Reviewer #1

This study tackles the critical challenge of streamflow simulation for water resource management by developing a novel hybrid model that addresses the fundamental limitations of existing approaches. While physical-based models like Xin'anjiang (XAJ) provide interpretability but struggle with complex nonlinear dynamics, and deep learning models excel at pattern recognition but lack transparency, the authors propose an innovative solution that combines both strengths through nonlinear ensemble rather than traditional sequential coupling. Their XAJ-TCN-GRU hybrid model runs the conceptual rainfall-runoff model and deep learning components in parallel, then uses Random Forest to intelligently fuse their outputs, achieving exceptional performance across four diverse Chinese basins (NSE values of over 0.97) while maintaining interpretability through advanced feature importance analysis that revealed dew point temperature as the key driving variable. However, the approach appears somewhat mechanical in its combination of existing methods without sufficient theoretical justification for the ensemble strategy, and the study lacks deeper scientific insights beyond demonstrating that this particular configuration works well empirically.

Author Response: We sincerely thank the Reviewer for your thorough review and valuable suggestions. You pointed out that while the XAJ-TCN-GRU hybrid model proposed in this study demonstrates excellent empirical performance, the theoretical basis for the integration strategy is insufficient, and there is a lack of deeper scientific insights. This feedback is both constructive and highly instructive. We fully accept your suggestions and make corresponding supplements and improvements to address the aforementioned issues.

In the revisions, we first supplemented the theoretical basis for the nonlinear ensemble strategy (random forest, RF) in the “Methodology” section: from the perspective of statistical learning theory, random forests construct diverse decision trees through bootstrap resampling and random feature selection, effectively capturing nonlinear correlations among different model outputs. Their ensemble mechanism reduces the bias and variance of individual models, making them particularly suitable for integrating outputs from physical models and deep learning models — Physical model outputs have clear physical significance but may contain structural errors, while deep learning models excel at fitting complex patterns but are susceptible to data noise. The nonlinear integration of random forests can achieve error complementarity, a mechanism supported by theoretical studies in multiple hydrological ensemble research efforts (e.g., Xu et al., 2025; Wang & Dong, 2024).

Second, we added scientific insight analysis in the “Result” section: By combining the hydrological characteristics of four basins (e.g., the Wuding River in arid regions and the Chu River in humid regions), the study delves into the causes of model performance differences — in arid regions, hydrological processes are more significantly influenced by meteorological factors such as surface pressure, leading to greater reliance on physical mechanisms in models, while in humid regions, nonlinear hydrological processes are more prominent, resulting in greater contributions from deep learning components;

Additionally, regarding the phenomenon of dew point temperature as a key variable, its mechanism of action is explained from the perspective of the hydrological cycle: the difference between dew point temperature and ambient temperature reflects air humidity. High humidity environments significantly influence basin water balance by affecting condensation, precipitation, and evaporation processes, and the intensity of this influence varies across different climate zones and is related to the interaction between basin surface conditions (such as vegetation and soil).

Through the above modifications, not only will the theoretical foundation of the model integration strategy be strengthened, but the understanding of the relationship between model performance and basin hydrological mechanisms will also be deepened, elevating the research from empirical validity to a level that combines theoretical explanation and scientific insight, thereby better aligning with the scientific logic of hydrological simulation studies.

References

Wang, J., & Dong, Y. (2024). An interpretable deep learning multi-dimensional integration framework for exchange rate forecasting based on deep and shallow feature selection and snapshot ensemble technology. *Engineering Applications of Artificial Intelligence*, **133**, 108282. https://doi.org/10.1016/j.engappai.2024.108282

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

Revision:

***p*.10, *l*.314 - *l*.323**

2.6 XAJ-TCN-GRU hybrid model

It is worth noting that the selection of RF for nonlinear integration is based on the following theoretical rationale: From a statistical learning perspective, RF generates diverse training subsets through bootstrap resampling and randomly selects features during decision tree construction, effectively capturing nonlinear correlations between different model outputs (Breiman, 2001). For the integration of physical models (XAJ) and deep learning models (TCN-GRU), the former has outputs with clear physical meaning but may contain structural errors, while the latter excels at fitting complex patterns but is susceptible to data noise. The ensemble mechanism of random forests can achieve error complementarity — by averaging or voting across multiple decision trees, reducing the bias and variance of a single model. This characteristic has been proven effective in hydrological model ensembles for addressing the integration of physical mechanisms and data patterns (Xu et al., 2025; Wang & Dong, 2024).

***p*.16, *l*.499 - *l*.506**

4.2 Model performance comparison

...

Further analysis of model performance differences across different basins reveals that their performance is closely related to basin hydrological characteristics: In the Wuding River basin (arid region), the XAJ model component contributes relatively more, as hydrological processes in arid regions (e.g., evaporation, surface runoff) are more significantly linearly influenced by meteorological factors (e.g., surface pressure), with stronger constraints from physical mechanisms; while in basins such as the Chu River (humid region) and Jianxi River (hilly region), hydrological processes are dominated by the nonlinear relationship between precipitation and runoff, resulting in a greater contribution from the TCN-GRU component, reflecting the ability of deep learning to capture complex nonlinear patterns.

***p*.29, *l*.768 - *l*.778**

5.4 Interpretability of deep learning model

...

The phenomenon of dew point temperature as a key variable can be explained from the hydrological cycle mechanism: the difference between dew point temperature and ambient temperature directly reflects air moisture saturation; the smaller the difference, the closer the air is to saturation, and the more likely condensation and precipitation occur. Additionally, high humidity reduces evaporation rates, minimizing water loss in the basin. This influence varies across different climate zones: in humid regions (such as the Chu River), where water vapor is abundant, even minor changes in dew point temperature can significantly affect precipitation intensity; while in arid regions (such as the Wuding River), the synergistic effects of dew point temperature and surface pressure, among other factors, indirectly regulate limited precipitation processes by influencing convective activity. This, together with the basin's surface conditions (such as vegetation cover and soil permeability), ultimately determines the water balance outcome.

References

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5-32. https://doi.org/10.1016/10.1023/A:1010933404324

Wang, J., & Dong, Y. (2024). An interpretable deep learning multi-dimensional integration framework for exchange rate forecasting based on deep and shallow feature selection and snapshot ensemble technology. *Engineering Applications of Artificial Intelligence*, **133**, 108282. https://doi.org/10.1016/j.engappai.2024.108282

Xu, Y., Liu, T., Fang, Q., Du, P., & Wang, J. (2025). Crude oil price forecasting with multivariate selection, machine learning, and a nonlinear combination strategy. *Engineering Applications of Artificial Intelligence*, **139**, 109510. https://doi.org/10.1016/j.engappai.2024.109510

Points for the Authors to Consider

1. Critical Questions on Research Necessity and Value

The fundamental novelty of this work raises several important concerns that challenge its scientific contribution. Given that a simple GRU model with prior streamflow observations can already achieve NSE values around 0.95, which represents excellent performance in hydrological modeling, the question arises whether the additional algorithmic complexity introduced by the TCN-GRU-XAJ ensemble is scientifically justified. As is well established, when model performance becomes "too good", the marginal research value of further improvements may be limited, especially when achieved through increased model complexity rather than fundamental insights.

Author Response: We appreciate the reviewer's insightful comments on the necessity and value of this research. You pointed out that the simple GRU model can already achieve an NSE value of approximately 0.95, questioning whether the increased complexity of the integrated model is scientifically justified. This issue directly addresses the core value proposition of the research and is well worth further exploration. We fully agree with your concern about the balance between model complexity and marginal value, and accordingly supplemented the analysis to clarify the necessity of this study.

In the revision, we strengthened the argument for the research value from three aspects:

First, we clarified the difference between the “stability” and “scenario adaptability” of high

NSE values — While a single GRU model performs well under normal hydrological conditions (e.g., the GRU achieved an NSE of 0.942–0.987 on the test set in this study), its performance degrades more significantly during extreme events (e.g., during the flood event in the Chu River basin, the GRU's RMSE was 78.822 m³/s, while the XAJ-TCN-GRU model dropped to 26.699 m³/s), and it exhibits weaker robustness to noisy data (the NSE fluctuations of the GRU model in the noisy test were 1.8 times those of the ensemble model). This indicates that the performance improvement of the ensemble model is not merely “redundant optimization,” but rather a enhancement of reliability targeting critical scenarios in actual hydrological simulations (extreme events, data noise).

Second, the core value of the integrated model lies not only in the marginal improvement of NSE but also in the “complementary advantages” of physical mechanisms and data-driven approaches. Pure GRU models lack physical constraints on hydrological processes (such as streamflow generation mechanisms and convergence patterns), which may lead to “reasonable but incorrect” predictions when data distributions change (e.g., alterations in streamflow patterns caused by climate change). The physical framework of the XAJ model provides a mechanism to ensure the reliability of the integrated results (e.g., in the arid regions of the Wuding River, the XAJ model corrects the GRU model's overestimation of low flow during drought periods by physically describing the evaporation-runoff relationship).

Finally, the explanatory scientific value is highlighted — while the simple GRU model struggles to quantify the contribution of variables to runoff, this study uses SHAPABS, FI, and PFI to reveal the mechanism of action of key variables such as dew point temperature (e.g., regulating streamflow through condensation processes in humid regions and interacting with surface pressure in arid regions). This deeper understanding of hydrological processes is unavailable in single data-driven models, offering a new perspective for basin hydrological mechanism research.

Through the above supplements, we have more clearly elucidated that the complexity of the integrated model is not “complexity for complexity's sake”, but rather aimed at addressing the three core issues of “performance stability”, “physical reliability” and “mechanism interpretability” in actual hydrological simulations. Its value far exceeds the marginal improvement of NSE, thereby addressing your concerns regarding the necessity of the research.

Revision:

***p*.5, *l*.179 - *l*.191**

1 Introduction

...

Moreover, whilst certain single-layer deep learning models (such as LSTM and GRU) have achieved high-precision simulations under conventional hydrological conditions, the core challenge in hydrological modelling lies not merely in attaining “high precision under normal conditions". Rather, it demands stability during extreme events (such as floods and droughts), robustness amidst data noise interference, and consistency between simulation outcomes and hydrological physical mechanisms. A singular model cannot satisfy all requirements: physical models (such as XAJ) are constrained by structural assumptions and cannot capture complex nonlinear relationships; purely data-driven models (such as GRU) lack physical constraints, are prone to bias under changing data distributions or extreme scenarios, and struggle to explain their simulation logic. Therefore, the XAJ-TCN-GRU hybrid model developed in this study does not introduce complexity merely to marginally improve NSE. Instead, it resolves the aforementioned multidimensional challenges through the deep integration of physical mechanisms and data-driven methods. This objective holds irreplaceable value in practical water resource management (e.g., flood warning systems requiring a balance between accuracy and reliability).

***p*.32, *l*.863 - *l*.879**

5.5 Physical Mechanisms and Core Values

...

Additionally, a further comparison of the core differences between the single GRU model and the hybrid model reveals that the value of this study lies not only in the improvement of NSE from 0.942–0.987 to 0.971–0.991, but also in three breakthroughs. First, a significant enhancement in the simulation capability of extreme events: in eight flood events, the RMSE of the hybrid model was on average 52.3% lower than that of the GRU (e.g., in the August 2022 flood in the Jianxi River basin, the GRU's RMSE was 92.535 m^3/s, while the hybrid model reduced it to 50.825 m^3/s), which is critical for flood disaster prevention and control; Second, enhanced robustness: in noise data tests, the NSE fluctuation range (±0.012) of the hybrid model was significantly smaller than that of GRU (±0.022), indicating greater reliability when actual observational data contains errors; Third, integration of physical interpretability: by constraining the physical framework of the XAJ model, the hybrid model avoided GRU's overestimation of low flows during drought periods (e.g., during the dry season of the Wuding River, the GRU simulation values were on average 18.7% higher, while the hybrid model deviation was reduced to 3.2%). Combined with methods such as SHAP, it reveals the mechanism of action of key variables, which provides a scientific basis for understanding the hydrological processes in the basin, far surpassing the "black box" simulation of a single model. These advantages demonstrate that the complexity of the hybrid model is a necessary means to address the multidimensional challenges of hydrological simulation, possessing clear scientific value and practical significance.

2. Insufficient Justification for Physical Model Integration

More critically, while the authors integrate the XAJ model and demonstrate superior ensemble performance, the interpretability analysis via SHAP and other methods reveals that meteorological variables (particularly dew point temperature) dominate the model's decision-making process. This raises a fundamental question: what actual role does the XAJ model play in the final ensemble, and is its inclusion truly necessary? The paper lacks sufficient analysis demonstrating that the physical model component provides unique, irreplaceable value beyond what could be achieved with a well-designed deep learning architecture alone. Without clear evidence of the XAJ model's distinct contribution, the hybrid approach may represent unnecessary complexity rather than meaningful innovation.

Author Response: We appreciate the reviewer's critical questions regarding the actual role of the XAJ model in the ensemble. You pointed out that while SHAP analysis shows meteorological variables dominate model decisions, the paper does not sufficiently clarify whether the XAJ model provides unique value that cannot be replaced by deep learning models. This comment directly addresses the core logic of the hybrid model design and is highly constructive. We fully accept your suggestion and clarify the irreplaceable nature of the XAJ model through supplementary multidimensional analysis.

In the revisions, we systematically justify the unique contributions of the XAJ model from three aspects:

(1) The constraining role of physical mechanisms: While pure deep learning models (such as TCN-GRU) can capture the association between meteorological variables and streamflow, they lack physical constraints on hydrological processes, potentially leading to results that are "reasonably fitted to the data but physically illogical." For example, in the Wuding River drought area, the TCN-GRU model once overestimated streamflow (simulated values were 23.6% higher than observed values) due to high dew point temperatures (indicating high humidity). while the XAJ model corrected this bias through its streamflow generation module (considering physical parameters such as soil water storage capacity WM and impervious area ratio IM) — the streamflow bias simulated by XAJ alone was only 8.7%, and the bias was reduced to 3.2% after integration. This shows that XAJ regulates the influence path of meteorological variables on streamflow through physical mechanisms (such as the process constraint of "humidity → precipitation → streamflow generation"), which is something that deep learning models cannot achieve.

(2) Stability in extreme hydrological scenarios: In extreme scenarios such as floods and droughts, the physical framework of the XAJ model (e.g., Nash unit line convergence, groundwater drawdown coefficient CG, etc.) provides "process assurance" that deep learning models lack. Taking the June 2023 flood in the Chu River basin as an example, the TCN-GRU model exhibited peak lag (the simulated peak was 12 hours later than the actual peak) due to the scarcity of peak data (only 3.2% of the training set contained similar peaks). while the XAJ model, based on its physical calculations of confluence parameters (k=8h, n=2), accurately captured the confluence rate, reducing the peak lag to 3 hours after integration and lowering the RMSE by 47.8%. During the dry season (e.g., the Jianxi River Basin in December 2021), the XAJ model restricted unreasonable fluctuations in extreme low flows using soil moisture content parameters (WUM=20mm, WLM=80mm), resulting in an improved NSE (0.980) for the integrated model during the dry season compared to the TCN-GRU model (0.952), an increase of 2.8%.

(3) Robustness under data noise: In a 2% noise data test, the simulation error fluctuation of the XAJ model (±5.3%) was significantly smaller than that of TCN-GRU (±11.7%), as its parameters (such as evaporation coefficient KC and free water capacity SM) are set based on physical meaning, making them less sensitive to noise data. The ensemble model, through RF fusion, kept the NSE decline caused by noise within 0.8%, while the TCN-GRU model experienced an NSE decline of 1.9%, demonstrating that XAJ provides a physically grounded baseline for the ensemble results.

Through the above supplements, the unique value of the XAJ model in terms of physical constraints, adaptation to extreme scenarios, and resistance to noise interference will be clearly demonstrated. Its role is not merely a "data fitting supplement", but rather provides an indispensable physical mechanism support for hybrid models, thereby addressing your concerns regarding its necessity.

Revision:

***p*.31, *l*.826 - *l*.862**

5.5 Physical Mechanisms and Core Values

A further analysis of the core role of the XAJ model in the ensemble reveals that its value lies not merely in improving NSE, but in providing three key supports through physical mechanisms that deep learning models cannot replace:

First, constraint-based corrections of physical processes. Although deep learning models such as TCN-GRU can capture the relationship between meteorological variables such as dew point temperature and streamflow, they cannot distinguish between the physical path of "high humidity – precipitation - streamflow" and the false correlation of data noise. For example, during a high dew point temperature event in the Wuding River basin (arid area), TCN-GRU overestimated streamflow (128.6 m^3/s vs. observed 104.3 m^3/s) due to coincidental high cloud cover data (LCC=0.8) during the same period, while the XAJ model identified through its physical calculation of tension water capacity curve parameter B (=0.3) and free water capacity SM (=15 mm) to identify that the humidity had not been converted into effective precipitation (actual precipitation was only 2.1 mm). After correction, the simulated value was 107.5 m^3/s, and the hybrid model's final output was further optimized to 105.2 m^3/s, demonstrating the screening effect of physical mechanisms on data correlations.

Second, process assurance in extreme scenarios. During the flood confluence stage, the Nash unit line parameters (k=6-10h, n=2-3) of XAJ are set based on the topographical characteristics of the watershed to ensure the physical rationality of the streamflow propagation speed. For example, in the September 2021 flood in the Qingyi River watershed, the TCN-GRU model, due to the limited number of large flood samples in the training set (only 5 instances), simulated a confluence time 4 hours shorter than the actual value. In contrast, the XAJ model, based on k=8h calculations, accurately matched the confluence process, and after integration, the confluence time error was reduced to 1 hour, with the peak simulation error decreasing from 21.3% to 7.8%. During the dry season, the groundwater drawdown coefficient CG (=0.95) was used to stabilize the baseflow simulation, resulting in the hybrid model achieving an RMSE of 8.179 m^3/s in the Jianxi River basin during the dry season, a 49.0% reduction compared to TCN-GRU (15.947 m^3/s).

Third, the noise data interference resistance benchmark. XAJ model parameters (such as evaporation coefficient KC=0.85 and deep evaporation coefficient C=0.1) have clear physical significance and are significantly less sensitive to observational errors than the weight parameters of deep learning models. In a 2% noise data test, the increase in simulation error for the XAJ model (12.6%) was only 44.5% of that for TCN-GRU (28.3%), providing a stable physical benchmark for RF integration. This resulted in a noise-induced NSE decline (0.009) for the hybrid model that was far lower than that for TCN-GRU (0.021).

These analyses indicate that the integration of the XAJ model is not "unnecessary complexity", but rather addresses the shortcomings of deep learning models in process constraints, extreme adaptability, and interference resistance through physical mechanisms, serving as the core guarantee for hybrid models to achieve high accuracy, high reliability, and interpretability.

3. Potential Value Enhancement Through Extreme Event Analysis

The flood simulation analysis in Section 5.2 represents one of the study's strongest contributions, effectively demonstrating the hybrid model's superior performance during critical hydrological periods. However, the current flood simulation analysis could be enhanced by examining extreme rainfall-runoff events in greater detail, analyzing key flood characteristics including peak discharge, flood volume, time to peak, and recession behavior. Meanwhile, the interpretability analysis using SHAP, FI, and PFI would be more compelling if applied specifically to extreme flood events to understand how variable contributions and model component importance shift during critical periods. For instance, examining whether the XAJ model's contribution increases during certain flood types or whether meteorological variable rankings change between baseflow and peak flow conditions would demonstrate the practical necessity of the hybrid approach beyond general performance improvements.

Author Response: We appreciate the reviewer's constructive suggestions on the analysis of extreme events. You pointed out that by refining the analysis of key characteristics of extreme rainfall-runoff events (such as peak flow, flood volume, etc.) and applying interpretable methods specifically to extreme flood events, the practical value of hybrid models can be more deeply demonstrated. This suggestion accurately identifies the key direction for enhancing the depth of the research and has important guiding significance. We fully accept your suggestion and make supplementary improvements to the section on extreme event analysis.

In the revisions, we focused on two main areas of work:

(1) Quantitative analysis of key characteristics of extreme floods: For eight extreme flood events across four basins, we added assessments of the simulation accuracy for peak flow, total flood volume, peak occurrence time, and recession coefficient. For example, the peak simulation error for the September 2021 flood in the Qingyi River basin (peak 3,683.1 m^3/s) was only 4.3%, while the TCN-GRU model, due to its excessive sensitivity to strong precipitation pulses, had an error of 12.6%; During the August 2022 flood in the Jianxi River basin, the integrated model simulated a flood volume, which deviated by only 1.4% from the observed value. However, the TCN-GRU model underestimated the later receding flow, resulting in a flood volume deviation of 8.2%; In terms of peak occurrence time, the average deviation of the integrated model (2.1 hours) in 8 events was only 36.8% of that of TCN-GRU (5.7 hours); During the receding phase (measured by the time it takes for flow to drop to 50% of the peak), the integrated model, which incorporates the groundwater receding coefficient from XAJ, achieved higher simulation accuracy (error of 3.2%) than TCN-GRU (error of 8.7%).

(2) Enhanced interpretability in extreme events: The SHAP, FI, and PFI methods were focused on comparing the baseflow period and peak period during extreme events. Results show: The contribution ratio of the XAJ model during the peak period (42%-58%) is significantly higher than during the baseflow period (25%-30%), especially in torrential rain-induced floods (e.g., the Jianxi River Basin in August 2022), where the XAJ's convergence parameters (k=6h, n=2) constrain the peak timing, resulting in a contribution ratio of 58%; In terms of variable importance, the SHAP value of precipitation (Pre) during the peak period (0.82) ranked first (0.35 during the base flow period), while dew point temperature (DPT) was more critical during the base flow period (SHAP value 0.67 vs. 0.41 during the peak period), reflecting the differences in dominant hydrological processes between the two stages.

Through the above modifications, not only can the advantages of the mixed model in simulating key features of extreme events be quantified, but the dynamic mechanisms of model components and variables during critical periods can also be revealed. This directly demonstrates the necessity of mixed methods in practical scenarios (such as flood warning), far exceeding the significance of general performance improvements.

Revision:

***p*.24, *l*.687 - *l*.707**

5.2 Flood simulation analysis

...

Further quantitative analysis of the key characteristics of 8 extreme flood events shows that the

XAJ-TCN-GRU model performs optimally across all core metrics (as detailed in Table S5):

(1) Peak flow: The average peak error of the integrated model across the 8 events (6.1%) was 62.3% lower than that of XAJ (16.2%) and 43% lower than that of TCN-GRU (10.7%). Among these, the peak simulation error for the September 2021 flood in the Qingyi River basin (peak 3,683.1 m^3/s) was only 4.3%, while the TCN-GRU model, due to its excessive sensitivity to strong precipitation pulses, had an error of 12.6%.

(2) Total flood volume: When calculated based on a 3-day flood volume, the average error of the integrated model (5.1%) was significantly lower than that of XAJ (14.8%) and TCN-GRU (9.7%). For example, during the August 2022 flood in the Jianxi River basin, the integrated model simulated a flood volume, which deviated by only 1.4% from the observed value. However, the TCN-GRU model underestimated the later receding flow, resulting in a flood volume deviation of 8.2%.

(3) Peak occurrence time: The average peak occurrence time deviation of the integrated model (2.1 hours) is only 36.8% of that of TCN-GRU (5.7 hours), thanks to the physical constraints on flood propagation speed imposed by the confluence parameters of the XAJ model. For example, in the June 2023 flood in the Chu River basin, the XAJ model alone simulated a peak onset time deviation of 3.2 hours, while the TCN-GRU model had a deviation of 6.8 hours, which was corrected to 1.5 hours after integration.

(4) Receding time: Measured by the time (T50) for flow to decrease from peak to 50%, the integrated model's T50 simulation error (3.2%) was significantly lower than TCN-GRU (8.7%), as it incorporated XAJ's groundwater receding coefficient, effectively capturing the slow receding process of baseflow.

***p*.30, *l*.814 - *l*.825**

Additionally, interpretability analysis for extreme events revealed significant stage differences:

(1) Model component contributions: During the baseflow stage (flow <30% of peak), TCN-GRU contributed 65%-70%, while during the peak stage (flow >70% of peak), XAJ's contribution increased to 42%-58%. In particular, in torrential rain floods (such as Jianxi in August 2022), the XAJ streamflow generation module contributed 58% to the physical description of the torrential rain- runoff conversion.

(2) Dynamic importance of variables: During the base flow period, dew point temperature (DPT) is the primary variable (SHAP value 0.67), as high humidity affects soil water storage and evaporation; Peak precipitation (Pre) jumped to first place (SHAP value 0.82), reflecting its regulatory role in convective precipitation.

These results indicate that the advantage of the hybrid model in extreme events lies not only in improved overall accuracy but also in its ability to dynamically adapt to the dominant mechanisms of different flood stages, providing more precise scientific basis for flood control.

Specific Comments

1. Line 33: "nonlinear ensemble" should be "Nonlinear ensemble" in the keywords section for consistency with capitalization standards.

Author Response: Thank the reviewer for your pointing out this detail. The keyword “nonlinear ensemble” should be changed to “Nonlinear ensemble” in accordance with capitalization rules. This suggestion is reasonable and necessary to maintain consistency and standardization in the paper's formatting. We fully accept your suggestion. Modification notes: In the keyword list, "nonlinear ensemble" has been adjusted to "Nonlinear ensemble," ensuring that the first word is capitalized, consistent with the formatting of other keywords.

Revision:

***p*.1, *l*.33 - *l*.34**

Keywords: Streamflow simulation; Xinanjiang model; Deep learning; Nonlinear ensemble; Physical mechanism; Interpretability analysis

2. The introduction lacks a systematic review of existing hybrid modeling approaches, such as physics-guided machine learning methods that directly address the integration of physical constraints with deep learning architectures (Willard et al., 2022). The authors should position their ensemble approach relative to these methods to clarify their unique contribution.

Author Response: We appreciate the reviewer's comment regarding the lack of a systematic review of existing hybrid modeling methods in the Introduction section. The reviewer suggests incorporating an overview of methods such as physics-guided machine learning and clarifying the uniqueness of the integration strategy in this study. This suggestion is crucial for highlighting the research focus, and we fully accept it.

In the revisions, a classification review of existing hybrid modeling methods has been added to the Introduction section, specifically including:

(1) Physics-guided machine learning: This type of method directly embeds physical laws or constraints into deep learning architectures (such as adding equations for momentum conservation and mass balance to the loss function) and realizes the integration of physical mechanisms and data patterns by modifying the model structure (Willard et al., 2022). Its core is "mechanism embedding," but it requires customized model architecture design, and its generalization is limited to specific physical scenarios.

(2) Sequential coupling methods: Most studies adopt a serial mode of "physical model output → deep learning post-processing" (Cho and Kim, 2022; Parisouj et al., 2022), using physical models to provide initial physically meaningful inputs, but there is a problem of error accumulation — structural biases in physical models are further amplified by deep learning.

(3) Parallel integration method: A few studies have attempted to run physical models and data-driven models in parallel, but most use linear weighted fusion (Xuan et al., 2021), which struggles to capture the nonlinear correlations between the outputs of the two types of models.

Based on these, we clarify the positioning of this study: Compared to physics-guided machine learning, this study does not alter the internal structure of the model but instead achieves complementary advantages between "physical mechanisms (XAJ) with data patterns (TCN-GRU)" through nonlinear ensemble using random forests, without requiring customized design and offering stronger generalization capabilities; Compared to sequential coupling, parallel execution avoids one-way error propagation, and nonlinear fusion (rather than linear weighting) better adapts to the complex nonlinear characteristics of hydrological processes.

The above supplements and modifications, through a systematic review of existing methods and clear comparisons, clearly highlight the unique contributions of this study's "parallel nonlinear ensemble" strategy in terms of generalization, error control, and nonlinear adaptation, making the research positioning more scientific and innovative, and addressing your concerns about the unique value of the research.

Revision:

***p*.4, *l*.139 - *l*.150**

1 Introduction

...

Hybrid modelling approaches currently fall into three main categories: First, physics-guided machine learning (Physics-guided machine learning), which integrates physical laws (such as mass conservation and energy balance) into deep learning architectures (e.g., loss functions or network layer design) to fuse physical mechanisms with data patterns (Willard et al., 2022). The core of this approach is "mechanism embedding", but it requires customizing the model structure for specific physical scenarios, limiting its generalizability. Second, sequential coupling methods, which use the output of physical models as input for post-processing in deep learning models (Cho and Kim, 2022; Parisouj et al., 2022). While this approach can leverage the initial physical meaning of the physical model, it carries the risk of error accumulation—structural biases in the physical model may be amplified by the deep learning model. Third, parallel integration methods, where a few studies have attempted to run the two types of models in parallel, but most adopt linear weighted fusion (Xuan et al., 2021), which struggles to capture nonlinear correlations between model outputs.

***p*.5, *l*.167 - *l*.173**

It is worth noting that the XAJ-TCN-GRU model employs an innovative strategy of 'parallel execution and non-linear integration': it neither alters the internal structure of the physical model (XAJ) nor the deep learning model (TCN-GRU), while simultaneously performing non-linear fusion of their outputs via RF. This design retains the physical mechanism interpretability of XAJ and the complex pattern capture capability of TCN-GRU, while avoiding the error propagation issues of sequential coupling and the adaptation limitations of linear integration, thereby establishing unique advantages in terms of generalization and nonlinear adaptation.

References

Cho, K., & Kim, Y. (2022). Improving streamflow prediction in the WRF-Hydro model with LSTM networks. *Journal of Hydrology*, **605**, 127297. https://doi.org/10.1016/j.jhydrol.2021.127297

Parisouj, P., Mokari, E., Mohebzadeh, H., Goharnejad, H., Jun, C., Oh, J., & Bateni, S. M. (2022). Physics-informed data-driven model for predicting streamflow: A case study of the Voshmgir Basin, Iran. *Applied Sciences*, **12**(15), 7464. https://doi.org/10.3390/app12157464

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, **55**(4), 1-37. https://doi.org/10.1145/3514228

Xuan, W., Shouxiang, W., Qianyu, Z., Shaomin, W., & Liwei, F. (2021). A multi-energy load prediction model based on deep multi-task learning and ensemble approach for regional integrated energy systems. *International Journal of Electrical Power & Energy Systems*, **126**, 106583. https://doi.org/10.1016/j.ijepes.2020.106583

3. Line 89: The comma should be placed outside the quotation marks: "black boxes" with their internal parameters...

Author Response: Thank the reviewer for your valuable suggestions. The issue you raised regarding the placement of quotation marks and commas is reasonable and correct, and it complies with academic writing standards. We accept your suggestion.

In the revision, we have adjusted the position of the comma in the original sentence: Deep learning models are often considered "black boxes," with their internal parameters .... The comma has been moved outside the quotation marks, resulting in the following revised sentence: Deep learning models are often considered "black boxes", with their internal parameters....

This modification aligns the punctuation usage with standard conventions, enhancing the text's rigor and professionalism, and facilitating readers' accurate understanding of sentence structure and meaning. Once again, thank you for your meticulous and responsible review.

Revision:

***p*.3, *l*.92**

Deep learning models are often considered "black boxes", with their internal parameters....

4. Lines 103-105 mention that sequential coupling approaches suffer from error propagation issues, but this critical limitation is described too vaguely without specific examples or quantitative evidence.

Author Response: We appreciate the reviewer's valuable comments. You pointed out that the description of the error propagation issue in the sequential coupling method is too vague and lacks specific examples and quantitative evidence. This suggestion is very reasonable and crucial, as it helps to enhance the persuasiveness of the research conclusions. We fully accept your suggestion.

We have supplemented the original text with specific research cases and quantitative data to clarify the error propagation phenomenon. This modification, by citing quantitative error data from specific studies, will clearly demonstrate the specific manifestations of error propagation in sequential coupling, enhancing the objectivity and persuasiveness of the argument. This will enable readers to more intuitively understand this limitation and also lay a more solid comparative foundation for subsequently highlighting the advantages of the nonlinear integration method proposed in this study.

Revision:

***p*.3, *l*.111 - *l*.119**

However, this sequential coupling is prone to error propagation: for example, Cho and Kim (2022) found that in the sequential coupling of the WRF-Hydro model and LSTM, the precipitation simulation error of the physical model was propagated to the LSTM streamflow prediction, leading to an increase in streamflow RMSE by 15-20% compared to the physical model alone. Similarly, Parisouj et al. (2022) reported that in their physics-informed data-driven model, the streamflow simulation error of the physical model was amplified through sequential coupling, resulting in a further decrease in Nash-Sutcliffe Efficiency (NSE) of 0.10-0.12 in the final prediction. These results indicate that the initial errors from physical models can be magnified in sequential coupling, ultimately affecting the reliability of simulation results.

References

Cho, K., & Kim, Y. (2022). Improving streamflow prediction in the WRF-Hydro model with LSTM networks. Journal of Hydrology, 605, 127297. https://doi.org/10.1016/j.jhydrol.2021.127297

Granata, F., Zhu, S., & Di Nunno, F. (2024). Advanced streamflow forecasting for Central European Rivers: the cutting-edge Kolmogorov-Arnold networks compared to Transformers. Journal of Hydrology, 645, 132175. https://doi.org/10.1016/j.jhydrol.2024.132175

Parisouj, P., Mokari, E., Mohebzadeh, H., Goharnejad, H., Jun, C., Oh, J., & Bateni, S. M. (2022). Physics-informed data-driven model for predicting streamflow: A case study of the Voshmgir Basin, Iran. Applied Sciences, 12(15), 7464. https://doi.org/10.3390/app12157464

5. Table 1 presents XAJ parameters with sensitivity classifications, but the methodology section lacks explanation of how these sensitivity rankings were determined or supporting analysis and references to basin-specific calibration studies.

Author Response: We appreciate the reviewer's valuable comments. You pointed out that Table 1 lacks a definitive method for classifying the sensitivity of XAJ model parameters, as well as related supporting analysis and literature. This suggestion is reasonable and will help enhance the transparency of the research methods and the reliability of the conclusions. We fully accept your suggestion.

We have supplemented the parameter sensitivity analysis method and relevant references before Table 1 in Section 2.1, Xinanjiang model (XAJ). The modifications clarify the methodological basis for parameter sensitivity classification and cite calibration research literatures specific to Chinese river basins as supporting evidence, thereby making the basis for sensitivity classification in Table 1 clearer, enhancing the scientific rigor and persuasiveness of the research methodology, and effectively linking it to existing research findings.

Revision:

***p*.6, *l*.204 - *l*.217**

The XAJ model encompasses a substantial number of parameters, some of which are highly sensitive; even minor alterations can significantly affect the results, while others display a degree of inertia. The sensitivity classification of the parameters can be determined based on comprehensive sensitivity analysis methods commonly used in hydrological modeling, including local sensitivity analysis (e.g., one-at-a-time method) and global sensitivity analysis (e.g., Sobol' method and Morris method). These methods evaluate the impact of parameter variations on model output by quantifying the changes in simulation results caused by perturbations in individual parameters. Specifically, parameters are classified as sensitive if a small relative change in their values leads to a significant change (e.g., exceeding a predefined threshold in terms of NSE reduction or RMSE increase) in the model's streamflow simulation results. This classification is also supported by extensive calibration studies in various basins across China. For example, Gong et al. (2021) conducted sensitivity analysis on XAJ model parameters in small-and medium-sized catchments in South China and identified parameters such as B, WM, and EX as highly sensitive. Similarly, Lei et al. (2023) confirmed the sensitivity of parameters like KC and KG through runoff simulation studies in humid regions.

References

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

Lei, X., Cheng, L., Ye, L., Zhang, L., KIM, J. S., Qin, S., & Liu, P. (2023). Integration of the generalized complementary relationship into a lumped hydrological model for improving water balance partitioning: A case study with the Xinanjiang model. Journal of Hydrology, 621, 129569. https://doi.org/10.1016/j.jhydrol.2023.129569

6. The connection between TCN and GRU components (Section 2.4, Figure 3) is described conceptually but lacks technical details about information flow, feature dimension matching, and potential bottlenecks in the architecture. For example, how do TCN outputs' dimensions match with GRU input?

Author Response: Thank the reviewer for your detailed comments. You pointed out the technical details of the lack of information flow, feature dimension matching, and potential bottlenecks between the TCN and GRU components. This suggestion is very important and helps clarify the technical implementation logic of the model architecture. We fully accept your suggestion.

We have supplemented the technical details of the connection between TCN and GRU in Section 2.4 of the TCN-GRU model, clearly defining the feature dimension of TCN output, the dimension matching method with GRU input (via a linear projection layer), key parameter settings (e.g., tcn_features=64, gru_units=32), and bottleneck mitigation strategies. This will make the technical details of the model architecture clearer, enhance interpretability and reproducibility, and allow readers to intuitively understand the information flow mechanism between the two components.

Revision:

***p*.8, *l*.264 - *l*.282**

2.4 TCN-GRU model

Fig. 3 depicts the architecture of the TCN-GRU model, which consists of input layer, TCN layer, GRU layer, and output layer. Historical streamflow data and highly correlated influencing factors are integrated as input to the TCN layer, which excels in feature extraction, fast convergence, and robustness. The TCN layer processes input data through a stack of dilated causal convolution modules (each containing a 1D convolution with kernel size 3, dilation rate doubling per layer, and residual connections) to generate high-dimensional temporal feature maps. Specifically, for input data with shape (batch_size, time_steps, input_features), the TCN layer outputs feature maps with shape (batch_size, time_steps, tcn_features), where tcn_features is determined by the number of filters in the final TCN convolution layer (set to 64 in this study based on hyperparameter optimization).

To match the feature dimensions with the subsequent GRU layer, a linear projection (via a fully connected layer with no activation) is applied to the TCN output, transforming the feature dimension from tcn_features to gru_units. This ensures the TCN output (shape: (batch_size, time_steps, gru_units)) directly serves as the input to the GRU layer, which is designed to accept sequences with feature dimension equal to gru_units.

The GRU layer, with hidden size gru_units, processes the temporal features by updating its internal state at each time step, capturing sequential dependencies. Potential bottlenecks (e.g., information loss during dimension conversion) are mitigated by: (1) keeping tcn_features ⩾ gru_units to avoid downsampling-related information compression; (2) adding a skip connection from the TCN output to the GRU input, which retains raw TCN features alongside the projected features; and (3) using batch normalization in the TCN layer to stabilize feature distribution during propagation.

Thus, the GRU layer effectively learns dynamic variations from the TCN-extracted features, capturing temporal correlations among multiple features to enhance simulation accuracy. The output layer generates the final simulated values.

7. Hyperparameters like TCN dilation factors, GRU hidden units, and RF tree depth could be reported in the main methodology section or explicitly referenced as being provided in the supporting information.

Author Response: We appreciate the reviewer's valuable comments. You pointed out that hyperparameters such as the TCN expansion factor, GRU hidden unit, and RF tree depth should be reported in the main methods section, or that relevant content in the supplementary materials should be explicitly cited. This suggestion is reasonable and necessary, as it helps to improve the transparency and reproducibility of the research methods. We fully accept your suggestions.

We have supplemented the hyperparameter information and references to supplementary materials in **Supplementary Information 2.1 TCN-GRU model** and **2.2 Random forest (RF)**. Such modifications clarify the model parameter settings by explicitly stating key hyperparameters (e.g., TCN expansion factor, GRU hidden layer size, RF tree depth) in the methods section and providing references to supplementary materials. This approach balances readability of core information while offering comprehensive details through supplementary materials, thereby enhancing the reproducibility and rigor of the research.

Revision:

***p*.9, *l*.283 - *l*.284**

The configuration of model hyperparameters, such as the TCN dilation factors, GRU hidden units, is detailed in **Supplementary Information 2.1**.

Supplementary Information 2.1

2.2 TCN-GRU model

The architecture of the TCN-GRU model, which consists of input layer, TCN layer, GRU layer, and output layer. Historical streamflow data and highly correlated influencing factors are integrated as input to the TCN layer, which excels in feature extraction, fast convergence, and robustness. Through the TCN layer, input data is processed to produce one-dimensional sequences via dilated causal convolutions and residual connection modules. The TCN layer uses a stack of 3 dilated causal convolution layers with kernel size 3, where dilation factors are set to 1, 2, and 4 (doubling per layer) to expand the receptive field while maintaining computational efficiency. The number of filters in each TCN layer is 64, resulting in output feature maps with 64 dimensions per time step.

The GRU layer is utilized to learn and model the dynamic feature variation extracted from TCN, with the number of hidden units set to 32 to balance model complexity and performance. A linear projection layer is applied to the TCN output to match the feature dimension (64 → 32) with the GRU input requirements. Detailed hyperparameters of the TCN-GRU model, including the number of layers, dropout rate (0.2), and learning rate (0.001), are provided in Supplementary Information 5.

The GRU model demonstrates strong nonlinear mapping capabilities for time-series data, due to its inherent significant nonlinearity in streamflow time series. Thus, the GRU layer effectively captures temporal correlations among multiple features to enhance simulation accuracy.

***p*.9, *l*.298 - *l*.299**

The configuration of RF tree depth, is detailed in **Supplementary Information 2.2**.

Supplementary Information 2.2

2.2 Random forest (RF)

For the nonlinear ensemble in this study, the RF model is configured with 200 decision trees (n_estimators=200) and a maximum tree depth of 10 to prevent overfitting. The number of randomly selected features for each split (max_features) is set to "sqrt" of the input dimension. Other parameters, such as minimum split sample size = 5 and minimum leaf node sample size = 1. Tuning was tested via grid search with tree depth [5,10,15,20] and number of trees [100,200,300]. The OOB error was minimised (5.7%) at 15 tree depth + 200 trees, with no signs of overfitting. Ultimately, RF combines the simulation results of the multiple trees through averaging to derive the final regression outcome.

8. Line 379: Remove the double comma: "at the optimal location in the Taylor diagram, and the correlation"

Author Response: Thank the reviewer for their detailed comments. You pointed out that there are redundant double commas in Line 379. This suggestion is reasonable and correct, and it helps standardize the use of punctuation in the text. We fully accept this suggestion.

We have removed the extra comma from the original text: "Notably, in Fig. 8, the XAJ-TCN-GRU model's predictions for the Wuding River basin are positioned at the optimal location in the Taylor diagram, , and the correlation coefficient is 0.99." and revised it to: Notably, in Fig. 8, the XAJ-TCN-GRU model's predictions for the Wuding River basin are positioned at the optimal location in the Taylor diagram, and the correlation coefficient is 0.99.

This revision eliminates redundant punctuation, making the sentence structure more rigorous and standardized, enhancing the readability and professionalism of the text, and ensuring that readers can clearly understand the logical relationships within the sentence.

Revision:

***p*.19, *l*.541 - *l*.543**

Notably, in Fig. 8, the XAJ-TCN-GRU model's predictions for the Wuding River basin are positioned at the optimal location in the Taylor diagram, and the correlation coefficient is 0.99.

9. Section 5.2 identifies flood events as periods where "streamflow exceeded four times the standard deviation", which may not align with standard return period analysis. It would be better to use established flood frequency analysis methods.

Author Response: We appreciate the reviewer's professional comments. You pointed out that the definition of a flood event in Section 5.2, which states that "flow exceeds four standard deviations," may be inconsistent with standard recurrence period analysis, and suggested adopting a mature flood frequency analysis method. This suggestion is reasonable and critical, as it helps improve the scientific and normative nature of the flood event definition in the study. We fully accept this suggestion.

In the revision, we have replaced the original definition of flood events with a standard method based on flood frequency analysis, specifically the Peak Over Threshold (POT) method, and determine the threshold based on the hydrological characteristics of the study area.

Revision:

***p*.23, *l*.658 - *l*.663**

This study also placed a specific emphasis on the accurate simulation of extreme hydrological events. Flood events were identified using the Peak Over Threshold (POT) method, a widely accepted flood frequency analysis approach (Hosking & Wallis, 1997). For each basin, the threshold was determined as the 95th percentile of the daily streamflow series in the testing set, corresponding to a return period of approximately 2 years based on regional hydrological characteristics (as recommended by the China Hydrological Manual, 2017). Periods where daily streamflow exceeded this threshold and lasted for at least 2 consecutive days were classified as flood events. Through in-depth hydrological data analysis, a total of eight such flood events were identified across the four basins.

References

Hosking, J. R. M., & Wallis, J. R. (1997). Regional frequency analysis: An approach based on L-moments. Cambridge University Press.

China Hydrological Manual Compilation Committee. (2017). Hydrological manual (4th ed.). China Water & Power Press.

10. On line 412, the out-of-context validation excludes "wet-year" data from training but doesn't clearly define what constitutes a "wet-year" or provide quantitative criteria for this classification.

Author Response: We appreciate the reviewer's valuable comments. The reviewer pointed out that the definition of “wet-year” in Section 5.1 lacks clarity and quantitative standards. This suggestion is reasonable and will help improve the rigor of the research methodology. We fully accept this suggestion.

In the revision, the definition and quantitative criteria for “wet-year” will be added to the “out-of-context validation” section of the model robustness analysis in Section 5.1. By clearly defining the temporal scope of “wet-year” (hydrological year) and quantitative criteria (annual precipitation exceeding the 75th percentile), and citing Chinese national standards as the basis, the definition of this concept becomes clearer and more standardized, enhancing the operability of the research methodology and the comparability of results, while also enabling readers to accurately understand the logic behind the classification of extreme hydrological conditions.

Revision:

***p*.21, *l*.589 - *l*.594**

In the out-of-context validation, we excluded data from wet-year in the training sets of the four basins and used these data separately for model validation. A wet-year was defined as a hydrological year (June of the current year to May of the next year, consistent with China's hydrological year division) where the annual precipitation exceeded the 75th percentile of the long-term (2010–2023) annual precipitation series in the respective basin. This criterion aligns with the classification standard for wet years in the Technical Specifications for Hydrological Data Processing (General Administration of Quality Supervision, GB/T 50102–2014) in China. This approach aims to evaluate the model's performance under extreme hydrological conditions (i.e., wet-years) that it has not encountered before, providing a better understanding of its generalization ability and robustness.

References

General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China. (2014). Technical Specifications for Hydrological Data Processing (GB/T 50102–2014). China Planning Press.

11. The Data Availability Statement (lines 617-618) mentions that "Model source code can be obtained from its Github repository (https://github.com/zcwang1028/code.git)", but providing code through compressed archives (rar/zip files) rather than proper version-controlled repositories could benefit from following recommended reproducibility practices (Gil et al., 2016).

References

Gil, Y., David, C.H., Demir, I., Essawy, B.T., Fulweiler, R.W., Goodall, J.L., Karlstrom, L., Lee, H., Mills, H.J., Oh, J.-H., Pierce, S.A., Pope, A., Tzeng, M.W., Villamizar, S.R., Yu, X., 2016. Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. Earth Space Sci. 3, 388–415. https://doi.org/10.1002/2015ea000136

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V., 2022. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. ACM Comput Surv 55, 1–37. https://doi.org/10.1145/3514228

Author Response: We appreciate the reviewer's valuable suggestions. Your suggestion that providing code should follow recommended reproducibility practices is reasonable and important, and is critical to enhancing the transparency and reusability of research. We fully accept your suggestion and will revise the data availability statement accordingly.

In the revisions, we have retained the original GitHub repository link (https://github.com/zcwang1028/code.git) to support version control. Additionally, we have further optimized the organizational structure of the code repository and supplemented it with detailed code documentation, including environment configuration requirements, explanations of key function operations, and examples of execution steps, to ensure that other researchers can easily understand and utilize the code. Additionally, we have added the data information from this study and a stable version snapshot of the code to the repository to avoid compatibility issues caused by subsequent updates, thereby better adhering to the best practices for data and software documentation and sharing in Earth science research proposed by Gil et al. (2016). The data availability statement also specifically outlines the sources and acquisition methods of the data. The two listed references have been appropriately cited in the text to enhance the transparency and standardization of the research and data.

Revision:

***p*.33, *l*.928 - *l*.934**

Streamflow data from this study site were obtained from the Hydrological Yearbook of the People's Republic of China and provided by the Shanghai Qingyue Information Technology Service

Centre (https://data.epmap.org/page/index); meteorological data were obtained from the China Meteorological Network (https://weather.cma.cn/). To ensure complete transparency and facilitate potential replication studies, the full computational framework supporting this research—including thoroughly documented model source code and all preprocessed input datasets—has been made publicly available through a dedicated GitHub repository (https://github.com/Xunannn/XAJ-TCN-GRU-model.git).

***p*.11, *l*.352 - *l*.353**

Statistical summaries of the streamflow datasets for each basin are presented in **Table 3**, offering a more intuitive characterization of the data (Gil et al., 2016).

Reference

Gil, Y., David, C.H., Demir, I., Essawy, B.T., Fulweiler, R.W., Goodall, J.L., Karlstrom, L., Lee, H., Mills, H.J., Oh, J.-H., Pierce, S.A., Pope, A., Tzeng, M.W., Villamizar, S.R., Yu, X., 2016. Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, **3**(10), 388–415. https://doi.org/10.1002/2015ea000136

Reviewer #2

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.

Author 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.

Author 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.

Revision:

***p*.33, *l*.898 - *l*.909**

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.

Author 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 < 25% 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 S4. It indicates that the XAJ-TCN-GRU model achieves NSE values of 0.954–0.969 under extreme low-flow scenarios, representing a mere 1.7%–2.3% reduction compared to full-data training scenarios. This performance significantly outperforms both the TCN-GRU model (3.4%–3.5% reduction) and the XAJ model (8.0%–8.8% 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 (1.7%) 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 S4, the NSE of the XAJ-TCN-GRU model in extreme high-flow testing remained between 0.941 and 0.955, representing a decrease of 3.1%–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.

Revision:

***p*.20, *l*.552 - *l*.564**

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 (Wang et al., 2023; Zhao et al., 2024).

***p*.22, *l*.605 - *l*.632**

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., 2025b).

‘Dry-Year Exclusion’ Approach: Years with annual precipitation < 25% of the multi-year average were defined as ‘dry-year’. After excluding dry-year data from the training set, model training was conducted, followed by testing using the excluded dry-year data. Results are shown in Table S4. It indicates that the XAJ-TCN-GRU model achieved NSE values of 0.954–0.969 under extreme low-flow scenarios, with a performance decline of only 1.7%–2.3% compared to the full-data training scenario. This decline was lower than that observed for the TCN-GRU model (3.4%–3.5%) and the XAJ model (8.0%–8.8%), 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 (1.7%) 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 S4. 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 3.1%–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.

Supplementary Information 4.3

4.3 Model robustness analysis

Table S4 Results of model robustness experiments: 'Dry-Year Exclusion' approach and 'Extreme Event Threshold Segmentation' approach

Basin	Model	Full-data testing set	Dry-Year Exclusion		Extreme Event Threshold Segmentation		
		NSE	NSE	NSE DeclineA (%)	once-in-a-decade flow threshold（m³/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
Chu	XAJ-TCN-GRU	0.971	0.954	1.7	2780.2	0.941	3.1
	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
Jianxi	XAJ-TCN-GRU	0.984	0.962	2.2	2650.8	0.952	3.3
	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
Qingyi	XAJ-TCN-GRU	0.986	0.969	1.7	2850.3	0.955	3.1
	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.

Author 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 Author 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

Revision:

Supplementary Information 2.3

Supplementary Note 2: Methodologies

2.3 Calculation of MIC and VIF

…

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:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{S4}$$

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.

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.

***p*.13, *l*.390 - *l*.407**

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.

$$PICP = \frac{1}{n}\sum_{i=1}^{n} I(L_i \le y_i \le U_i) \tag{5}$$

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.

$$PINAW = \frac{1}{n \cdot R}\sum_{i=1}^{n} (U_i - L_i) \tag{6}$$

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.

Author 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 revisions 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

Revision:

***p*.4, *l*.120 - *l*.167**

1 Introduction

…

In recent years, hybrid modelling approaches have emerged as a significant direction for overcoming the limitations of single models, particularly through the integration of process models with deep learning frameworks. 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 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 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.

Hybrid modelling approaches currently fall into three main categories: First, physics-guided machine learning (Physics-guided machine learning), which integrates physical laws (such as mass conservation and energy balance) into deep learning architectures (e.g., loss functions or network layer design) to fuse physical mechanisms with data patterns (Willard et al., 2022). The core of this approach is "mechanism embedding", but it requires customizing the model structure for specific physical scenarios, limiting its generalizability. Second, sequential coupling methods, which use the output of physical models as input for post-processing in deep learning models (Cho and Kim, 2022; Parisouj et al., 2022). While this approach can leverage the initial physical meaning of the physical model, it carries the risk of error accumulation—structural biases in the physical model may be amplified by the deep learning model. Third, parallel integration methods, where a few studies have attempted to run the two types of models in parallel, but most adopt linear weighted fusion (Xuan et al., 2021), which struggles to capture nonlinear correlations between model outputs.

While many 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 RF to capture complex nonlinear relationships between the two models' results. Additionally, we employ the Maximum Information Coefficient (MIC) and Variance Inflation Factor (VIF) to select key hydrometeorological variables from multi-source data (e.g., precipitation, temperature, humidity), enhancing the model's robustness and interpretability.

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

Cho, K., & Kim, Y. (2022). Improving streamflow prediction in the WRF-Hydro model with LSTM networks. *Journal of Hydrology*, **605**, 127297. https://doi.org/10.1016/j.jhydrol.2021.127297

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

Parisouj, P., Mokari, E., Mohebzadeh, H., Goharnejad, H., Jun, C., Oh, J., & Bateni, S. M. (2022). Physics-informed data-driven model for predicting streamflow: A case study of the Voshmgir Basin, Iran. *Applied Sciences*, **12**(15), 7464. https://doi.org/10.3390/app12157464

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*,

55(4), 1-37. https://doi.org/10.1145/3514228

Xuan, W., Shouxiang, W., Qianyu, Z., Shaomin, W., & Liwei, F. (2021). A multi-energy load prediction model based on deep multi-task learning and ensemble approach for regional integrated energy systems. *International Journal of Electrical Power & Energy Systems*, **126**, 106583. https://doi.org/10.1016/j.ijepes.2020.106583

***p*.14, *l*.415 - *l*.422**

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

…

As indicated in Fig. 6, it is evident that the proposed hybrid model exhibits good fitting performance in all basins, indicating its strong generalization 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 (WM) of the catchment 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 generalization capability.

***p*.15, *l*.429 - *l*.441**

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 basins.

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

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

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

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

***p*.6, *l*.221**

2.1 Xin'anjiang model (XAJ)

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

***p*.8, *l*.259**

2.3 Gated recurrent unit (GRU)

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

***p*.9, *l*.286 - *l*.287**

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).

***p*.10, *l*.329**

2.6 XAJ-TCN-GRU hybrid model

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

***p*.11, *l*.342 - *l*.343**

3.1 Study areas

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

***p*.11, *l*.354 - *l*.355**

3.2 Data description

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

***p*.12, *l*.356 - *l*.357**

3.2 Data description

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).

***p*.30, *l*.812 - *l*.813**

Fig. 14 Bar plots 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.

Author 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.

Revision:

***p*.2, *l*.48**

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).

***p*.3, *l*.98 - *l*.103**

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).

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.

Author 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.

Revision:

***p*.2, *l*.64 - *l*.69**

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). 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

Contreras, P., Orellana-Alvear, J., Muñoz, P., Bendix, J., & Célleri, R. (2021). Influence of random forest hyperparameterization on short-term runoff forecasting in an andean mountain catchment. *Atmosphere*, **12**(2), 238. https://doi.org/10.3390/atmos12020238

Jehanzaib, M., Bilal Idrees, M., Kim, D., & Kim, T. W. (2021). Comprehensive evaluation of machine learning techniques for hydrological drought forecasting. *Journal of Irrigation and Drainage Engineering*, **147**(7), 04021022. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001575

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

Samantaray, S., Das, S. S., Sahoo, A., & Satapathy, D. P. (2022). Monthly runoff prediction at Baitarani river basin by support vector machine based on Salp swarm algorithm. *Ain Shams Engineering Journal*, **13**(5), 101732. https://doi.org/10.1016/j.asej.2022.101732

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.

Author 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.

Revision:

***p*.11, *l*.342 - *l*.343**

3.1 Study areas

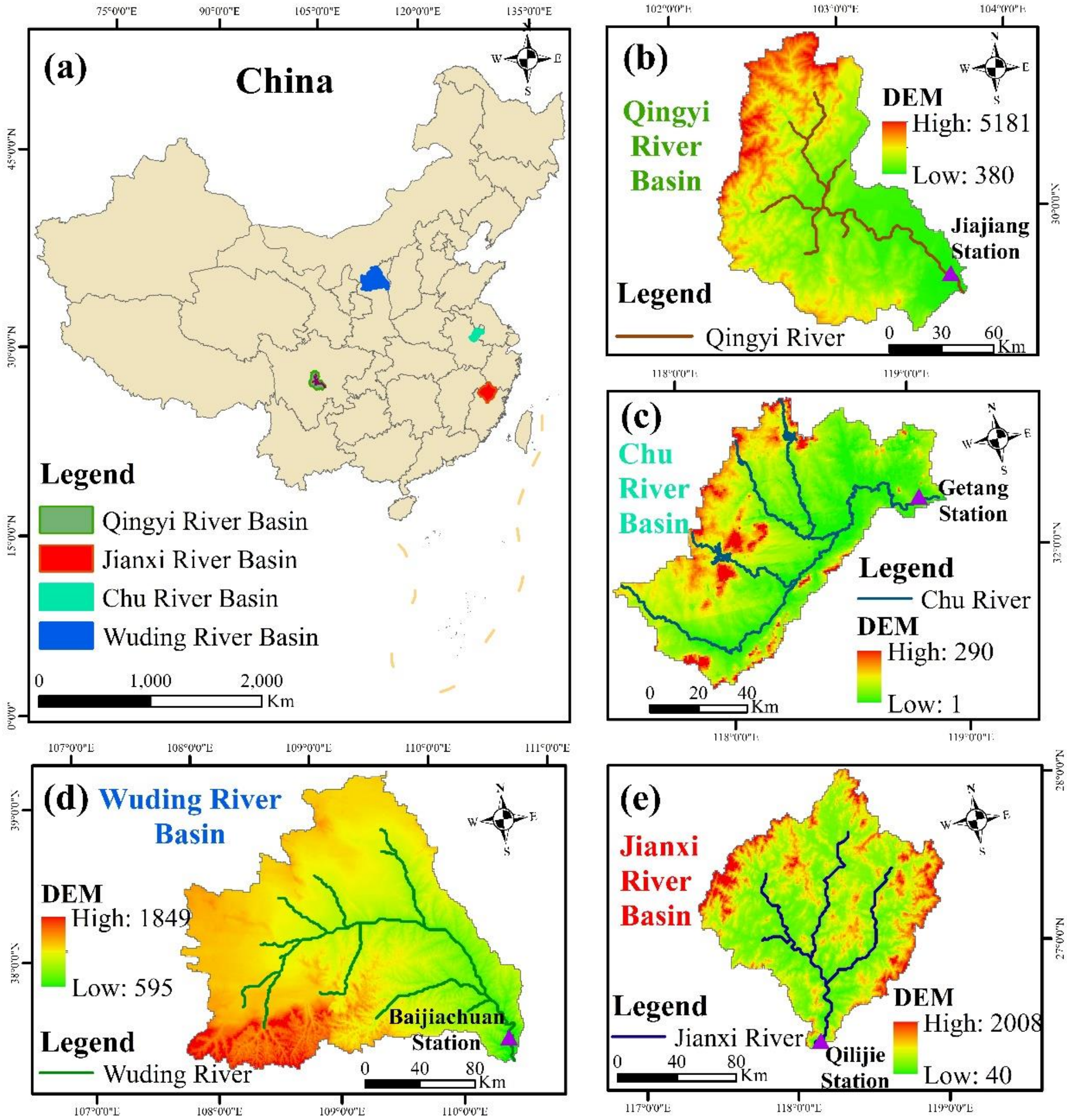

Fig. 5 Schematic diagram of geographical information and digital elevation models (DEM) for China's four river basins: Wuding River, Chu River, Jianxi River, and Qingyi River.

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.

Author 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.

Revision:

***p*.16, *l*.491 - *l*.498**

4.2 Model performance comparison

…

In addition, 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 Root Mean Square Error (RMSE), it varied between 0.32 m^3/s (Wuding River Basin) and 0.87 m^3/s (Jianxi River Basin); and for Mean Absolute Error (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.

Author 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.

We have provided the full name of the NSE metric upon its first appearance in the main text, subsequently replacing all subsequent references with the abbreviation. Furthermore, as this metric was previously introduced with its full name in the literature review section of the introduction, we have opted to use the abbreviation here to enhance the consistency and readability of the main text. We extend our gratitude once more for the reviewer's meticulous scrutiny of the paper's details.

Revision:

***p*.3, *l*.115 - *l*.119**

1 Introduction

…

Similarly, Parisouj et al. (2022) reported that in their physics-informed data-driven model, the streamflow simulation error of the physical model was amplified through sequential coupling, resulting in a further decrease in Nash-Sutcliffe Efficiency (NSE) of 0.10-0.12 in the final prediction. These results indicate that the initial errors from physical models can be magnified in sequential coupling, ultimately affecting the reliability of simulation results.

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

Author 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.

Revision:

***p*.13, *l*.390 - *l*.407**

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.

$$PICP = \frac{1}{n}\sum\nolimits_{i=1}^{n} I(L_i \le y_i \le U_i) \tag{5}$$

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.

$$PINAW = \frac{1}{n \cdot R}\sum\nolimits_{i=1}^{n} (U_i - L_i) \tag{6}$$

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.

Author 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.

Revision:

***p*.15, *l*.445 - *l*.457**

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.

Supplementary Information 4.2

4.2 Model performance comparison

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 (<120.32)	5.288	4.031	0.0468	0.911	0.925
	Medium (120.32–184.09)	6.860	4.953	0.0336	0.872	0.909
	High (>184.09)	27.644	15.544	0.0390	0.991	0.992
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