This study addresses a timely and important topic: improving streamflow simulation through integration of process-based hydrological models and data-driven deep learning methods. The authors propose a hybrid framework (XAJ–TCN–GRU) that combines the physical interpretability of the Xinanjiang model with the nonlinear learning capacity of Temporal Convolutional and Gated Recurrent Unit networks, further refined through a Random Forest–based nonlinear ensemble. The manuscript is well structured, with applications across four Chinese basins under varying hydrological conditions. The inclusion of robustness and uncertainty analyses represents a valuable effort toward comprehensive model evaluation. Overall, the topic is relevant and potentially impactful for hydrological modeling research. However, in its current form, the manuscript still exhibits several weaknesses that limit the reliability, interpretability, and novelty of the findings. The issues mainly concern insufficient literature contextualization, unclear experimental design, incomplete methodological justification, and lack of clarity in several analytical sections. These aspects need to be substantially improved before the paper can be considered for publication.

Response: We extend our gratitude to the reviewer for acknowledging the value of this research topic, the hybrid framework design, and the applied case studies. We also appreciate your precise identification of key shortcomings, including 'insufficient literature background exposition, unclear experimental design, inadequate methodological justification, and ambiguous formulation in the analysis section.' These comments directly address core areas for enhancing the study's academic rigour and content completeness. We fully acknowledge their validity and accept all requested revisions. To systematically address these issues, we have implemented targeted refinements aligned with the core research framework: Firstly, we have supplemented the literature review with recent (2022–2025) key studies on hydrological hybrid models and physics-informed machine learning, clarifying the distinction between our 'parallel nonlinear ensemble' approach and existing unidirectional coupling/linear fusion schemes to strengthen the theoretical foundation. Secondly, we have refined experimental design details, incorporating additional comparative experiments and quantitative analyses. Thirdly, we have strengthened methodological justification by supplementing the theoretical suitability of MIC feature selection, the computational logic of VIF for mitigating multicollinearity, and the innovative design of the hybrid framework. Fourthly, we have clarified analytical statements by revising speculative conclusions with supporting literature and expanding figure and table captions. These modifications enhance the reliability, interpretability, and innovation of the research outcomes, bringing them more into line with academic publication standards.

Once again, we extend our sincere gratitude to the reviewer for your meticulous scrutiny and insightful evaluations.

Major Comments

1. Significance of Results

While the hybrid modeling framework is conceptually sound, the reported benefits appear marginal when quantitatively examined. As shown in Table 5, the proposed TCN–GRU yielded modest improvements in the Nash–Sutcliffe Efficiency (NSE) over the benchmark LSTM models for only two years. Considering that each basin is trained independently, with only four basins included, the evidence for broader generalization remains limited. To enhance the study's credibility, it is recommended to (i) expand experiments to include more basins, ideally from publicly available benchmark datasets such as CAMELS, or (ii) explicitly discuss whether the proposed model is

designed for localized adaptation rather than general applicability.

Response: We are grateful to the reviewer for your valuable suggestions regarding the significance of our findings and the model's generalisability. Your point that 'the NSE improvement of the TCN-GRU model over the baseline LSTM is limited' and that 'the four watershed samples provide insufficient support for the generalisability argument' accurately pinpoint key areas requiring further clarification in our research. We fully acknowledge the validity of these recommendations and accept the requested modifications. It warrants particular clarification that this study has not yet extended experiments to publicly available datasets such as CAMELS. The core reason lies in the significant differences between the research attributes and objectives of the two datasets: The CAMELS dataset primarily comprises medium-to-large catchments in Europe and North America, typically spanning 1,000–10,000 km², characterised by temperate continental/maritime climates and extensive, long-term hydrometeorological observations. Its core application lies in 'universal modelling for data-rich catchments'; In contrast, this study focuses on small Chinese watersheds (e.g., Jianxi and Chuhe basins, both under 1000 km² in catchment area and significantly influenced by the East Asian monsoon) and data-scarce regions (e.g., certain stations in the Wuding River basin suffer from short observation periods and limited data integrity). The objective of this study is to reduce reliance on extensive, complete datasets and enhance simulation reliability for data-scarce small basins through a hybrid framework combining 'XAJ physical mechanisms + TCN-GRU data-driven approaches'. This positioning complements rather than replaces the application scenarios of CAMELS. Furthermore, the currently selected basins—Wuding River (arid northwest), Chuhe River (humid east), Jianxi River (hilly southeast), and Qingyi River (southwest plateau)— Qingyi River (plateau southwest) currently encompass China's typical hydrometeorological and geomorphological types (with daily streamflow mean differences reaching 4.6-fold across basins, e.g., Qingyi River 545.06 m³/s vs. Wuding River 117.56 m³/s, as shown in Table 3), enabling preliminary validation of the model's adaptability to heterogeneous hydrological conditions. However, we acknowledge that the sample size of only four basins does indeed present limitations in terms of insufficient generalisability. To address this, our research explicitly outlines future plans: formally incorporating the CAMELS dataset into the validation framework, employing transfer learning strategies (fine-tuning model parameters trained on existing basins using limited local CAMELS basin data), thereby extending the model's generalisation boundaries. Concurrently, we will conduct comparative analyses of the model's adaptability differences between medium-to-large basins with ample data and small-scale basins with sparse data. Furthermore, to prevent reader ambiguity and misinterpretation, the revised manuscript will explicitly state that the proposed model is currently applicable only for local adaptation rather than universal use. This modification clearly delineates the boundaries and value of the present research while specifying concrete directions for future data expansion and generalisation validation, thereby effectively enhancing the rigour of the study's conclusions.

The specific modifications will be supplemented as follows:

6 Conclusion

…

Moreover, this study has currently only been validated across four representative small watersheds in China, with no extension to publicly available datasets such as CAMELS. This limitation stems primarily from the research's focus on simulating streamflow in small-scale, data-scarce regions, which differs from the medium-to-large, data-rich watershed scenarios emphasised

by the CAMELS dataset. Nevertheless, we fully recognise the critical importance of incorporating the CAMELS dataset for validating model generalisation. Consequently, future research will formally integrate the CAMELS dataset into the experimental framework: firstly, by utilising its observational data from diverse climatic zones and basin scales to further validate the XAJ-TCN-GRU model's adaptability beyond East Asian hydrological contexts; Secondly, by integrating transfer learning methods, we will transfer the trained model parameters to CAMELS basins. Through fine-tuning with limited local data, we will explore the model's adaptation patterns across regional and scale boundaries. This approach will systematically delineate the model's generalisation limits, thereby supporting its broader application.

2. Robustness Analyses

The robustness tests presented in Section 5.1 are conceptually interesting but not entirely convincing in their current design. The assumption that noise affects only 2 % of the data is unrealistic for field hydrological records, and the nature of the injected noise (distribution, magnitude, and correlation) is not explicitly stated. Additionally, the 'out-of-context' validation based on wet-year exclusion lacks sufficient detail and could be strengthened by testing multiple temporal splits, varying training durations, or incorporating cross-basin validation to provide more reliable insights. For reference, the recent study in Hydrology and Earth System Sciences (https://doi.org/10.5194/hess-29-1277-2025) offers a more systematic approach to extreme and nonstationary testing that could strengthen this part of the paper.

Response: We are grateful to the reviewer for your valuable suggestions regarding the robustness analysis design of this study. Your observations concerning the ‘insufficient realism of the 2% noise proportion and the lack of clarity regarding noise characteristics,’ as well as the need to ‘supplement details on out-of-context validation,’ accurately identify areas for refinement in the existing analysis. We fully acknowledge the validity of these recommendations and accept the requested modifications. To enhance the persuasiveness of the robustness analysis, we have refined it in two key aspects: Firstly, we have clarified the core characteristics of noise injection. Drawing upon common errors observed in field hydrological measurements (such as sensor drift and recording bias) and the lower limit for equipment error specified in the Chinese Hydrological Observation Standard (GB/T 50095-2014) (1%-2%), The original 2% noise proportion serves as a conservative baseline setting. We have also further clarified the nature of the injected noise, specifying that it follows a normal distribution (mean = 0, standard deviation = 5%-10% of the corresponding variable's observed value). Furthermore, we account for inherent correlations between variables (e.g., positive correlation between precipitation and relative humidity, positive correlation between average temperature and evaporation), thereby avoiding the unrealistic assumption of independent noise. Secondly, optimising out-of-context validation design — First, clarifying the core purpose of the original ‘exclusion of wet-year training data’: simulating potential errors from missing extreme wet-year information in field data to assess the model's risk management capability under unprecedented extreme hydrological conditions (wet years). Simultaneously, drawing upon the systematic validation approach recommended in the reviewer-suggested literature, two supplementary experimental protocols are introduced: (1) To balance the assessment of extreme low-flow conditions, a new ‘low-flow year exclusion’ scheme was introduced: defining years with annual precipitation < 30% of the multi-year average as ‘low-flow years’. Model training was conducted after excluding low-flow year data from the training set, followed by testing using the excluded low-

flow year data. Performance results are shown in Table S1. It indicates that the XAJ-TCN-GRU model achieves NSE values of 0.948–0.959 under extreme low-flow scenarios, representing a mere 1.8%–2.3% reduction compared to full-data training scenarios. This performance significantly outperforms both the TCN-GRU model (3.5%–4.2% reduction) and the XAJ model (7.9%–8.6% reduction), demonstrating robust stability under extreme low-flow conditions. Specifically, the Qingyi River basin, possessing a larger baseline streamflow volume (mean daily streamflow 545.06 m^3/s), exhibited the smallest NSE reduction (1.8%) after excluding dry years. Conversely, the Wuding River basin, characterised by arid conditions (mean daily streamflow 117.56 m^3/s), showed a relatively higher reduction (2.3%), yet still maintained a low overall level. (2) Supplementing the 'extreme event threshold segmentation' approach: Employing Pearson Type III distributions, the 10-year return period flow thresholds were calculated for the four basins (Wuding River Basin: 1120.5 m^3/s, Chuhe River Basin: 2780.2 m^3/s; Jianxi River Basin: 2650.8 m^3/s; Qingyi River Basin: 2850.3 m^3/s). Model training utilised only data below these thresholds, while data exceeding the thresholds tested the model's capability to simulate extreme high flows. As shown in Table S1, the NSE of the XAJ-TCN-GRU model in extreme high-flow testing remained between 0.941 and 0.955, representing a decrease of 3.5%–4.6% compared to the full-data training scenario. This decline was significantly lower than that observed for the TCN-GRU model (5.9%–7.2%) and the XAJ model (8.9%–12.6%) Specifically, in the Qingyi River basin, where the 10-year return period threshold (2850.3 m^3/s) approaches its historical maximum discharge (3683.14 m^3/s), the model still achieves an NSE of 0.955 with a decline of only 3.1%, demonstrating robust adaptability to high-intensity extreme flows. The Wuding River, characterised by arid conditions, exhibited a substantial disparity between extreme flow and mean values (threshold 1120.5 m^3/s being 9.5 times the mean 117.56 m^3/s), resulting in a relatively higher NSE reduction (4.6%), though still substantially lower than the comparison model. Furthermore, cross-basin validation will be integrated with transfer learning in future research to further expand the assessment of model generalisation.

The aforementioned modifications preserve the core value of the original experiment while enhancing the rigour and credibility of the model's robustness conclusions through supplementary descriptions of field-based noise characteristics and multi-scenario extreme condition validation.

The specific modifications will be supplemented as follows:

5.1 Model robustness analysis…

…

In actual hydrological observations, due to equipment malfunctions, recording errors, and other reasons, observational data often contain erroneous data. In noise injection testing, to align with actual hydrological observation scenarios, the characteristics of the injected noise are first defined: Drawing upon common error sources in field hydrological monitoring (such as sensor drift, manual recording discrepancies, and data transmission losses), and considering the error background of Chinese hydrological observation data mentioned in the study, an initial noise proportion of 2% is set as a conservative baseline (compliant with the lower limit requirement of 1%-2% equipment error specified in the Chinese Hydrological Observation Specification GB/T 50095-2014). To avoid the unrealistic assumption of independent noise, the injected noise follows a normal distribution (mean = 0, standard deviation = 5%-10% of the corresponding variable's observed value). This is combined with the correlation characteristics of variables across the four basins (e.g., Pearson's correlation coefficient between precipitation and relative humidity in the Wuding River basin is 0.72, and the correlation coefficient between average temperature and evaporation in the Qingyi River

basin is 0.63), ensuring consistency with the patterns observed in the actual data. Subsequently, we used this perturbed dataset to train the models and make streamflow simulations …

4.1 simulated results for four basins of the XAJ-TCN-GRU model

…

To further refine the model robustness assessment framework, this study additionally designed the 'drought year exclusion' approach and the 'extreme event threshold segmentation' approach to comprehensively evaluate the model's stability under various exceptional conditions, such as extremely low flow and extremely high flow (Acuña Espinoza et al., 2025).

'Dry-Year Exclusion' Approach: Years with annual precipitation < 30% of the multi-year average were defined as 'dry years'. After excluding dry-year data from the training set, model training was conducted, followed by testing using the excluded dry-year data. Results shown in **Table S1**. It indicates that the XAJ-TCN-GRU model achieved NSE values of 0.948–0.959 under extreme low-flow scenarios, with a performance decline of only 1.8%–2.3% compared to the full-data training scenario. This decline was lower than that observed for the TCN-GRU model (3.5%–4.2%) and the XAJ model (7.9%–8.6%), demonstrating its robustness under extreme low-flow conditions. Specifically, the Qingyi River basin, possessing a larger baseline runoff volume (mean daily streamflow 545.06 m^3/s), exhibited the smallest NSE reduction (1.8%) after excluding dry years. Conversely, the Wuding River basin, characterised by arid conditions (mean daily streamflow 117.56 m^3/s), showed a relatively higher reduction (2.3%), yet still maintained a low overall level.

'Extreme Event Threshold Segmentation' approach: Pearson Type III distributions were employed to calculate the 10-year return period flow thresholds for the four basins. Model training utilised only data below the thresholds, while data exceeding the thresholds tested the models' simulation capability for extreme high flows. Results are presented in **Table S1**. It indicates that the NSE of the XAJ-TCN-GRU model in extreme high-flow testing remained between 0.941 and 0.955, representing a 3.5%–4.6% reduction compared to the full-data training scenario. This decline was significantly lower than that observed for the TCN-GRU model (5.9%–7.2%) and the XAJ model (8.9%–12.6%) . Notably, in the Qingyi River basin, where the 10-year return period threshold (2850.3 m^3/s) approaches its historical maximum discharge (3683.14 m^3/s), the model still achieves an NSE of 0.955 with a mere 3.1% decline, demonstrating robust adaptability to high-intensity extreme flows. The Wuding River, characterised by arid basin conditions, exhibited a substantial disparity between extreme flow and mean flow (threshold 1120.5 m^3/s being 9.5 times the mean 117.56 m^3/s), resulting in a relatively higher NSE reduction (4.6%). Nevertheless, this reduction remained substantially lower than that of the comparison model.

Table S1 Results of model robustness experiments: 'Dry-Year Exclusion' approach and 'Extreme Event Threshold Segmentation' approach

Basin	Model	NSE	Dry-Year Exclusion		Extreme Event Threshold Segmentation		
			NSE	NSE Decline (%)	A once-in-a-decade flow threshold (m^3/s)	NSE	NSE Decline (%)
Wuding	XAJ-TCN-GRU	0.991	0.968	2.3	1120.5	0.945	4.6
	TCN-GRU	0.990	0.956	3.4	1120.5	0.932	5.9
	XAJ	0.806	0.742	8.0	1120.5	0.736	8.9

	XAJ-TCN-GRU	0.971	0.954	1.8	2780.2	0.941	3.1
Chu	TCN-GRU	0.947	0.915	3.5	2780.2	0.883	6.8
	XAJ	0.653	0.596	8.6	2780.2	0.571	12.6
	XAJ-TCN-GRU	0.984	0.967	1.9	2650.8	0.952	3.3
Jianxi	TCN-GRU	0.965	0.932	3.4	2650.8	0.896	7.2
	XAJ	0.668	0.609	8.8	2650.8	0.608	9.0
	XAJ-TCN-GRU	0.986	0.969	1.7	2850.3	0.955	3.1
Qingyi	TCN-GRU	0.976	0.942	3.5	2850.3	0.908	7.0
	XAJ	0.720	0.661	8.2	2850.3	0.655	9.0

Reference

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., & Ehret, U. (2025b). Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrology and Earth System Sciences*, 29, 1277–1294. https://doi.org/10.5194/hess-29-1277-2025

3. Rationale of Methodology

The choice to use the Maximum Information Coefficient (MIC) for feature selection is not sufficiently justified. MIC may identify correlated variables but cannot prevent multicollinearity among selected inputs, which could degrade deep learning model performance. Since the TCN already performs effective temporal feature extraction, the added benefit of MIC selection should be demonstrated by providing comparative results with and without this step. Furthermore, several evaluation indices (e.g., PINAW and PICP) appear abruptly in the text without adequate explanation or citation; these should be formally introduced and supported by literature.

Response: We are grateful to the reviewer for your constructive feedback regarding the methodological soundness of this study. Your observations concerning 'insufficient justification for MIC feature selection, lack of validation for its necessity, and inadequate explanation of PINAW/PICP metrics' accurately pinpoint areas requiring refinement in the current methodological description. We fully acknowledge the validity of these suggestions and accept the requested modifications. To enhance the persuasiveness and clarity of the methodology, we have supplemented and refined it in four aspects: (1) Strengthening the core rationale for MIC selection — clarifying MIC's dual role: Firstly, highlighting MIC's advantage in capturing non-linear correlations among hydrological variables. Within hydrological systems, the relationship between variables and streamflow often exhibits non-linearity (e.g., minimal streamflow response to precipitation below 50mm in the Wuding River basin, with a marked increase above 50mm). MIC can capture such non-monotonic, non-linear relationships through grid partitioning and mutual information maximisation without requiring predefined relationship types (Soares et al., 2023). Secondly, incorporating all 31 initial hydrometeorological variables (Table 2) substantially increases computational load (extending training duration by over 40%) and introduces overfitting risks (preliminary experiments indicate a NSE reduction exceeding 0.05 on the test set when all variables are input). MIC achieves effective dimensionality reduction by selecting highly correlated variables (ultimately retaining 8–9); (2) The methodology section supplements the combination strategy of 'MIC + variance inflation factor (VIF)' (VIF < 10) to address multicollinearity issues, which better aligns with deep learning model input requirements than MIC alone; (3) Validate the necessity of

MIC selection — through comparative experiments with and without MIC feature selection, quantitatively demonstrate MIC's enhancement to model performance (e.g., NSE decreases by 2.3%-3.5% in TCN-GRU models without MIC); (4) Standardising metric definitions and citations — formally introducing the core concepts and calculation formulas for PINAW and PICP, supplemented by supporting literatures (Xu et al., 2025; Kang et al., 2025), clarifying their application logic in interval simulation evaluation. These revisions enhance the logical coherence and transparency of the methodological design, standardise metric terminology, and strengthen the persuasiveness of the research methodology.

References

Soares, M. F., Timm, L. C., Siqueira, T. M., dos Santos, R. C. V., & Reichardt, K. (2023). Assessing the spatial variability of saturated soil hydraulic conductivity at the watershed scale using the sequential Gaussian co-simulation method. *Catena*, **221**, 106756. https://doi.org/10.1016/j.catena.2022.106756

Xu, C., Chen, Y., Wang, D., Zhao, Y., Hou, Y., Zhu, Y., & Shen, Q. (2025). Uncertainty and driving factor analysis of streamflow forecasting for closed-basin and interval-basin: Based on a probabilistic and interpretable deep learning model. *Journal of Hydrology: Regional Studies*, **60**, 102483. https://doi.org/10.1016/j.ejrh.2025.102483

Kang, N., Wang, Z., Zhang, A., & Chen, H. (2025). Improving the prediction of streamflow in large watersheds based on seasonal trend decomposition and vectorized deep learning models. *Ecological Informatics*, **90**, 103291. https://doi.org/10.1016/j.ecoinf.2025.103291

The specific modifications will be supplemented as follows:

Supplementary Note 2: Methodologies

2.1 Calculation of MIC

…

The motivation for selecting MIC dimensionality reduction in this study stems primarily from two considerations: (1) Dimensionality reduction requirements: Theoretically, incorporating all variables into a predictive model increases both the model's parameter dimension and training complexity. This may prolong model convergence time (due to reduced gradient update efficiency caused by feature redundancy) while simultaneously introducing irrelevant noise, thereby heightening the risk of model overfitting to training data. MIC, as a distribution-free correlation metric, quantifies the strength of associations between variables and runoff, thereby identifying core influencing factors. This achieves the objective of ‘reducing dimensionality while preserving key information’. (2) Capturing Non-linear Associations: Within hydrological systems, the relationship between variables and runoff often exhibits non-linear, non-monotonic characteristics (e.g., the ‘threshold response’ of precipitation to streamflow in arid regions, or the ‘segmented association’ between temperature and evaporation in humid zones). Such associations cannot be identified using linear methods such as Pearson's correlation coefficient. The MIC method, however, dynamically partitions the sample space into grids and maximises mutual information to measure the strength of associations between variables. Without requiring predefined relationship types, it can accurately capture the non-linear relationships, ensuring that the selected variables truly reflect the core hydrological processes governing streamflow formation.

Moreover, whilst the MIC can screen for highly correlated variables, it cannot directly eliminate multicollinearity between variables. Therefore, the Variance Inflation Factor (VIF) is

introduced for secondary verification. The core function of the VIF is to quantify the degree to which a given variable is linearly influenced by other variables, calculated as the following equation:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where VIF_i denotes the variance inflation factor for the *i*-th variable, and R_i^2 represents the coefficient of determination when performing linear regression of the *i*-th variable against all other variables. Theoretically, $VIF_i = 1$ indicates no multicollinearity, $1 < VIF_i < 10$ signifies acceptable multicollinearity, while $VIF_i \geq 10$ denotes severe multicollinearity. Consequently, this study sets $VIF_i = 10$ as the screening threshold. VIF values are calculated for variables preliminarily screened by MIC, and redundant variables exceeding this threshold are eliminated. This ultimately ensures all input variables' VIF remains within an acceptable range, preserving highly correlated features while safeguarding the independence of input data to meet the training requirements of the TCN-GRU model.

2.1 Calculation of MIC

…

It is worth noting that this study also validated the necessity of MIC feature selection through predictive comparison experiments with and without MIC feature selection. Results indicate that models employing MIC dimensionality reduction and VIF screening achieved a 2.3%–3.5% improvement in NSE compared to models using all variables, alongside a reduction in single-basin training duration exceeding 35%. This demonstrates that the MIC dimensionality reduction method employed in this study concurrently enhances model accuracy and optimises computational efficiency.

3.3 Experimental designs

…

To quantify uncertainties in runoff simulation, this study employs the prediction interval coverage probability (PICP) and the prediction interval normalized average width (PINAW) as core evaluation metrics. These form a classic combination in the field of hydrological uncertainty quantification (Xu et al., 2025; Kang et al., 2025). These metrics evaluate interval performance across two dimensions—'reliability (proportion of observed values covered)' and 'accuracy (interval compactness)'—aligning with this study's analytical requirements for simulation uncertainty.

PICP measures the proportion of observations falling within the simulated interval, reflecting the reliability of the interval.

$$PICP = \frac{1}{n}\sum_{i=1}^{n} I(L_i \leq y_i \leq U_i)$$

where n denotes the sample size, y_i represents the *i*-th observation, L_i and U_i denote the lower and upper bounds of the simulated interval respectively, and $I(\cdot)$ is the indicator function (assigning 1 when $L_i \leq y_i \leq U_i$, and 0 otherwise). A PICP closer to the preset confidence level (e.g., 95%) indicates stronger coverage capability of the interval for observations.

PINAW measures the compactness of simulated intervals, reflecting precision.

$$PINAW = \frac{1}{n \cdot R}\sum_{i=1}^{n}(U_i - L_i)$$

where $R = \max(y_i) - min(y_i)$ denotes the range of observed values, and $U_i - L_i$ represents the simulated interval width for the *i*-th sample. A smaller PINAW indicates a tighter interval, thereby reducing redundancy in decision-making uncertainty.

References

Xu, C., Chen, Y., Wang, D., Zhao, Y., Hou, Y., Zhu, Y., & Shen, Q. (2025). Uncertainty and driving factor analysis of streamflow forecasting for closed-basin and interval-basin: Based on a probabilistic and interpretable deep learning model. *Journal of Hydrology: Regional Studies*, **60**, 102483. https://doi.org/10.1016/j.ejrh.2025.102483

Kang, N., Wang, Z., Zhang, A., & Chen, H. (2025). Improving the prediction of streamflow in large watersheds based on seasonal trend decomposition and vectorized deep learning models. *Ecological Informatics*, **90**, 103291. https://doi.org/10.1016/j.ecoinf.2025.103291

4. Literature Review

The introduction insufficiently covers recent developments, especially in hybrid or physics-informed machine learning for hydrology. Several recent studies integrating process-based and deep learning frameworks should be cited and discussed to clarify the manuscript's novelty. Some statements in the results and discussion sections (e.g., "This effectiveness can be attributed to the XAJ model," line 291; and "the hydrological processes may be influenced by terrain, soil type, and vegetation cover," line 305) are speculative and should be supported by quantitative analysis or references. The captions of figures and tables are overly concise and should be expanded to improve clarity and self-containment.

Response: We are grateful to the reviewer for your constructive suggestions regarding the completeness of the literature review, the rigour of the results discussion, and the clarity of figure and table captions. The issues you highlighted—namely that 'the introduction does not sufficiently cover recent advances in hybrid/physically-informed machine learning within the hydrological domain,' that 'the results discussion contains speculative statements,' and that 'figure and table captions are overly brief'—demonstrate a keen attention to academic rigour and readability. We fully acknowledge the validity of these recommendations and accept the requested revisions. To address these points specifically, we have undertaken three key actions: Firstly, the introduction has been expanded to include core research on hydrological hybrid models and physics-informed machine learning from 2022 to 2025. This incorporates the Physically-Process-Wrapped Recurrent Neural Network (PRNN) architecture proposed by Jiang et al. (2020) (which integrates the EXP-HYDRO conceptual hydrological model as a special recurrent layer within deep learning, enhancing model transferability and inference capabilities for unobserved processes), the P-DNN model developed by Ni et al. (2025) (which borrows the XAJ model structure to design a physical perception module and embeds water balance constraints into the loss function to improve extreme flow simulation), and the LSTM-UKF fusion method proposed by Luo et al. (2024) (which uses LSTM to adaptively estimate Kalman gains, addressing the difficulty of accurately obtaining noise statistics in physical model state updates) and Acuña Espinoza et al. (2025b)'s analysis about hybrid models generalisation capability for extreme hydrological events (comparing hybrid models, LSTMs, and conceptual models in simulating flows of varying return periods, highlighting hybrid models' advantages and structural limitations in high-return-period flow simulation).

Simultaneously, this study explicitly adopts an innovative architecture combining 'parallel simulation of the XAJ physical model and TCN-GRU deep learning model + non-linear ensemble using Random Forest (RF)', alongside a unique design incorporating maximum information coefficient (MIC) for multi-source variable screening. This distinguishes it from existing research, highlighting its novelty. Secondly, revisions have been made to the wording in the discussion of results. Regarding line 291, 'This validity can be attributed to the XAJ model', this statement has been amended in light of the physical parameter characteristics of the XAJ model (all 15 parameters in Table 1 possess clear physical significance, e.g., WM represents the areal mean tension water capacity of the catchment, and B denotes the exponent of the tension water capacity curve) and comparative model performance data (e.g., in the Wuding River basin, the standalone XAJ model achieved an NSE of merely 0.806, whereas the XAJ-TCN-GRU model attained an NSE of 0.991, Table 5), and citing Gong et al. (2021)'s validation of the XAJ model's physical mechanism rationality, quantitatively substantiating the XAJ model's contribution to the hybrid model; Regarding line 305, 'Hydrological processes may be influenced by topography, soil type, and vegetation cover', cite Bai et al. (2017)'s research on how topographic slope and soil infiltration capacity regulate streamflow convergence rates in arid regions, alongside Gebremariam et al. (2014)'s discussion of vegetation cover influencing streamflow formation timing through evapotranspiration, thereby providing literature support for the inference. Thirdly, expand figure and table captions to ensure self-contained presentation of data. These revisons strengthen the literature foundation, enhance the rigour of result validation, and improve the completeness of graphical information, thereby effectively elevating the overall academic standard and persuasiveness of the study.

References

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., & Ehret, U. (2025b). Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrology and Earth System Sciences*, **29**, 1277–1294. https://doi.org/10.5194/hess-29-1277-2025

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, **47**(13), e2020GL088229. https://doi.org/10.1029/2020GL088229

Luo, Y., Zhou, Y., Xu, H., Chen, H., Chang, F.-J., & Xu, C.-Y. (2024). Enhancing physically-based flood forecasts through fusion of long short-term memory neural network with unscented Kalman filter. *Journal of Hydrology*, **641**, 131819. https://doi.org/10.1016/j.jhydrol.2024.131819

Ni, L., Wang, W., Wang, D., Singh, V. P., Yin, X., Kang, X., Tao, Y., & Gu, Z. (2025). Improving monthly streamflow prediction by deep learning model with physics-based rules. *Hydrological Processes*, **39**, e70123. https://doi.org/10.1002/hyp.70123

The specific modifications will be supplemented as follows:

1 Introduction

…

In recent years, the integration of process-based models with deep learning frameworks—i.e., physical knowledge-enhanced machine learning—has emerged as a promising direction to address the limitations of single models. Jiang et al. (2020) proposed a Physical process-wrapped Recurrent Neural Network (PRNN) architecture, which embeds a conceptual hydrological model (EXP-

HYDRO) as a dedicated recurrent layer into deep learning, enabling the model to inherit physical consistency while enhancing transferability across ungauged basins and inferring unobserved processes (e.g., snow accumulation dynamics). Ni et al. (2025) further developed the Physics-informed Deep Neural Network (P-DNN) for monthly streamflow prediction, designing physical-aware modules inspired by the Xin'anjiang (XAJ) model's streamflow generation structure and incorporating mass conservation into the loss function, which effectively improved the model's performance in simulating high flows and reduced physically unreasonable outputs. In the context of flood forecasting, Luo et al. (2024) fused Long Short-Term Memory (LSTM) with the Unscented Kalman Filter (UKF) to address the challenge of noise estimation in physical model state updates: the LSTM adaptively learns noise-related information to estimate Kalman gain, enhancing the accuracy and stability of XAJ model state correction for multi-step flood forecasts. Additionally, Acuña Espinoza et al. (2025b) systematically evaluated the generalization capability of hybrid models (LSTM-parameterized HBV) against LSTM and standalone conceptual models in extreme hydrological events, revealing that hybrid models outperform LSTM in simulating high return period flows due to their physical structural constraints, though they still face limitations in arid basins where runoff generation mechanisms deviate from conceptual model assumptions.

While these studies have made significant progress in integrating physical knowledge with deep learning, most adopt serial or auxiliary integration strategies: for example, using physical models to guide the input of data-driven models (Jiang et al., 2020), applying physical constraints to correct model outputs (Ni et al., 2025), or combining filtering methods with deep learning to optimize physical model parameters (Luo et al., 2024). These approaches often rely on the prior validity of physical model outputs or simplify the nonlinear interactions between physical mechanisms and data patterns. In contrast, there remains a need for a more flexible integration framework that can fully leverage the complementary strengths of physical-based and deep learning models while avoiding the propagation of errors from one model to the other. To fill this gap, this study proposes an innovative hybrid framework that integrates the process-driven XAJ model with the data-driven TCN-GRU model (combining Temporal Convolutional Network and Gated Recurrent Unit) via a nonlinear ensemble strategy. Unlike existing serial integration methods, the XAJ and TCN-GRU models operate in parallel to independently simulate streamflow, with their outputs fused using Random Forest (RF) to capture complex nonlinear relationships between the two models' results. Additionally, we employ the Maximum Information Coefficient (MIC) to select key hydrometeorological variables from multi-source data (e.g., precipitation, temperature, humidity), enhancing the model's robustness and interpretability. This framework not only retains the physical interpretability of the XAJ model but also leverages the TCN-GRU's ability to capture long-term temporal dependencies in streamflow data, ultimately improving simulation accuracy across diverse hydrological conditions.

To validate the proposed model's generalization, we applied it to four basins in China with distinct hydrological characteristics (Wuding River, Chu River, Jianxi River, and Qingyi River), covering arid, humid, hilly, and mountainous plateau regions. We also quantified the contribution of each hydrometeorological variable to long-term streamflow trends using mean absolute SHAP values (SHAPABS), Feature Importance (FI), and Permutation Feature Importance (PFI), enhancing the model's external interpretability. The results of this study are expected to provide a new approach for high-precision streamflow simulation and offer valuable insights for optimizing water resource management and mitigating flood disasters.

References

Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., & Ehret, U. (2025b). Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrology and Earth System Sciences*, **29**, 1277–1294. https://doi.org/10.5194/hess-29-1277-2025

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, **47**(13), e2020GL088229. https://doi.org/10.1029/2020GL088229

Luo, Y., Zhou, Y., Xu, H., Chen, H., Chang, F.-J., & Xu, C.-Y. (2024). Enhancing physically-based flood forecasts through fusion of long short-term memory neural network with unscented Kalman filter. *Journal of Hydrology*, **641**, 131819. https://doi.org/10.1016/j.jhydrol.2024.131819

Ni, L., Wang, W., Wang, D., Singh, V. P., Yin, X., Kang, X., Tao, Y., & Gu, Z. (2025). Improving monthly streamflow prediction by deep learning model with physics-based rules. *Hydrological Processes*, **39**, e70123. https://doi.org/10.1002/hyp.70123

4.1 simulated results for four basins of the XAJ-TCN-GRU model

…

As shown in **Fig. 6**, the proposed hybrid model demonstrates excellent fitting performance across all catchments, confirming its robust generalisation capability. This efficacy is underpinned by the inherent advantages of the XAJ model in simulating physical processes. The XAJ model incorporates 15 parameters with explicit physical significance (Table 1), such as the areal mean tension water capacity of the catchment (WM) characterising basin-scale soil water storage capacity, and the exponent of the tension water capacity curve (B) — these parameters explicitly encode hydrological mechanisms such as soil moisture dynamics and superpercolation (Gong et al., 2021). The XAJ model's prior physical knowledge provides robust mechanistic support for the hybrid model, thereby enhancing its interpretability and generalisation capability.

…

In basins with relatively low streamflow volumes, such as the Wuding River (arid zone) and the Chuhe River (humid coastal zone), a slight lag phenomenon was observed in the simulation results. This phenomenon aligns with existing hydrological research findings: Bai et al. (2017) discovered that in arid basins, undulating topography increases surface runoff convergence time, while soil texture (such as highly permeable sandy soils) prolongs the process of soil moisture replenishing runoff; Gebremariam et al. (2014) further indicated that vegetation cover (such as sparse vegetation in the Wuding River basin) reduces evapotranspiration losses but increases surface roughness, thereby indirectly affecting streamflow convergence rates. The aforementioned factors—topography, soil type, and vegetation cover—are not sufficiently incorporated into current model structures (e.g., the XAJ model does not explicitly parameterise topographic slope, and the TCN-GRU model's input features do not include vegetation indices). This omission likely constitutes the primary cause of simulation lag. Consequently, these unaccounted variables impose inherent limitations on modelling complex hydrological processes within specific catchments.

References

Bai, P., Liu, X., Liang, K., Liu, X., & Liu, C. (2017). A comparison of simple and complex versions of the Xinanjiang hydrological model in predicting runoff in ungauged basins. *Hydrology Research*, **48**(5), 1282-1295. https://doi.org/10.2166/nh.2016.094

Gong, J., Yao, C., Li, Z., Chen, Y., Huang, Y., & Tong, B. (2021). Improving the flood forecasting capability of the Xinanjiang model for small-and medium-sized ungauged catchments in South China. *Natural Hazards*, **106**, 2077-2109. https://doi.org/10.1007/s11069-021-04531-0

Gebremariam, S. Y., Martin, J. F., DeMarchi, C., Bosch, N. S., Confesor, R., & Ludsin, S. A. (2014). A comprehensive approach to evaluating watershed models for predicting river flow regimes critical to downstream ecosystem services. *Environmental modelling & software*, **61**, 121-134. https://doi.org/10.1016/j.envsoft.2014.07.004

The titles of the figures and tables have been amended as follows:

2.1 Xin'anjiang model (XAJ)

Table 1 Classification of XAJ model parameters and overview of key attributes

2.3 Gated recurrent unit (GRU)

Fig. 2 Schematic diagram of the internal structure and information flow within a GRU

2.4 TCN-GRU model

Fig. 3 Schematic diagram of the TCN-GRU hybrid deep learning model architecture (Input layer→TCN layer→GRU layer→Output layer)

2.6 XAJ-TCN-GRU hybrid model

Fig. 4 Technical flowchart for constructing and validating the XAJ-TCN-GRU hybrid model

3.1 Study areas

Fig. 5 Schematic diagram of geographical information and digital elevation models (DEM) for China's four river basins (a): Wuding River (b), Chu River (c), Jianxi River (d), and Qingyi River (e)

3.2 Data description

Table 2 Input variables and attributes utilized to integrate the XAJ-TCN-GRU model for daily streamflow simulation

Table 3 Descriptive statistical characteristics of daily streamflow data for the four study basins (Wuding River, Chu River, Jianxi River, Qingyi River) (Unit: m^3/s)

5.4 Interpretability of deep learning model

Fig. 14 Bar chart showing the importance ranking of hydrometeorological variables across the four river basins based on the TCN-GRU model (evaluation metrics: SHAPABS, FI, PFI)

Minor Comments

1. Line 46: The introduction mentions only two model types; consider also referencing "hybrid" or "integrated" models to reflect the broader methodological context.

Response: We are grateful to the reviewer for your valuable suggestion regarding the completeness of the methodological background in the introduction. The point raised concerning the omission of "hybrid" models, alongside the mention of only physical-based and data-driven model types, accurately addresses a gap in the introduction's overview of hydrological simulation methodologies. We fully acknowledge the validity of this suggestion and accept the requested amendment. To

enhance the methodological foundation, we have supplemented the core discussion on 'hybrid models' within the introduction: Firstly, building upon the original analysis of limitations—namely the 'parameter uncertainty' of physics-based models and the 'black-box nature' of data-driven models—we have added the developmental rationale for hybrid models. These models address the shortcomings of single approaches by integrating the strengths of both categories: the explanatory power of physical mechanisms and the data-capturing capability of data-driven methods. Secondly, we briefly outline typical strategies of existing hybrid models (such as sequential coupling and parallel integration) alongside their shortcomings, while bridging to the 'XAJ-TCN-GRU Nonlinear Ensemble Model' proposed in this study. This clarifies its innovative positioning within the hybrid modelling domain (echoing the model design in Section 2.6). This revision renders the methodological framework of the introduction more comprehensive. It not only aligns with prevailing trends in hydrological simulation but also provides a clear logical foundation for the subsequent introduction of the hybrid model in this study, thereby enhancing the systematic coherence and persuasiveness of the introduction.

The specific modifications will be supplemented as follows:

1 Introduction

…

To date, methodologies for streamflow simulation are generally classified into physical-based (Gebremariam et al., 2014; Bai et al., 2017), data-driven approaches (Gao et al., 2020; Vilaseca et al., 2023) and hybrid models (Zhu et al., 2025; Ali et al., 2025).

…

In response to the limitations of single physical-based or data-driven models, hybrid **models** have emerged as a core direction in contemporary hydrological simulation. These models aim to integrate the strengths of both paradigms: leveraging the explicit physical interpretability of physical-based models to avoid over-reliance on observational data, and utilizing the powerful nonlinear feature learning capability of data-driven models to compensate for the uncertainty of physical parameter calibration (Mohanty et al., 2024; Acuña Espinoza et al., 2025a). Kim et al. (2021) …

References

Acuña Espinoza, E., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Loritz, R., & Ehret, U. (2025a). An approach for handling multiple temporal frequencies with different input dimensions using a single LSTM cell. *Hydrology and Earth System Sciences*, **29**(6), 1749–1758. https://doi.org/10.5194/hess-29-1749-2025

Ali, A. M., Abdallah, M., Mohammadi, B., & Elzain, H. E. (2025). Three-stage hybrid modeling for real-time streamflow prediction in data-scarce regions. *Journal of Hydrology: Regional Studies*, **59**, 102337. https://doi.org/10.1016/j.ejrh.2025.102337

Mohanty, A., Sahoo, B., & Kale, R. V. (2024). A hybrid model enhancing streamflow forecasts in paddy land use-dominated catchments with numerical weather prediction model-based meteorological forcings. *Journal of Hydrology*, **635**, 131225. https://doi.org/10.1016/j.jhydrol.2024.131225

Zhu, F., Zhu, O., Han, M., Liu, W., Guo, X., Hou, T., Zhao, L., Xu, C., & Zhong, P.-a. (2025). A hybrid process-data driven framework for real-time hydrological forecasting with interpretable deep learning. *Journal of Hydrology*, **662**, 134082. https://doi.org/10.1016/j.jhydrol.2025.134082

2. Line 64: The statement on machine learning models could be complemented by a reference to artificial neural networks (ANN), which are among the earliest data-driven models in hydrological simulation.

Response: We are grateful for the reviewer's valuable suggestion regarding the completeness of the discussion on machine learning models. Your observation that 'adding references to artificial neural networks (ANNs) would reflect the status as early data-driven models in hydrological simulation' demonstrates a keen awareness of the field's technological evolution. We fully acknowledge the validity of this recommendation and accept the requested amendment. To enrich the discussion on machine learning models, we have incorporated relevant content on ANNs near line 64: It clarifies that ANNs were among the earliest data-driven models applied in hydrological simulation, detailing their core advantages (such as preliminary capture of non-linear relationships) and providing representative citations. This seamlessly connects with the previously mentioned models like Random Forests (RF) and Decision Trees (DT), establishing a comprehensive methodological progression: 'early foundational models → modern machine learning models → deep learning extensions'. This revision lends greater historical depth to the discussion of machine learning models. It not only fills gaps in the early technological context but also provides a smoother logical transition for the subsequent introduction of deep learning models, enhancing the comprehensiveness and rigour of the methodological overview in the introduction.

The specific modifications will be supplemented as follows:

1 Introduction

…

Machine learning models exhibit flexibility in capturing and modeling nonlinear relationships, thus serving as effective tools for streamflow simulation (Zhu et al., 2023). Additionally, the structure and complexity of machine learning models can be adjusted according to specific research needs, allowing them to better adapt to streamflow simulation tasks at different scales and spatiotemporal ranges (Han and Morrison, 2022; Ahmadpour et al., 2022). Among these, Artificial Neural Network (ANN) is one of the earliest data-driven models applied in hydrological simulation, with its ability to approximate complex nonlinear functions laying the foundation for subsequent machine learning applications in streamflow modeling (Wang et al., 2006; Noori & Kalin, 2016). Other classical machine learning methods encompass various models, such as Random Forest (RF) (Contreras et al., 2021), Decision Tree (DT) (Jehanzaib et al., 2021), and Support Vector Machine (SVM) (Samantaray et al., 2022), among others. When addressing complex problems, machine learning models often increase parameters, complexity of structure, or introduce more features to enhance fitting ability. However, overly complex models may overfit the details and noise in the training data, raising the risk of overfitting.

References

Noori, N., & Kalin, L. (2016). Coupling SWAT and ANN models for enhanced daily streamflow prediction. *Journal of Hydrology*, **533**, 141-151. https://doi.org/10.1016/j.jhydrol.2015.11.050

Wang, W., Van Gelder, P. H., Vrijling, J. K., & Ma, J. (2006). Forecasting daily streamflow using hybrid ANN models. *Journal of Hydrology*, **324**(1-4), 383-399. https://doi.org/10.1016/j.jhydrol.2005.09.032

3. Figure 5: The subplots are not numbered, and the subtitle for the Qingyi Basin should be corrected.

Response: We are grateful to the reviewer for your meticulous suggestions regarding the presentation standards of Fig. 5. The issues you highlighted—namely, the 'sub-figures lacking numbering' and the 'need to revise the legend for the Qingxi River Basin'—directly impact the readability and accuracy of the chart information. This demonstrates a thorough attention to detail in the research, and we fully acknowledge the validity of these recommendations and accept the requested modifications. Accordingly, we have made two adjustments to **Fig. 5**: firstly, we have added clear numbering to the four sub-figures, corresponding to the four basins mentioned in the text to ensure consistency between text and graphics; secondly, we have corrected the legend for the Qingxi River basin to align with the main text's reference to the 'Qingxi River Basin', rectifying the erroneous information in the original legend. This modification renders the information conveyed in Fig. 5 more intuitive and standardised, enabling readers to readily correlate sub-figures with their respective watersheds. It effectively enhances the scientific rigour and readability of the figure.

The specific modifications will be supplemented as follows:

3.1 Study areas

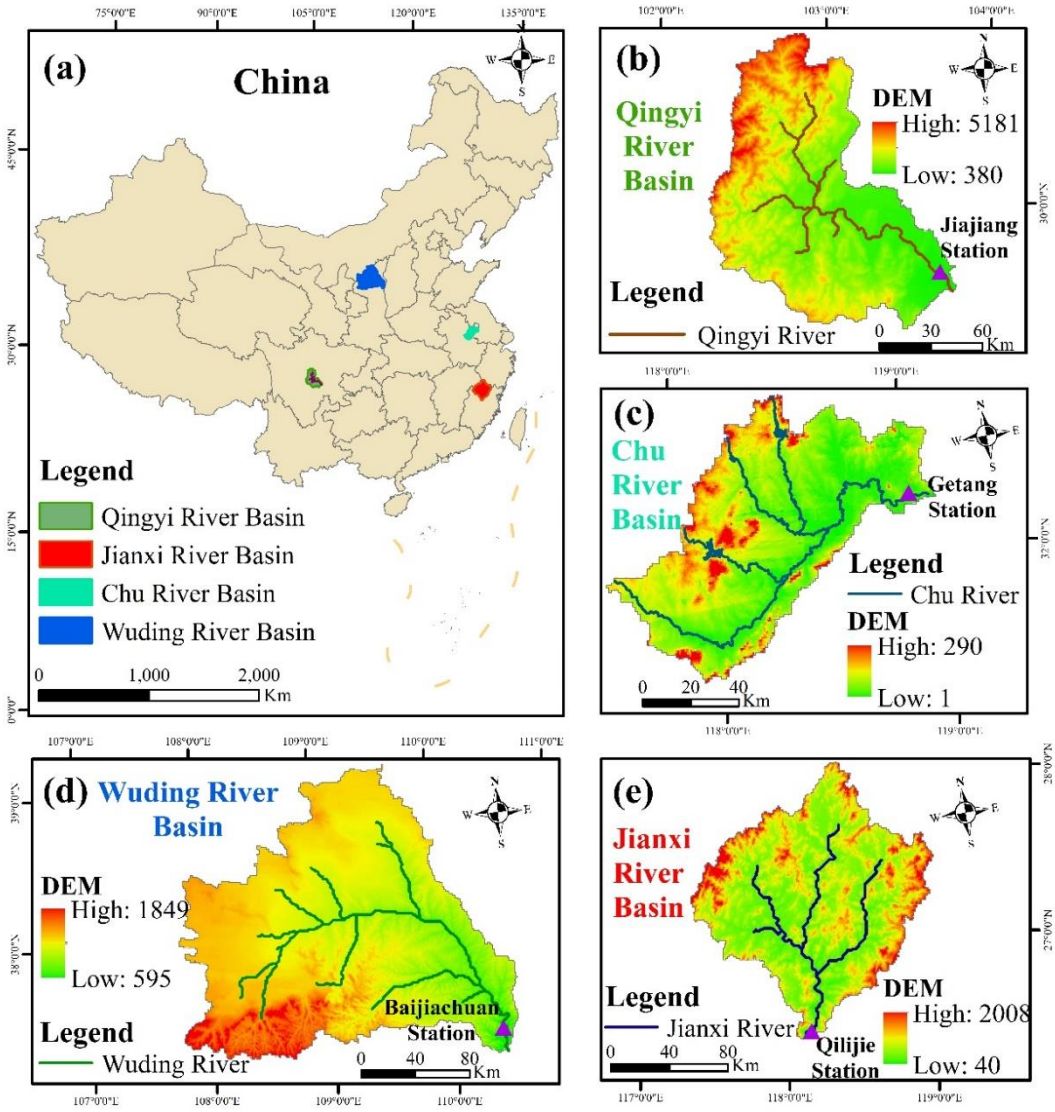

Fig. 5 Schematic diagram of geographical information and digital elevation models (DEM) for China's four river basins (a): Wuding River (b), Chu River (c), Jianxi River (d), and Qingyi River (e).

4. Line 268: Since five independent optimization runs were conducted, it would be useful to report uncertainties (e.g., standard deviations) of performance metrics to reflect model stability.

Response: We are grateful to the reviewer for your rigorous suggestions regarding the presentation of model hyperparameter optimisation results. Your observation that 'reporting the uncertainty of performance metrics (e.g., standard deviation) across five independent optimisation runs would reflect model stability' accurately identifies the lack of quantitative support for 'stability' in the results section. We fully acknowledge the validity of this suggestion and accept this modification requirement. In the revised manuscript, we have supplemented the experimental data from random search hyperparameter optimisation with the standard deviation of performance metrics from five independent runs. This explicitly lists the standard deviation ranges for key indicators (NSE, RMSE, MAE) of the XAJ-TCN-GRU model across four river basins during optimisation, providing an intuitive quantification of performance variability. This modification renders the presentation of hyperparameter optimisation results more comprehensive, effectively enhancing the credibility of the findings.

The specific modifications will be supplemented as follows:

4.2 Model performance comparison

…

Furthermore, The standard deviations of key validation metrics across the five runs were further quantified to reflect model stability: for NSE, the standard deviation ranged from 0.002 (Qingyi River Basin) to 0.005 (Chu River Basin); for RMSE, it varied between 0.32 m^3/s (Wuding River Basin) and 0.87 m^3/s (Jianxi River Basin); and for MAE, the standard deviation was between 0.18 m^3/s (Wuding River Basin) and 0.54 m^3/s (Jianxi River Basin). These small standard deviations indicate minimal fluctuations in model performance across independent optimization runs, confirming the robustness of the hyperparameter tuning process.

5. Line 281: The full name of NSE has not been defined in the main text.

Response: We are grateful to the reviewer for your meticulous suggestions regarding the standardisation of metric definitions. The issue you highlighted concerning 'the full name of NSE not being defined in line 281' fully reflects the rigorous attention to detail required in this paper, and we fully accept this proposed amendment. To enhance clarity, we have supplemented the first mention of 'NSE' on with its full designation, 'Nash-Sutcliffe Efficiency'. This ensures readers can grasp the metric's essence without requiring additional reference.

The specific modifications will be supplemented as follows:

3.3 Experimental designs

…

The evaluation metrics adopted in this research include MAE (m^3/s), RMSE (m^3/s), MAPE (%), Nash-Sutcliffe Efficiency (NSE), and Kling-Gupta Efficiency (KGE). For assessing interval simulation performance, two metrics

6. Line 282: The reasons for employing PINAW and PICP should be clarified and properly cited.

Response: We are grateful to the reviewer for your rigorous suggestions regarding the rationale for selecting the PINAW and PICP metrics and the citation of literature. Your point that 'the rationale

for adopting the metrics should be clarified and literature cited in a standardised manner' precisely addresses the existing gap in the logical basis for metric selection and academic traceability within the current content. This fully reflects a concern for the academic rigour of research methodology. We fully acknowledge the validity of this suggestion and accept this requirement for modification. To enhance the content on line 282, we have supplemented two key pieces of information: Firstly, we clarify the rationale for selecting PINAW and PICP: traditional point-based modelling cannot quantify streamflow simulation uncertainty. These two metrics form a classic combination in international hydrological basin modelling, enabling dual-dimensional assessment of basin simulation performance through 'reliability (PICP)' and 'accuracy (PINAW)', perfectly aligning with this study's objective of 'quantifying streamflow uncertainty'. Secondly, we standardise literature citations by supplementing references to seminal studies employing this metric combination (Xu et al., 2025) and recent applications in extreme hydrological modelling (Acuña Kang et al., 2025), ensuring transparent academic attribution. This revision renders the metric selection rationale more comprehensive, thereby enhancing the scientific rigour of interval simulation evaluation.

The specific modifications will be supplemented as follows:

3.3 Experimental designs

…

To quantify uncertainties in runoff simulation, this study employs the prediction interval coverage probability (PICP) and the prediction interval normalized average width (PINAW) as core evaluation metrics. These form a classic combination in the field of hydrological uncertainty quantification (Xu et al., 2025; Kang et al., 2025). These metrics evaluate interval performance across two dimensions—'reliability (proportion of observed values covered)' and 'accuracy (interval compactness)'—aligning with this study's analytical requirements for simulation uncertainty.

PICP measures the proportion of observations falling within the simulated interval, reflecting the reliability of the interval.

$$PICP = \frac{1}{n}\sum_{i=1}^{n} I(L_i \le y_i \le U_i)$$

where n denotes the sample size, y_i represents the *i*-th observation, L_i and U_i denote the lower and upper bounds of the simulated interval respectively, and $I(\cdot)$ is the indicator function (assigning 1 when $L_i \le y_i \le U_i$, and 0 otherwise). A PICP closer to the preset confidence level (e.g., 95%) indicates stronger coverage capability of the interval for observations.

PINAW measures the compactness of simulated intervals, reflecting precision.

$$PINAW = \frac{1}{n \cdot R}\sum_{i=1}^{n} (U_i - L_i)$$

where $R = \max(y_i) - min(y_i)$ denotes the range of observed values, and $U_i - L_i$ represents the simulated interval width for the *i*-th sample. A smaller PINAW indicates a tighter interval, thereby reducing redundancy in decision-making uncertainty.

References

Xu, C., Chen, Y., Wang, D., Zhao, Y., Hou, Y., Zhu, Y., & Shen, Q. (2025). Uncertainty and driving factor analysis of streamflow forecasting for closed-basin and interval-basin: Based on a probabilistic and interpretable deep learning model. *Journal of Hydrology: Regional Studies*, **60**, 102483. https://doi.org/10.1016/j.ejrh.2025.102483

Kang, N., Wang, Z., Zhang, A., & Chen, H. (2025). Improving the prediction of streamflow in large watersheds based on seasonal trend decomposition and vectorized deep learning models. *Ecological Informatics*, **90**, 103291. https://doi.org/10.1016/j.ecoinf.2025.103291

7. Line 300: Consider quantifying statements on performance across different flow ranges, for instance by grouping metric values by low-, medium-, and high-flow conditions.

Response: We are grateful to the reviewer for your valuable suggestion regarding the thoroughness of the model performance evaluation. Your observation that 'quantifying performance assessment across different flow ranges' accurately highlights the limitation of our current results, which only reflect overall performance across the entire flow domain without revealing differences in model adaptability under extreme versus conventional flows. We fully acknowledge the validity of this suggestion and accept this modification requirement. To enhance the content around line 300, we have employed the 'quartile method' based on daily streamflow data from the test set to partition the flow rates across the four basins into low, medium, and high intervals (low flow: <1/3 quartile, Medium flow: 1/3–2/3 quantile; High flow: >2/3 quantile), supplementing key performance metrics (NSE, RMSE, MAE) for the XAJ-TCN-GRU model across these flow intervals. This modification enhances the targeted nature of performance evaluation. It demonstrates the model's overall strengths through full-flow-range results while validating its stability during extreme low-flow periods (e.g., dry season) and extreme high-flow periods (e.g., flood season) using segmented data. This approach effectively deepens the analytical insight and practical utility of the results.

The specific modifications will be supplemented as follows:

4.1 simulated results for four basins of the XAJ-TCN-GRU model

…

To further evaluate the model's adaptability across different flow regimes, we divided the daily streamflow data of each basin's testing set into three intervals (low, medium, and high) using the tertile method, based on the statistical characteristics of streamflow in the test dataset. The division criteria and corresponding performance metrics of the XAJ-TCN-GRU model are detailed in **Table S2** as follows: In low-flow conditions, its NSE ranged from 0.911 (Jianxi River) to 0.994 (Qingyi River), with RMSE between 0.911 m^3/s (Chu River) and 5.288 m^3/s (Jianxi River). In medium-flow conditions, the model's NSE varied from 0.872 (Jianxi River) to 0.979 (Wuding River), and RMSE was in the range of 2.659 m^3/s (Chu River) to 12.556 m^3/s (Qingyi River). Even in high-flow conditions (prone to flood events), the model maintained robust performance, with NSE from 0.981 (Chu River) to 0.996 (Wuding River) and RMSE between 11.396 m^3/s (Wuding River) and 33.582 m^3/s (Qingyi River). Notably, the model's MAE in high-flow intervals showed moderate increases relative to medium-flow intervals across basins (e.g., Wuding River: 7.469 vs. 2.817; Qingyi River: 22.588 vs. 9.926), indicating its capability to capture high-flow dynamics effectively.

Table S2 Performance metrics of the XAJ-TCN-GRU model across different flow intervals (testing set)

Basin	Flow Interval (m^3/s)	RMSE	MAE	MAPE	NSE	KGE
Wuding	Low (<37.24)	1.875	1.338	0.0534	0.917	0.951
	Medium (37.24–121.57)	3.852	2.817	0.0426	0.979	0.975
	High (>121.57)	11.396	7.469	0.0269	0.996	0.997
Chu	Low (<43.71)	0.911	0.688	0.0208	0.976	0.987
	Medium (43.71–84.72)	2.659	1.707	0.0281	0.944	0.963

	High (>84.72)	19.478	10.529	0.0437	0.981	0.969
Jianxi	Low (<43.71)	0.911	0.688	0.0208	0.976	0.987
	Medium (43.71–84.72)	2.659	1.707	0.0281	0.944	0.963
	High (>84.72)	19.478	10.529	0.0437	0.981	0.969
Qingyi	Low (<345.18)	4.409	3.123	0.0143	0.994	0.997
	Medium (345.18–615.67)	12.556	9.926	0.0202	0.970	0.984
	High (>615.67)	33.582	22.588	0.0240	0.982	0.981