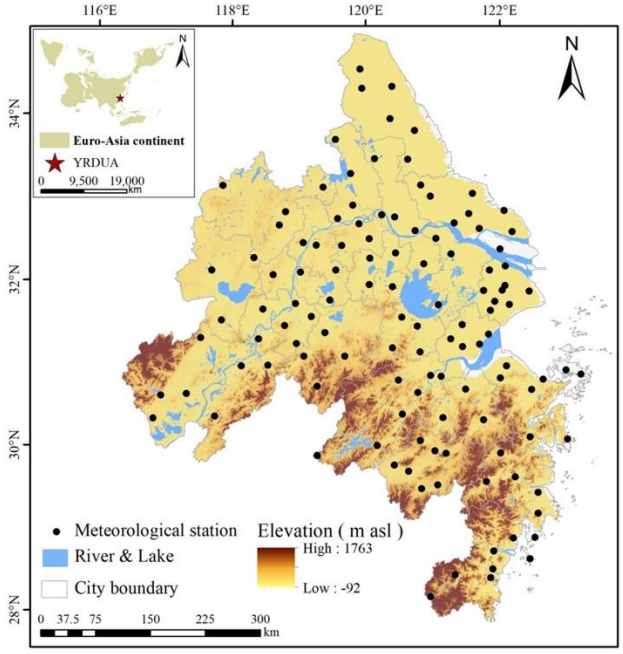

Figure 1: The schematic map of the YRDUA.

b) Section "Author contribution": please use initials for the authors' names.

Response: Thank you for your suggestion. We have revised the Author Contributions section to use initials for all authors' names, in accordance with the journal's formatting requirements.

The revised author contributions are as follows:

YG: Writing - original draft preparation, Validation, Software, Methodology, Conceptualization **HL:** Writing-review & editing, Visualization, Supervision, Formal analysis. **YZ:** Methodology, Formal analysis. **HJ:** Writing - review & editing, Methodology. **SW:** Software, Formal analysis. **YG:** Visualization, Software. **SZ:** Writing-review & editing, Resources, Project administration, Funding acquisition, Conceptualization .

Reply on RC1

1. Figure 2. Please change m to m asl, as I suggested before.

Response: Thank you very much for your constructive suggestions.

We have made the following modifications to Figures 2:

(1) The elevation unit has been corrected from "m" to "m asl" (meters above sea level) **in Figure 2**.
(2) The inset map in the upper-left corner has been replaced: it now shows the position of the YRDUA within the Euro-Asia continent, and no longer includes any disputed territories.
(3) The location of the study area is now marked with a red five-pointed star, making it easily identifiable.

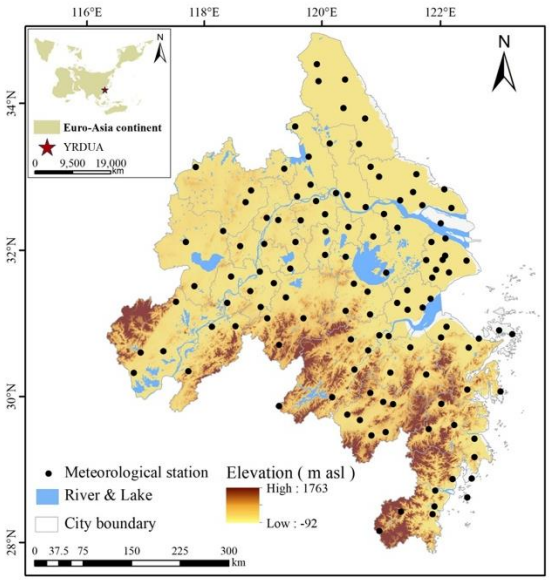

Figure 2: The schematic map of the YRDUA.

2. L165: please add more information about which kind of kriging you use and information about the semivariogram adopted.

Response: Thank you for your helpful suggestion. We have revised the manuscript to specify the type of Kriging interpolation used and the semivariogram model employed.

In this study, we applied ordinary Kriging, which assumes a constant but unknown mean within the local neighborhood. The semivariogram model adopted was the spherical model, which is widely used in environmental geostatistics due to its smooth and bounded nature.

This clarification has been added to the revised manuscript in Section 2.2 (Data Preprocessing) as follows:

"To generate continuous spatial surfaces from discrete data points, we applied the Ordinary Kriging interpolation method, which assumes a constant but unknown local mean (Cressie, 1990). A spherical semivariogram model was adopted to capture spatial autocorrelation, as it is widely used in environmental geostatistics for its bounded range and smooth continuity (Webster and Oliver, 2007). The interpolation process was carried out using ArcGIS 10.8."

We believe this addition improves methodological transparency and will help readers better understand the spatial processing procedure.

3. I'm curious about why the testing performance in several metrics is better in the testing dataset (e.g., extra trees or decision trees) than with the training set. The authors should try to explain that strange behavior.

Response: Thank you for this insightful observation. We agree that it is uncommon for a model to perform better on the test set than on the training set. However, we believe this phenomenon in our study is justifiable due to several technical and methodological considerations, as detailed below.

In our study, AutoML (Auto-sklearn) was employed to construct a binary classification model for flood hazard prediction. Only the six second-level hazard indicators were used as input features: Average Annual Precipitation (PREC), Annual Cumulative Heavy Rainfall Duration (DURA), Digital Elevation Model (DEM), Slope (SLOPE), Drainage Density (DD), and Normalized Difference Vegetation Index (NDVI). These variables represent the natural drivers of flood events. The classification task was based on 556 points (278 flooded + 278 non-flooded) selected from the historical inundation areas, verified using remote sensing and flood databases. The sample set was split into a 70% training set and a 30% test set (as detailed in Section 2.3 and Section 3.1.1).

Regarding your question, several technical factors may explain the slightly better performance observed in some metrics on the test set:

(1) Small Sample Size and Statistical Variability:

The total sample size (556) is relatively limited. With a test set of ~167 samples, minor variations (e.g., one or two easier-to-classify cases) can noticeably influence performance metrics like precision or F1-score. These fluctuations are within the range of statistical randomness and are not indicative of data leakage or overfitting.

(2) Random Partitioning and Class Balance:

Although the dataset was balanced (1:1 flooded to non-flooded), the training and test sets were split randomly. This could lead to the test set containing slightly "easier" or more representative instances, whereas the training set might include a few borderline or noisy samples. Thus, the model may generalize better to the test set purely by chance.

(3) Regularization and AutoML Behavior:

All models were trained under AutoML's hyperparameter tuning pipeline, which uses internal cross-validation and enforces regularization (e.g., tree depth limits or minimum leaf size). As a result, the model may slightly underfit the training data to avoid overfitting. This behavior, particularly with models like Extra Trees that average multiple randomized trees, can result in slightly higher test performance under certain conditions.

(4) Evaluation Metric Sensitivity:

Precision, recall, and F1-score are sensitive to class distribution and small sample shifts. A few additional true positives or fewer false positives in the test set can raise these metrics, especially in a dataset of this size.

(5) No Data Leakage:

We confirm that the training/test split was performed before any modeling or hyperparameter tuning, and all parameter optimization was confined to the training set. The test set was never seen during training or model selection, ruling out data leakage. If leakage had occurred, we would expect all metrics to be uniformly higher—not just marginally so in a few models.

(6) Consistency with Other Studies:

Similar patterns of test performance exceeding training performance in individual metrics have been observed in other flood-related studies. For instance, Wang et al. (2024) reported a comparable result in their study on the Spatio-temporal evolution of public opinion on urban flooding during the 7.20 Henan extreme flood event.

To clarify this in the manuscript, we have added the following sentence to Section 3.1.1 (AutoML optimal model selection):

“Interestingly, in a few cases (e.g., Extra Trees), test set performance slightly exceeded that of the training set in certain metrics. This is not uncommon in small, balanced datasets and may result from a combination of factors such as random sampling variation, slightly easier test samples, or appropriate regularization that reduces overfitting in the training set."

In conclusion, the slight outperformance on the test set is a result of normal statistical variability, model regularization, and careful design of a balanced dataset. This behavior is well-documented in prior literature and does not indicate methodological issues.

4. Data availability statement. Please take care of this section and write all the availability statements, as I requested before. Nowadays, it is not enough to write “Data will be made available on request."

Response: The revised “Data availability” section now explicitly lists each dataset and its corresponding provider, along with web links to public repositories or official data platforms where applicable. We have also removed the previous general statement “Data will be made available on request,” as advised.

Administrative boundaries and river network density were obtained from the Resource and Environmental Science and Data Center, Chinese Academy of Sciences (https://www.resdc.cn/DataList.aspx). Digital elevation data were derived from the SRTM1 dataset provided by USGS (https://earthexplorer.usgs.gov/). Land use data were sourced from the China Land Cover Dataset (CLCD) developed by Wuhan University (https://zenodo.org/records/8176941). Hourly precipitation data from 120 meteorological stations were obtained from the National Meteorological Information Center, China Meteorological Administration (https://data.cma.cn/). Historical flood inundation data were obtained from the MODIS-based Global Flood Database and validated using the EM-DAT disaster database (https://developers.google.com/earthengine/datasets/catalog/GLOBAL_FLOOD_DB_MODIS_EVENTS_V1).

Reply on RC2

1. Line 57 - What are the limitations that have been addressed by ensemble methods? What are ensemble methods?

Response: Thank you for this helpful comment. In the revised manuscript, we have clarified the definition of ensemble methods and explicitly listed the limitations of traditional machine learning models that ensemble methods aim to address. Specifically, we now explain that ensemble methods are a class of ML techniques that combine multiple base learners to form a stronger predictive model, and they are designed to mitigate issues such as high variance,

overfitting, sensitivity to noise, and poor generalization.

The revised sentence now reads:

"Ensemble methods are a class of machine learning techniques that combine multiple base learners to form a stronger predictive model (Webb and Zheng, 2004). They are designed to overcome several limitations of individual models, such as high variance, overfitting, sensitivity to noise, and poor generalization(Yang et al., 2013). By aggregating the outputs of weak learners, ensemble methods significantly enhance model stability, accuracy, and robustness—especially in high-dimensional and complex classification or regression tasks (Kazienko et al., 2015)."

2. Line 58 - What are integrated ML methods?

Response: Thank you for pointing out this ambiguity. In the original manuscript, the term "integrated ML methods" was intended to refer to ensemble machine learning techniques. To avoid confusion, we have revised the sentence to use the more accurate and standard term "ensemble ML techniques." Additionally, we have provided specific examples of ensemble methods such as bagging (Random Forest), boosting (XGBoost, CatBoost), and stacking, and clarified their relevance to hydrological modeling.

The revised sentence reads:

"Various ensemble ML techniques, including bagging (e.g., Random Forest), boosting (e.g., XGBoost, CatBoost), and stacking, have been widely used in hydrology, with boosting algorithms in particular showing strong performance in flood prediction and risk assessment (Shafizadeh-Moghadam et al., 2018; Mirzaei et al., 2021; Yan et al., 2024)."

3. Line 71 - The statement is unclear and seems incorrect - the authors have themselves pointed out that by carefully configuring ML models, they can in fact perform well across varied problems.

Response: Thank you for your careful reading and insightful comment. We acknowledge that the original statement may appear contradictory. In the revised version, we have clarified our point by emphasizing that while machine learning algorithms can achieve strong performance in many cases, no single algorithm universally outperforms others across all tasks. Therefore, careful configuration of each component in the ML pipeline—such as feature engineering, model selection, and hyperparameter tuning—is essential to adapt to different problems.

The revised sentence reads:

"While machine learning algorithms have demonstrated strong performance in many domains, no single algorithm consistently performs best across all types of problems. Therefore, to achieve optimal performance, it is essential to carefully configure key components of the ML

pipeline, including feature engineering, model selection, and hyperparameter tuning (Li et al., 2017; Raschka, 2020)."

4. Line 76 - What does experience mean?

Response: Thank you for your insightful comment. We have revised the sentence to explicitly define what "experience" refers to in the context of machine learning. We also revised the structure of the sentence for clarity and to maintain a logical flow into the key research challenge addressed in our study.

The updated sentence now reads:

"The effectiveness of ML improves with experience, where "experience" refers to the model's iterative exposure to training data and its ability to learn patterns from labeled examples (Jordan and Mitchell, 2015; Nagarajah and Poravi, 2019). One key challenge addressed in this study is how to automatically optimize model components such as feature selection and algorithm configuration in flood risk prediction, while maintaining high accuracy and adaptability across complex hydrological conditions. "

5. Line 80 - What are some of the various problems being addressed by AutoML?

Response: Thank you for your helpful comment. We agree that the original wording was too general and did not adequately specify the kinds of problems AutoML addresses. In the revised manuscript, we now explicitly state that AutoML automates several key processes in the machine learning pipeline, including feature selection, model selection, hyperparameter tuning, and ensemble learning. These steps are critical to improving model performance and reducing subjectivity.

In addition, we have clarified the relevance of AutoML in the context of this study. Specifically, AutoML enables automatic optimization of hazard factor selection, model construction, and parameter adjustment in the flood risk assessment workflow.

The revised passage reads:

"AutoML is an innovative machine learning framework that automates key stages of the model development pipeline, including feature selection, model selection, hyperparameter tuning, and ensemble learning. By addressing these challenges, AutoML reduces reliance on expert knowledge and minimizes subjectivity in model building (He et al., 2021; Consuegra-Ayala et al., 2022). In the context of this study, AutoML enables the automatic optimization of hazard factor selection, model construction, and parameter adjustment for flood risk assessment tasks, thereby improving efficiency, objectivity, and reproducibility in model development."

6. Line 95 - The authors have not provided any evidence to support this statement.

7. Line 96 - The statement is unsupported.

Response: Thank you for your valuable comments. We agree that the original statements in Lines 95–96 lacked sufficient supporting evidence. In the revised manuscript, we have restructured this section to improve clarity and provide appropriate citations. Specifically, we now cite Guo et al. (2022c) to support the effectiveness of AutoML in flood hazard prediction, and Hutter et al. (2019) and He et al. (2021a) to support its methodological advantages.

The revised text now reads:

"In the field of flood risk assessment, AutoML has been preliminarily demonstrated to perform well in flood hazard prediction (Guo et al., 2022). As an efficient 'black-box' modeling approach, AutoML provides strong support for flood risk modeling through automated feature selection, model training, and parameter optimization (Hutter et al., 2019; He et al., 2021). "

We believe this revision strengthens the foundation for the subsequent discussion on the methodological limitations of AutoML and the motivation for incorporating MCDA.

8. Line 99 -Which complex decision making problems are the authors referring to?

Response: Thank you for your insightful comment. We agree that the original sentence did not sufficiently clarify what types of complex decision-making problems are involved. In the revised manuscript, we have elaborated on this point by explaining that flood risk assessment in urban agglomerations involves multiple natural and socioeconomic indicators—such as rainfall, topography, land use, drainage, and population density — that originate from heterogeneous, often multi-source datasets and differ in type. These indicators frequently interact in non-linear and uncertain ways, making it difficult to evaluate their relative contributions to overall flood risk.

To address these challenges, we now clarify that multicriteria decision analysis (MCDA) provides a structured framework for integrating such diverse indicators into a unified evaluation system by constructing weighting schemes that align the results with real-world conditions and expert knowledge (Fernández and Lutz, 2010). Furthermore, in cases where certain data are missing or difficult to quantify, MCDA allows for the incorporation of expert judgment through tools such as scoring systems and pairwise comparison matrices, thus improving the robustness and practical applicability of the assessment (Hites et al., 2006).

The revised passage now reads:

In urban agglomerations, flood risk assessment is a highly complex task involving diverse natural and socioeconomic factors derived from heterogeneous and often multi-source datasets (Wang et al., 2023). These factors—such as rainfall, topography, land use, drainage, and population density—differ in type and often interact in non-linear and uncertain ways (Shuster

et al., 2005; Zhang et al., 2017; Wang et al., 2018). Under such complex circumstances, AutoML struggles to systematically evaluate the multi-dimensional indicators of flood risk. To address this limitation, this study introduces a multicriteria decision analysis (MCDA) approach to quantify the importance of various indicators within the evaluation framework (Pham et al., 2021). MCDA facilitates the integration of such heterogeneous indicators into a unified evaluation framework by constructing structured weighting schemes, thereby aligning the assessment results more closely with real-world conditions and expert knowledge (Fernández and Lutz, 2010). In cases where data are limited or certain indicators are difficult to quantify, MCDA methods allow for the incorporation of expert judgment through scoring systems and pairwise comparison matrices, enhancing the practical applicability and robustness of the model (Hites et al., 2006).

9. Line 171 - Was the normalization done prior to or after splitting the data into training and test sets? Doing it prior to split can give incorrectly optimistic results due to data leakage. For example, if training data only had GDP up to $1000 per capita, then the normalized value of 1.0 will refer to that. In test data, if a value of $1500 is observed, it would now be input as a value > 1.0 using the normalization logic from the training set. However, by normalizing prior to the split, the normalization logic would have already "seen" the $1500 value and assigned 1.0 to it, hence the data has been leaked.

Response: Thank you for your valuable comment. We appreciate your observation and acknowledge that the original wording in the manuscript may have led to misunderstanding.

We would like to clarify that in this study, the normalization was conducted after splitting the dataset into training and testing subsets, which effectively avoids data leakage. Specifically, the minimum and maximum values used for normalization were computed only from the training set, and the same values were used to normalize both the training and test sets. As a result, the normalized values in the training set are guaranteed to lie within the range [0,1], while values in the test set may exceed this range if they fall outside the value range of the training data.

To ensure clarity and correctness, we have revised the relevant sentence in the manuscript as follows:

(2) Normalization of the numerical range can be achieved using a normalization process. In this study, the Min-Max normalization method is applied. Specifically, the minimum and maximum values of each feature are computed only from the training set, and both the training and test sets are then normalized using these training-derived parameters. This ensures that the normalized values in the training set are scaled to the range [0,1], while the values in the test set may exceed this range if they fall outside the training set's value distribution. The formula is as follows:

$$x' = \frac{x - x_{min}^{train}}{x_{max}^{train} - x_{min}^{train}}$$

10. Line 177 - What is the definition of a flooded point given that multiple historical floods are being considered?

Response: Thank you for your insightful comment. We agree that the definition of a "flooded point" should be clearly stated, especially when multiple historical flood events are involved. In the revised manuscript, we have clarified that:

A flooded point is defined as a location that lies within the inundation extent of at least one recorded flood event during the study period.

This definition reflects the cumulative spatial footprint of flood events over time. All flooded points were selected from validated inundated areas identified through flood traces in the Global Flood Database and the EM-DAT database, and further verified through satellite imagery, Google Earth, and historical flood records. This ensures that the flooded points used in model training accurately represent locations that were affected by at least one historical flood.

11. Figure 4 - How were the second-level factors selected? Surely, there are 10's-100's of other factors each of which could contribute to H-E-V-C, so what led to the specific choices to select this subset?

Response: Thank you for your valuable comment. We agree that each dimension of the H-E-V-R framework includes a wide range of potential indicators.

In this study, the selection of second-level indicators was based on the following considerations:

(1) Theoretical framework and literature support: The indicator system was developed with reference to studies based on the H-E-V-R framework or similar vulnerability–resilience analytical approaches. It was specifically adapted to the characteristics of the Yangtze River Delta urban agglomeration. The selected variables have been widely used in existing flood risk assessments to quantify meteorological hazards, exposure, vulnerability, and resilience, and are proven to be representative and practical (Gain et al., 2015; Criado et al., 2019; Liu et al., 2019; Hsiao et al., 2021; Bin et al., 2023).

(2) Data availability and spatial applicability: All indicators used in this study are based on publicly available datasets with broad spatial coverage. Their resolution is appropriate for regional-scale analysis, particularly at the urban agglomeration level. This ensures consistency and comparability across the entire study area.

(3) Representativeness and reduction of redundancy: For each dimension, we prioritized key

indicators that reflect distinct aspects while avoiding strong correlations or overlapping meanings among variables. For example, in the exposure dimension, flood-related damage is typically correlated with the concentration of people, economic activity, infrastructure, and other assets in affected areas. Therefore, we selected indicators such as land area, population density, GDP density, and building density to capture different perspectives on the degree of exposure of natural and social systems to flood hazards.

(4) Modeling feasibility and efficiency: While ensuring comprehensive coverage of key factors, we also aimed to maintain interpretability and computational efficiency in the machine learning models. This avoids potential overfitting caused by excessive input variables. As a result, a representative and empirically grounded subset of indicators was finalized for use in this study.

Importantly, we would like to note that these principles and the rationale for indicator selection were already addressed in the previous revision of Section 2.4 ("Establishment of a flood risk assessment indicator system"). In that section, we clearly outlined the indicator selection process for each of the four dimensions (hazard, exposure, vulnerability, and resilience), explaining their relevance, definitions, and interpretation within the flood risk assessment framework. This earlier addition reflects our attention to scientific transparency and methodological justification.

12. Line 300 - How is the balance of Precision, Recall, and F1-score quantified?

Response: Thank you for your valuable comment. We have revised the relevant paragraph in the manuscript to clarify how the balance between Precision and Recall was quantified during model selection.

Specifically, the F1-score was used as the primary evaluation metric, as it provides a harmonic mean of Precision and Recall and is widely used to capture the trade-off between the two in binary classification tasks. For each set of hyperparameter combinations, k-fold cross-validation was performed within the training set, and the model with the highest average F1-score across all folds was selected as the optimal configuration. Precision and Recall were also reported independently to support interpretability, but F1-score served as the key criterion for quantifying the overall balance.

This clarification has been added to the revised manuscript (Section 2.5.2).

13. Line 419 - How was this 5-fold cross-validation performed, and how was it different than the k-fold used for hyperparameter tuning?

Response: Thank you for this important comment. We appreciate the opportunity to clarify the two distinct uses of cross-validation in our study.

In response to your suggestion, we have added a clarification at the end of Section 2.5.2 Model

training and hyperparameter optimization, explaining that 5-fold cross-validation was applied at two distinct stages:

(1) During hyperparameter optimization, 5-fold cross-validation was used within the training set only, to evaluate and select the best-performing hyperparameter combination.
(2) Following final model selection, an independent 5-fold cross-validation was applied to the entire dataset to evaluate the generalization performance of the model and identify potential overfitting.

The data splits used in these two stages were entirely separate, and no data leakage occurred. This clarification has been added to the revised manuscript to improve the transparency and reproducibility of the experimental procedure.

14. Line 485 - Did the authors populate the judgement matrix? How were the values selected?

Response: Thank you for your insightful question. We have revised Section 3.1.3 of the manuscript to clarify how the judgment matrices were constructed and how the values were determined.

Specifically, the judgment matrices were populated by the authors using a hybrid approach. For the hazard indicators, the pairwise comparisons were informed by feature importance scores obtained through the AutoML model. For the exposure, vulnerability, and emergency resilience indicators, the relative weights were derived based on a combination of expert judgment, a review of relevant literature, and the socio-environmental characteristics of the YRDUA. The Saaty 1–9 scale was applied to express the relative importance of each pair of indicators.

All judgment matrices underwent consistency checks, and the calculated consistency ratio (CR) was 0.0058, which is well below the 0.1 threshold, indicating that the matrices were logically consistent. These clarifications have been incorporated into the revised manuscript to enhance transparency and reproducibility.

15. Section 3.1.2 - Were only the hazard second-level factors used as features for the AutoML method, or were all 19 factors used, as only the importance of hazard factors is presented?

Response: Thank you for your important comment. We appreciate the opportunity to clarify the structure and feature input strategy of the AutoML model.

In this study, only the six second-level indicators under the hazard dimension were used as input features for the AutoML model. This is because flood hazard—defined as the physical likelihood of flooding—is primarily determined by natural environmental factors such as precipitation, elevation, slope, and drainage density. The AutoML algorithm was used to build a binary classification model (flooded vs. non-flooded points), which is a task it performs particularly well due to its strengths in automatic feature selection, high computational

efficiency, and predictive accuracy.

However, AutoML, as a black-box model, lacks the interpretability required for evaluating social or systemic dimensions such as exposure, vulnerability, and resilience. It cannot provide structured weights across multiple dimensions of flood risk. To address this, we incorporated the Analytic Hierarchy Process (AHP), a widely used multi-criteria decision-making method in flood risk assessment, to assign weights to these additional dimensions based on expert judgment and literature review.

The judgment matrix for the hazard indicators was constructed based on the feature importance scores output by the AutoML model. In contrast, the matrices for the exposure, vulnerability, and resilience dimensions were constructed using the 1–9 Saaty scale, informed by domain experts and relevant literature.

To ensure transparency, we have explicitly clarified this distinction in the revised manuscript in two places:

(a) In the final paragraph of the Introduction, we further clarified the overall model structure:

"This study develops a flood risk assessment model for the YRDUA by analyzing the factors influencing flood risk and integrating AutoML and AHP methods. In this model, AutoML is employed to construct the flood hazard sub-model, using indicators that represent natural environmental drivers as input features. The hazard is modeled as a binary classification problem (i.e., whether flooding occurs), and the resulting feature importance rankings provide an objective basis for subsequent indicator weighting. Nevertheless, as a data-driven approach, AutoML alone cannot structurally interpret the relative influence of social and systemic factors within a multi-dimensional flood risk assessment framework. Therefore, this study incorporates the AHP to calculate the weights of flood exposure, vulnerability, and resilience in the YRDUA, based on expert knowledge and existing literature. A regional flood risk zoning map is then generated. A comparative analysis with observed inundation points data shows a strong spatial alignment between the distribution of flooded points and the high to medium-high risk zones, highlighting the reliability and applicability of the proposed model. The remainder of this paper is structured as follows: Section 2 describes the study area, data sources, and methodology; Section 3 presents the results and analysis; Section 4 discusses the findings and their implications; and Section 5 concludes the study with key insights and recommendations."

(b) At the beginning of Section 3.1.2, we added the following explanation:

"In this study, the AutoML model was used specifically to assess flood hazard, which represents the physical likelihood of flood occurrence and is directly driven by environmental factors such as rainfall, topography, and drainage characteristics. Therefore, only the six second-level indicators under the hazard dimension were used as input features in the AutoML model. This

approach allowed us to focus the model on identifying the key natural drivers of flooding, while the other dimensions—exposure, vulnerability, and resilience—were later incorporated via the AHP method for comprehensive flood risk evaluation."

We hope this explanation and revision adequately address your concern.

16. Section 3.2 - Some of the factors like population and building density are expected to consistently increase and have likely done so between 1990 to 2010. What were the ranges of these factors during model training vs when using the model for calculating spatio-temporal variation? A table showing typical values both during training and later implementation will be helpful, and e.g., could be included in Table 1.

Response: Thank you for this thoughtful question. We believe this comment raises an important point regarding the temporal evolution of socio-economic indicators such as population density (DPOP) and building density (DBUI), and their implications for model training and application. We appreciate the opportunity to clarify how these variables were used in our study.

We would like to clarify that in this study, only the six second-level indicators under the flood hazard dimension—namely Average Annual Precipitation (PREC), Annual Cumulative Heavy Rainfall Duration (DURA), Digital Elevation Model (DEM), Slope (SLOPE), Drainage Density (DD), and NDVI—were used as input features for AutoML model training. These indicators are driven by natural environmental conditions and were used to train a binary classification model to predict flood occurrence (flooded vs. non-flooded points).

In contrast, other indicators such as population density, GDP density, and building density, which are known to evolve over time, were not included in the AutoML training process. Instead, they were integrated into the comprehensive flood risk assessment using the Analytic Hierarchy Process (AHP). This method enabled us to incorporate both expert judgment and literature references to assign weights to all second-level indicators across the three dimensions: exposure, vulnerability, and resilience.

This design reflects our overall methodological strategy of combining data-driven modeling (AutoML) with expert-informed decision analysis (AHP), enabling both predictive accuracy and model interpretability. Importantly, the AutoML component, which could be affected by variable shifts over time, was isolated from these time-sensitive socio-economic features, ensuring the model's temporal consistency and preventing leakage or extrapolation issues.

To further clarify this distinction, we have added an explanatory paragraph at the beginning of Section 3.1.2 of the revised manuscript.

We hope this explanation adequately addresses your concern.

References

Amini, A., Abdollahi, A., and Hariri-Ardebili, M. A.: An automated machine-learning-assisted stochastic-fuzzy multi-criteria decision making tool: Addressing record-to-record variability in seismic design, Applied Soft Computing, 154, 111354, https://doi.org/10.1016/j.asoc.2024.111354, 2024.

Bin, L., Xu, K., Pan, H., Zhuang, Y., and Shen, R.: Urban flood risk assessment characterizing the relationship among hazard, exposure, and vulnerability, Environ Sci Pollut Res, 30, 86463–86477, https://doi.org/10.1007/s11356-023-28578-7, 2023.

Cressie, N.: The origins of kriging, Math Geol, 22, 239–252, https://doi.org/10.1007/BF00889887, 1990.

Criado, M., Martínez-Graña, A., San Román, J. S., and Santos-Francés, F.: Flood risk evaluation in urban spaces: The study case of Tormes River (Salamanca, Spain), International journal of environmental research and public health, 16, 5, 2019.

Fernández, D. S. and Lutz, M. A.: Urban flood hazard zoning in Tucumán Province, Argentina, using GIS and multicriteria decision analysis, Eng. Geol., 111, 90–98, 2010.

Gain, A. K., Mojtahed, V., Biscaro, C., Balbi, S., and Giupponi, C.: An integrated approach of flood risk assessment in the eastern part of Dhaka City, Nat. Hazards, 79, 1499–1530, https://doi.org/10.1007/s11069-015-1911-7, 2015.

Guo, Y., Quan, L., Song, L., and Liang, H.: Construction of rapid early warning and comprehensive analysis models for urban waterlogging based on AutoML and comparison of the other three machine learning algorithms, Journal of Hydrology, 605, 127367, https://doi.org/10.1016/j.jhydrol.2021.127367, 2022.

He, X., Zhao, K., and Chu, X.: AutoML: A survey of the state-of-the-art, Knowledge-Based Systems, 212, 106622, https://doi.org/10.1016/j.knosys.2020.106622, 2021.

Hites, R., De Smet, Y., Risse, N., Salazar-Neumann, M., and Vincke, P.: About the applicability of MCDA to some robustness problems, European Journal of Operational Research, 174, 322–332, https://doi.org/10.1016/j.ejor.2005.01.031, 2006.

Hsiao, S.-C., Chiang, W.-S., Jang, J.-H., Wu, H.-L., Lu, W.-S., Chen, W.-B., and Wu, Y.-T.: Flood risk influenced by the compound effect of storm surge and rainfall under climate change for low-lying coastal areas, Science of the total environment, 764, 144439, 2021.

Hutter, F., Kotthoff, L., and Vanschoren, J. (Eds.): Automated Machine Learning: Methods, Systems, Challenges, Springer Nature, https://doi.org/10.1007/978-3-030-05318-5, 2019.

Liu, Y., Li, L., Zhang, W., Chan, P., and Liu, Y.: Rapid identification of rainstorm disaster risks based on an artificial intelligence technology using the 2DPCA method, Atmospheric Research, 227, 157–164, https://doi.org/10.1016/j.atmosres.2019.05.006, 2019.

Pham, B. T., Luu, C., Dao, D. V., Phong, T. V., Nguyen, H. D., Le, H. V., von Meding, J., and Prakash, I.: Flood risk assessment using deep learning integrated with multi-criteria decision analysis, Knowledge-Based Systems, 219, 106899, https://doi.org/10.1016/j.knosys.2021.106899, 2021.

Shuster, W. D., Bonta, J., Thurston, H., Warnemuende, E., and Smith, D. R.: Impacts of impervious surface on watershed hydrology: A review, Urban Water Journal, https://doi.org/10.1080/15730620500386529, 2005.

Wang, P., Deng, X., Zhou, H., and Qi, W.: Responses of urban ecosystem health to precipitation extreme: A case study in Beijing and Tianjin, Journal of Cleaner Production, 177, 124–133, https://doi.org/10.1016/j.jclepro.2017.12.125, 2018.

Wang, W., Zhu, X., Lu, P., Zhao, Y., Chen, Y., and Zhang, S.: Spatio-temporal evolution of public opinion on urban flooding: Case study of the 7.20 Henan extreme flood event, International Journal of Disaster Risk Reduction, 100, 104175, https://doi.org/10.1016/j.ijdrr.2023.104175, 2024.

Wang, Y., Li, C., Hu, Y., Lv, J., Liu, M., Xiong, Z., and Wang, Y.: Evaluation of urban flooding and potential exposure risk in central and southern Liaoning urban agglomeration, China, Ecological Indicators, 154, 110845, https://doi.org/10.1016/j.ecolind.2023.110845, 2023.

Webb, G. I. and Zheng, Z.: Multistrategy ensemble learning: reducing error by combining ensemble learning techniques, IEEE Transactions on Knowledge and Data Engineering, 16, 980–991, https://doi.org/10.1109/TKDE.2004.29, 2004.

Webster, R. and Oliver, M. A.: Geostatistics for Environmental Scientists, John Wiley & Sons, 333 pp., 2007.

Yang, J., Zeng, X., Zhong, S., and Wu, S.: Effective Neural Network Ensemble Approach for Improving Generalization Performance, IEEE Transactions on Neural Networks and Learning Systems, 24, 878–887, https://doi.org/10.1109/TNNLS.2013.2246578, 2013.

Zhang, H., Wu, C., Chen, W., and Huang, G.: Assessing the Impact of Climate Change on the Waterlogging Risk in Coastal Cities: A Case Study of Guangzhou, South China, https://doi.org/10.1175/JHM-D-16-0157.1, 2017.