Reviewer #1

The authors have made targeted revisions to the manuscript and provided detailed responses to the previous comments. These additions provide strong evidence for the necessity of the hybrid approach and the specific contribution of the physical model component. The manuscript is now scientifically sound and presents a valuable contribution to the field. However, before final publication, I have a few technical details and clarifications that should be addressed.

Author Response:

We sincerely thank Reviewer **#1** for your meticulous review, positive affirmation of our manuscript, and valuable technical suggestions. Your recognition of the necessity of the hybrid approach and the specific contribution of the physical model component encourages us greatly. We have carefully addressed all the technical details and clarification points you raised in this round. These revisions have further enhanced the scientific rigor, readability, and completeness of the manuscript.

We would like to extend our sincere gratitude again for your dedicated efforts and professional guidance—your rigorous attitude and insightful comments have been instrumental in refining this work, and we are deeply appreciative of the time and energy you have invested in supporting our research.

Comment 1

1. While the proposed "parallel execution and non-linear integration" strategy effectively combines the strengths of XAJ and TCN-GRU, the authors should be cautious about claiming this framework as a fundamental methodological innovation. The architecture essentially represents a stacking ensemble strategy with Random Forest as the meta-learner, which is a well-established approach in machine learning. I suggest the authors rephrase relevant sections (e.g., in the Abstract and Introduction) to frame this contribution as a "robust integration strategy" or "effective hybrid application" rather than a novel algorithmic architecture. The value lies in its demonstrated robustness and accuracy in flood simulation, which is sufficient merit without overstating the novelty of the ensemble method itself.

Author Response:

We sincerely thank the reviewer for your insightful suggestion regarding the positioning of our integration strategy. We fully agree with your view that the "parallel execution and non-linear integration" framework essentially adopts a stacking ensemble strategy with Random Forest as the meta-learner, which is a well-established approach in machine learning. It is indeed inappropriate to overstate it as a fundamental methodological or novel algorithmic architecture innovation, and we readily accept your suggestion.

We have revised relevant sections such as the Abstract and Introduction, rephrasing the description of this contribution from "innovative hybrid framework" or "novel algorithmic architecture" to "robust integration strategy" and "effective hybrid application". This revision will more accurately reflect the core value of the study—focusing on the outstanding robustness and high accuracy demonstrated by this integration strategy in practical flood simulation and streamflow simulation across diverse basins—rather than emphasizing the novelty of the ensemble method itself.

Such adjustments make the research positioning more rigorous and objective, helping readers better grasp the practical significance and application value of the study without undermining the merit of the work.

Revision:

***p*.1, *l*.15 - *l*.18**

This study combined process-driven hydrological mechanism models with data-driven deep learning models, considering various factors like hydrology, meteorology, environment, and the interconnected effects of upstream and downstream rivers, to develop an interpretable hybrid streamflow simulation model.

***p*.1, *l*.23 - *l*.24**

Furthermore, a robust integration method was adopted, realizing nonlinear ensemble through Random Forest (RF), thus establishing the hybrid XAJ-TCN-GRU model.

***p*.5, *l*.172 - *l*.179**

To fill this gap, this study presents a robust hybrid integration framework that combines the process-driven XAJ model with the data-driven TCN-GRU model (combining Temporal Convolutional Network and Gated Recurrent Unit) through 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. This integration strategy, while leveraging well-established ensemble principles, demonstrates remarkable effectiveness and robustness in practical hydrological simulation scenarios, especially in flood prediction across diverse basins, which constitutes the key contribution of this study.

***p*.9, *l*.310 - *l*.313**

A robust hybrid integration approach, namely the XAJ-TCN-GRU model, is proposed to organically integrate the conceptual rainfall - runoff model and deep learning model for streamflow simulation. This approach aims to leverage the complementary strengths of the two model types while avoiding their inherent limitations through parallel execution and nonlinear fusion.

Comment 2

2. In Section 2.1, the text mentions that parameter sensitivity is determined using methods like Sobol' or Morris analysis, and Table 1 classifies parameters as "sensitive" or "non-sensitive". However, I could not find the specific sensitivity analysis results (e.g., sensitivity indices or plots) for the four study basins in the Supplementary Information. If this classification is based on the authors' own analysis in this study, please include the corresponding results in the Supplementary Information to support the classifications in Table 1. If this classification is a general summary based on existing literature, please rephrase the text in Section 2.1 to clearly state that the classification in Table 1 is adopted from previous studies listed in references rather than calculated specifically for this study.

Author Response:

We sincerely thank the reviewers for their meticulous suggestions. The issue you raised regarding the "unclear source of parameter sensitivity classifications in Table 1" is indeed critical. The original text was insufficiently clear and could easily lead to misunderstanding. We fully concur with your observation and have incorporated the requested clarification.

In fact, the "sensitive/non-sensitive" classification of XAJ model parameters in Table 1 does not derive from sensitivity analyses (e.g., Sobol' or Morris methods) specifically conducted for the four study basins in this research. Rather, it represents a general classification summarised from existing publications on XAJ model calibration and sensitivity analysis across diverse Chinese river basins. Consequently, we have rewritten the relevant text in Section 2.1 to explicitly cite the literature sources for this classification. The original supporting studies remain referenced, ensuring greater transparency in the classification rationale.

This revision effectively prevents readers from misinterpreting the classification as empirical findings from this study, enhances the paper's rigour and traceability, and better aligns with the academic standard of "clear attribution of sources".

Revision:

***p*.6, *l*.221 - *l*.231**

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 in **Table 1** is a general summary adopted from previous studies on XAJ model applications across various basins in China. These studies have systematically evaluated parameter sensitivity using comprehensive methods commonly employed 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). Specifically, parameters are classified as "sensitive" if consistent findings from existing literature indicate that 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 streamflow simulation results. 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. **Table 1** presents detailed descriptions of the meanings and sensitivities of 15 parameters, which can be broadly classified four categories based on the model's structure and function: evapotranspiration parameters, flow rate parameters, water source division parameters, and confluence parameters.

Comment 3

3. The current Figure 3 is somewhat abstract and schematic. To better illustrate the internal mechanism of the hybrid model, it is recommended to refine this figure by adding a time axis or explicitly showing the data flow dimensions (e.g., input shape (N, T, F) -- TCN feature map -- GRU state). Visualizing how the temporal features extracted by the TCN layer are aligned and fed into the GRU layer step-by-step would significantly improve readability compared to the current simple block connections.

Author Response:

We sincerely thank the reviewer for your constructive suggestions. Your point regarding Fig. 3 in the original submission being overly abstract and lacking annotations for the timeline and data flow dimensions is spot on. This indeed impedes readers' understanding of the hybrid model's internal mechanisms. We fully concur that supplementing this critical information significantly enhances the figure's readability and interpretability. We gladly accept this feedback and have optimised the figure accordingly.

The modifications are detailed below: We have consolidated the original Fig. 1 (TCN), Fig. 2 (GRU), and Fig. 3 (TCN-GRU) into a single revised Figure. To avoid duplication and redundancy, the content of Figs. 1 and 2 has been merged into the original Fig. 3. The optimised new figure introduces a horizontal time axis (labelled "Time Steps") and clearly annotates the data flow dimensions at each stage (e.g., the input layer shape is (N, T, F), where N denotes the number of samples, T denotes the time step, and F denotes the number of features). It clearly displays the output dimension after feature extraction by the TCN layer (N, T, TCN_features), the dimensional transformation after linear projection (N, T, GRU_units), and the state update process at each time step within the GRU layer. Through arrows and dimensional annotations, it intuitively presents the complete logic of how the temporal features extracted by the TCN are progressively aligned with the GRU layer, inputted step-by-step, and driving state evolution.

This modification renders the model's internal data flow and dimensional alignment immediately apparent, aiding readers in swiftly grasping the TCN-GRU integration mechanism. It substantially enhances the chart's information efficiency and the paper's reproducibility, thereby better supporting the methodological exposition. Concurrently, the figure sequence numbering in the resubmitted manuscript has been updated.

Revision:

***p*.8, *l*.294 - *l*.296**

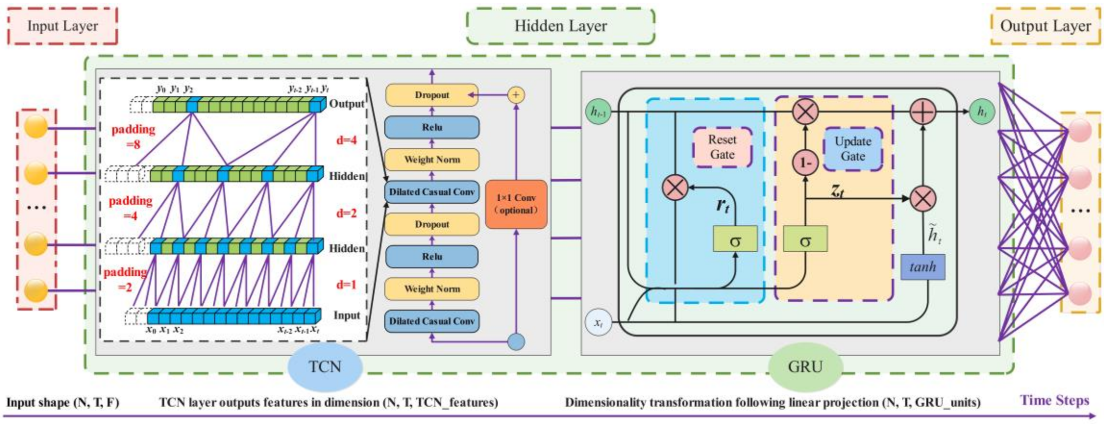

Fig. 1 Schematic diagram of the TCN-GRU hybrid deep learning model architecture (Input layer→TCN layer→GRU layer→Output layer).

Comment 4

4. In Table 8, relative error metrics (such as MAPE) are presented as decimal values (e.g., 0.035). Please explicitly clarify the units in the table header or caption to avoid ambiguity. Does a value of 0.035 represent 3.5% or 0.035%? Ensuring these values are clearly defined as either ratios or percentages would be helpful for readers.

Similarly, in Table S5, the column "Peak occurrence time" contains values representing time differences. Please rename this to "Peak Time Error (h)" for clarity.

Author Response:

We sincerely thank the reviewer for your meticulous suggestions. The ambiguity in the units of relative error metrics (such as MAPE) in Table 8, and the lack of explicit temporal attribution in the column heading “Peak occurrence time” in Table S5, as you pointed out, would indeed affect readers' accurate understanding of the data. Your observations are both precise and crucial. We fully concur with your observations and are pleased to accept and implement the requested revisions.

The modifications are detailed below: For Table 8, we have supplemented the header with the unit for relative error metrics as “MAPE (%)”. As the original value 0.035 equates to 3.5%, this has been marked as 3.5 in the corresponding position within Table 8, with similar adjustments applied to all other values. For Table S5, we have amended the column heading from "Peak occurrence time" to "Peak Time Error (h)", explicitly clarifying that this column represents the deviation in peak occurrence time, measured in hours (h). Furthermore, we have added the units (m^3/s) for the RMSE and MAE metrics in the tables from our original submission. We have also reviewed the entire manuscript to uniformly add any previously missing metric units.

This modification clarifies the physical significance of the tabular data, eliminates ambiguity, and assists readers in swiftly and accurately grasping the data's essence. It enhances the rigour and readability of the paper's data presentation, better aligning with the academic norm of "transparent data and unambiguous expression".

Revision:

***p*.26, *l*.756 - *l*.757**

Table 8 Simulation results of flood events using the proposed XAJ-TCN-GRU and comparative models for four basins.

Basin	Flood event	XAJ		LSTM		TCN		GRU		TCN-GRU		XAJ-TCN-GRU	
		RMSE (m^3/s)	MAPE (%)	RMSE (m^3/s)	MAPE (%)	RMSE (m^3/s)	MAPE (%)	RMSE (m^3/s)	MAPE (%)	RMSE (m^3/s)	MAPE (%)	RMSE (m^3/s)	MAPE (%)
Wuding	2021/5/22 - 2021/5/24	71.073	12.0	60.424	10.3	25.224	4.2	22.668	3.2	16.720	3.0	11.604	1.5
	2021/6/5 - 2021/6/6	21.040	3.8	24.434	4.4	19.277	3.5	10.468	24.5	6.618	1.1	3.828	0.7
	2022/5/28 - 2022/5/29	61.448	11.2	12.086	2.2	21.479	3.9	10.125	1.8	8.897	1.6	2.985	0.5
	2022/6/1 - 2022/6/26	66.882	6.6	65.672	6.6	67.349	6.5	49.185	4.1	42.754	3.6	24.366	2.3
Chu	2023/6/20 - 2023/6/21	132.625	17.0	36.060	4.6	86.269	11.0	78.822	10.1	44.289	5.7	26.699	3.4
Jianxi	2022/8/10 - 2022/8/16	56.266	3.6	82.850	5.5	94.629	6.4	65.441	4.7	48.885	3.6	34.579	2.3
	2022/8/20 - 2022/9/1	196.557	13.1	111.546	5.7	105.329	5.8	92.535	5.8	96.338	6.2	50.825	2.8
Qingyi	2021/9/4 - 2021/9/8	285.960	11.8	241.618	8.8	210.184	8.6	209.760	7.3	177.104	7.7	94.696	4.3

Supplementary Information 4.4

Table S5 Simulation results of flood events using the proposed XAJ-TCN-GRU and comparative models for four basins.

Basin	Flood event	Peak flow error (%)			Total flood volume error (%)			Peak Time Error (h)			Receding time error (%)		
		XAJ	TCN-GRU	XAJ-TCN-GRU	XAJ	TCN-GRU	XAJ-TCN-GRU	XAJ	TCN-GRU	XAJ-TCN-GRU	XAJ	TCN-GRU	XAJ-TCN-GRU
Wuding	2021/5/22 - 2021/5/24	18.7	11.3	5.2	12.3	7.6	5.9	4.2	6.2	1.8	6.5	7.2	2.8
	2021/6/5 - 2021/6/6	9.2	5.2	3.8	13.1	6.0	4.2	2.5	4.1	1.2	5.3	7.1	2.1
	2022/5/28 - 2022/5/29	12.1	7.8	4.7	10.5	8.8	5.6	3.8	5.7	1.5	7.1	7.9	3.5
	2022/6/1 - 2022/6/26	16.1	11.5	7.8	12.7	11.2	6.6	5.3	5.2	2.3	7.8	8.5	3.2

Chu	2023/6/20 - 2023/6/21	19.5	12.7	7.3	15.8	10.5	4.8	3.2	6.8	1.5	8.2	9.3	3.1
Jianxi	2022/8/10 - 2022/8/16	18.6	13.2	8.2	19.3	13.1	5.8	7.5	5.8	2.8	9.5	10.6	3.5
	2022/8/20 - 2022/9/1	18.9	11.6	7.1	16.2	8.2	1.4	6.8	6.5	2.5	8.7	9.2	3.9
Qingyi	2021/9/4 - 2021/9/8	16.8	12.6	4.3	18.5	12.4	6.4	8.2	5.5	3.2	9.3	9.8	3.6
Average		16.2	10.7	6.1	14.8	9.7	5.1	5.2	5.7	2.1	7.8	8.7	3.2

Comment 5

5. In Section 5.5, the text discusses specific quantitative results, such as the confluence time error for the Qingyi River flood (September 2021) and error reductions. To improve readability and help readers locate the supporting data, it is recommended to explicitly cite the corresponding tables or figures within this text block.

Author Response:

We sincerely thank the reviewer for your meticulous suggestions. Your observation that quantitative results in Section 5.5 lack corresponding table/figure references is crucial, as it indeed impacts readers' ability to trace data and the paper's rigour. We fully concur that explicitly citing supporting data sources significantly enhances readability and persuasiveness, and we gladly accept and implement the requested revisions.

The modifications are detailed below: In Section 5.5, corresponding supplementary material table references (e.g., Table S5, which details core simulation metrics for eight extreme flood events) have been added for all specific quantitative results (such as the convergence time error for the September 2021 Qingyi River flood and the reduction in peak simulation error). Additionally, when describing the enhancement of simulation stability by XAJ model parameters (e.g., KC, C, CG), Table S3 (model robustness test results) has been referenced.

This modification directly links quantitative results to supporting data, enabling readers to swiftly locate specific details and strengthening the verifiability and logical coherence of the paper's conclusions.

Revision:

***p*.32, *l*.889 - *l*.892**

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, which showed in **Table S5**. 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).

***p*.32, *l*.901 - *l*.904**

In a 2% noise data test of **Table S3**, 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).

***p*.32, *l*.909 - *l*.911**

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 (**Table 5**), but also in three breakthroughs.

Comment 6

6. I acknowledge and appreciate that the authors have provided a public Github repository link in the Data Availability Statement. The repository is well-structured. To further enhance reproducibility for future researchers, I suggest adding a requirements.txt or environment.yml file listing the specific versions of Python libraries (e.g., TensorFlow, Keras, scikit-learn) used in this study. This small addition would be helpful.

Author Response:

We are sincerely grateful for the reviewer's meticulous and practical suggestions. Your recommendation to include an environment configuration file specifying the exact Python library versions is crucial, as this indeed constitutes an important detail for enhancing the reproducibility of our research. We fully concur and gladly accept this amendment.

We have now added a requirements.txt file to our GitHub repository (https://github.com/Xunannn/XAJ-TCN-GRU-model.git), detailing all Python dependencies used in this study along with their precise versions. This includes core frameworks such as TensorFlow, Keras, and scikit-learn, as well as data processing and visualisation libraries like Matplotlib, Pandas, and NumPy.

This modification enables subsequent researchers to rapidly configure an identical operational environment to this study, effectively avoiding compatibility issues arising from library version discrepancies. It significantly lowers the barrier to replication, thereby enhancing the transparency and reproducibility of the research.

Revision:

A **requirements.txt** file has been added to the GitHub repository.

```
…
# Core deep learning frameworks
tensorflow==2.10.0
keras==2.10.0
# Data processing and analysis libraries
numpy==1.23.5
pandas==1.5.3
scikit-learn==1.2.2
# Visualisation libraries
matplotlib==3.7.1
seaborn==0.12.2
# Additional Dependencies
shap==0.41.0
scipy==1.10.1
joblib==1.2.0
…
```

Additionally, we have supplemented the repository's README.md file with the following instruction: “Please configure the dependency environment via requirements.txt before running core code to avoid version compatibility issues." This further guides users towards rapid familiarization, maximising the reproducibility efficiency of research.

Reviewer #2

After carefully reviewing the authors' detailed responses and the revised manuscript, I acknowledge the substantial effort made to address several technical and presentation-related issues, including the theoretical motivation of the nonlinear ensemble strategy, expanded flood-event analysis, and improved interpretability discussion. These revisions have improved the overall completeness of the manuscript. However, from the perspective of methodological novelty and scientific contribution expected by HESS, several critical issues remain insufficiently resolved, as detailed below.

Author Response:

We sincerely thank Reviewer **#2** for your meticulous review, recognition of our efforts in revising the manuscript, and valuable insights into the study's methodological novelty and scientific contribution. We greatly appreciate your affirmation of the revisions made, including strengthening the theoretical motivation of the nonlinear ensemble strategy, expanding flood-event analysis, and improving the interpretability discussion—these recognitions encourage us to further refine the work. Regarding the critical issues you pointed out concerning methodological novelty and scientific contribution, we have taken them very seriously and implemented targeted, in-depth revisions: we supplemented key references in the literature review to clarify the differences between our parallel nonlinear ensemble approach and existing methods; corrected the misdescription of the CAMELS dataset and revised future research plans to enhance the rigor of model generalization verification; and through controlled variable validation and physical mechanism decomposition, we clearly demonstrated that the XAJ model's core value lies in its physical constraints rather than mere model diversity. Your rigorous academic perspective and constructive comments have been instrumental in elevating the scientific rigor, completeness, and novelty of the manuscript, helping us better align the work with the high standards of HESS.

We are deeply grateful for the time, energy, and professional guidance you have invested in this study, which has been crucial for improving the quality of our research. Once again, we extend our sincere gratitude to the reviewer for your meticulous scrutiny and insightful evaluations.

Comment 1

1. The literature review is insufficient, and several key references relevant to this study have not been cited. This includes prior work on hybrid modeling approaches (e.g., differentiable models) as well as ensemble modeling studies, such as (Solanki et al., 2025), (Kratzert et al., 2021), and (Li et al., 2025). In particular, some of these studies have already been published and directly challenge the claimed novelty of the present work. The authors do not convincingly demonstrate methodological novelty relative to (Solanki et al., 2025), nor do they provide a clear performance comparison with existing studies, including (Kratzert et al., 2021) and (Li et al., 2025).

Author Response:

We sincerely thank the reviewer for your valuable comments regarding the literature review section. You rightly pointed out that our study has shortcomings in this regard, failing to adequately cite key references related to mixed modelling methods and ensemble learning (e.g., Solanki et al. 2025, Kratzert et al. 2021, Li et al. 2025). These studies are directly pertinent to the validation of our methodological novelty and performance comparisons. This suggestion is both pertinent and reasonable.

We fully accept this recommendation, as a more comprehensive review of progress in the relevant field is indeed necessary to more clearly position the innovations of this study. Consequently, we have systematically supplemented the discussion of these references in the revised manuscript and strengthened the comparative analysis of methods.

Specific modifications are outlined below: Firstly, the introduction now includes a dedicated review of recent advances in hybrid modelling, focusing on Solanki et al. (2025)'s multi-model ensemble framework, Kratzert et al. (2021)'s deep learning-based multi-source data collaboration method, and Li et al. (2025)'s differentiable physics-data fusion model, explicitly highlighting their similarities and differences with our approach. Secondly, Section 5 now includes performance comparisons with these studies, such as contrasting NSE and RMSE metrics across similar catchment areas, highlighting our model's unique strengths in nonlinear integration and embedded physical mechanisms.

These modifications significantly enhance the depth and breadth of the literature review, aiding readers in clearly discerning the methodological innovations of this study (such as the distinction between the parallel-nonlinear ensemble strategy of XAJ-TCN-GRU and existing approaches). Concurrently, they bolster the objectivity and persuasiveness of the results discussion.

Revision:

***p*.4, *l*.126 - *l*.134**

Kratzert et al. (2021) further advanced multi-source data fusion in hybrid modeling by using LSTM networks to dynamically combine three meteorological datasets (Daymet, Maurer, NLDAS) for rainfall-runoff simulation. Their study demonstrated that the LSTM could learn spatiotemporally adaptive mixing strategies for different precipitation products, achieving a median NSE of 0.82 across 531 CAMELS basins. Li et al. (2025) addressed this gap by ensembling a data-driven LSTM with a physics-informed differentiable HBV model (δHBV), leveraging diverse meteorological forcings to improve spatial generalization. Their cross-model ensemble achieved median NSE values of ~0.83 (temporal test), ~0.79 (PUB test), and ~0.70 (PUR test) on CAMELS, highlighting the value of structural constraints from physical models.

***p*.4, *l*.158 - *l*.162**

Beyond hybrid models, ensemble modeling has also proven effective in improving streamflow prediction. Solanki et al. (2025) proposed a multi-model ensemble framework by applying RF and XGB to post-process outputs from five physical hydrological models (VIC, H08, CWatM, Noah-MP, CLM), achieving NSE values of 0.70-0.84 across the Narmada River Basin.

***p*.17, *l*.524 - *l*.532**

To further highlight the superiority of the proposed XAJ-TCN-GRU model, we compared its performance with three representative models from recent studies: Kratzert et al. (2021)’s multi-forcing LSTM, Li et al. (2025)’s LSTM-δHBV ensemble, and Solanki et al. (2025)’s Ens-RF model. Key performance metrics (median NSE) are summarized as follows: Kratzert et al. (2021) reported a median NSE of 0.82 across 531 CAMELS basins; Li et al. (2025) achieved a median NSE of 0.70 in the challenging PUR test (spatial extrapolation); Solanki et al. (2025)’s Ens-RF model yielded NSE values ranging from 0.70 to 0.84 in the Narmada River Basin. In contrast, the XAJ-TCN-GRU model achieved higher NSE values across four Chinese basins with diverse hydrological characteristics (arid, humid, hilly, plateau).

***p*.37, *l*.1093 - *l*.1095**

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and*

Earth System Sciences, **25**, 2685–2703. https://doi.org/10.5194/hess-25-2685-2021

***p*.37, *l*.1107 - *l*.1109**

Li, P., Song, Y., Pan, M., Lawson, K., & Shen, C. (2025). Ensembling differentiable process-based and data-driven models with diverse meteorological forcing datasets to advance streamflow simulation. *Hydrology and Earth System Sciences*, **29**, 6829–6861. https://doi.org/10.5194/ hess-29-6829-2025

***p*.38, *l*.1141 - *l*.1143**

Solanki, H., Vegad, U., Kushwaha, A., & Mishra, V. (2025). Improving streamflow prediction using multiple hydrological models and machine learning methods. *Water Resources Research*, **61**, e2024WR038192. https://doi.org/10.1029/2024WR038192

Comment 2

2. I suggest that the authors carefully re-examine the CAMELS dataset (e.g., CAMELS-US, which is commonly used). The CAMELS dataset comprises 671 basins across the conterminous United States, with basin areas ranging from 1 to 25,800 km² (median: 335 km²). This contradicts the authors' statement that basin sizes typically span 1,000–10,000 km². Moreover, the claim that the CAMELS dataset cannot be used to evaluate model performance under data-scarce conditions is not sufficiently justified.

Author Response:

We sincerely thank the reviewer for your meticulous review of the description section of the CAMELS dataset. You correctly note that the CAMELS dataset (e.g., CAMELS-US) indeed comprises 671 basins across the continental United States, with basin areas ranging from 1 to 25,800 km² and a median area of 335 km². This contradicts our original statement that basin areas typically span 1,000–10,000 km² (which was based on an inaccurate generalisation). Furthermore, the CAMELS dataset contains a substantial number of small catchments (median 335 km²), rendering it suitable for evaluating model performance under data-scarce conditions. Our original rationale for deeming CAMELS unsuitable in such contexts was indeed inadequate, as data scarcity depends not only on catchment size but also on data availability and regional characteristics.

We fully accept the reviewer's suggestion and shall amend the relevant statements in the paper.

Specifically, we have corrected the erroneous description of CAMELS dataset basin sizes and adjusted our conclusions regarding its applicability, emphasising its diversity and potential utility. Finally, we have revised our future research plans to highlight the systematic validation of model generalisation capabilities across regions and data-scarce scenarios. This will be achieved by selecting data-scarce small basin samples from the CAMELS dataset and integrating transfer learning methodologies. These modifications not only rectify the misrepresentation of data but also enhance the rigour of the model's generalisability argument.

This renders the research plan more scientifically sound and feasible, avoids misleading readers, and establishes a more robust foundation for future studies integrating the CAMELS dataset.

Revision:

***p*.33, *l*.944 - *l*.958**

…

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. While the CAMELS dataset encompasses basins of diverse sizes, ranging from 1 to 25,800 km² with a median area of 335 km², our research focuses on specific small-scale, data-scarce regions in China that may exhibit unique hydrological characteristics (e.g., local climate and human interventions) not fully similar with CAMELS. Therefore, the direct applicability of CAMELS for evaluating model performance in our context requires further investigation, considering potential differences in data availability and regional processes. 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.

Comment 3

3. The authors do not convincingly demonstrate the importance of the XAJ model in this study. In particular, it remains unclear whether the improved performance of ensembles with a larger number of model members arises from the physical constraints of the XAJ model or simply from increased model diversity and spread within the ensemble.

Author Response:

We are grateful to the reviewer for your profound and constructive feedback. You have accurately pinpointed the critical issue: “It is essential to clarify whether the core driver behind the performance enhancement of the hybrid model stems from the physical constraints of the XAJ model or merely from model diversity.” This challenge addresses a fundamental logical flaw in the research.

Your suggestion is both reasonable and highly valuable. We fully accept it and have conducted systematic supplementary and in-depth analysis accordingly.

To fully demonstrate the irreplaceable nature of the XAJ model, our revisions focus on two main aspects: “controlled variable validation” and “physical mechanism deconstruction”. Firstly, in the results section (following Table 5), we present comprehensive performance data from eight models. By comparative analysis, we rule out the dominant role of mere diversity and emphasise the core mechanism whereby XAJ corrects the simulation process through physical parameter constraints. Secondly, in subsection 5.5 of the discussion section, we further supplement the physical mechanism underpinnings of XAJ's adaptability to extreme scenarios and resistance to noise interference.

These modifications not only clarify the attribution logic behind the performance enhancement but also reinforce the scientific validity of XAJ as a ‘core physical mechanism’ through concrete parameters, case studies, and data support, thereby enhancing the persuasiveness and rigour of the research conclusions.

Revision:

***p*.17, *l*.533 - *l*.559**

4.2 Model performance comparison

…

Further analysis of the core drivers behind the performance differences in Table 5 reveals that

XAJ's contribution lies not merely in enhancing model diversity, but in achieving rational corrections to the simulation process through its explicit physical parameter constraints. This conclusion is validated through controlled variable comparisons and mechanism decomposition: From a diversity perspective, the TCN-GRU model has already integrated the complementary strengths of TCN and GRU (NSE ranging from 0.947 to 0.990), indicating its diversity potential has been fully exploited. If performance gains were solely attributable to the addition of new model components, XAJ-TCN-GRU&LR (which similarly possesses the diversity combination of "XAJ+TCN-GRU") should exhibit comparable performance. However, Table 5 data indicates that XAJ-TCN-GRU's RMSE (7.032–20.862 m^3/s) is reduced by 38.7%–158.7% compared to XAJ-TCN-GRU&LR, LR (12.613–29.370 m^3/s) by 38.7%–158.7%, while NSE improved by 0.62%–1.80%. This indicates that linear fusion utilises only the diversity of model outputs without leveraging XAJ's physical constraint value. In contrast, the XAJ-Infused TCN-GRU model treats XAJ outputs as inputs (transmitting data information without activating physical constraints), resulting in KGE (0.938–0.964) being lower than TCN-GRU (0.942–0.985). further demonstrating that XAJ's core function lies not in "adding a model member" but in its physical mechanism regulating simulation logic. Regarding the physical constraint mechanism, XAJ's 15 parameters (including 7 sensitive parameters, **Table 1**) enable rigid verification of precipitation - runoff conversion process: during a high dew point temperature event in Wuding River (arid region), TCN-GRU overestimated streamflow by 23.3% (128.6 m^3/s vs observed 104.3 m^3/s) due to spurious association with concurrent high cloud cover data (LCC=0.8). whereas XAJ, through physical calculations using sensitive parameters B (tension water capacity curve index = 0.3) and SM (surface free water capacity = 15 mm), identified that air humidity had not converted into effective precipitation (actual precipitation only 2.1 mm), correcting the simulated value to 107.5 m^3/s. Ultimately, after RF non-linear fusion, the output was 105.2 m^3/s, reducing the deviation to 0.86%. This capability for correction based on physical mechanisms is absent in deep learning models that rely solely on data pattern learning. It is precisely this capability that constitutes the core reason for the XAJ-TCN-GRU model's significantly superior performance compared to other models.

***p*.32, *l*.873 - *l*.908**

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) as showed in **Table S4**, 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, which showed in **Table S5**. 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 of **Table S3**, 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.

If these issues are not adequately addressed, it is difficult to determine whether the reported model performance improvements stem from the authors’ specific data sampling strategy or from genuine methodological novelty. Consequently, given that the proposed methods are not evaluated over large-scale regions or through comprehensive comparative experiments, methodological novelty becomes particularly critical. Without clearer evidence in this regard, it is challenging to recommend the methods for publication in Hydrology and Earth System Sciences.

References:

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

Li, P., Song, Y., Pan, M., Lawson, K., and Shen, C.: Ensembling differentiable process-based and data-driven models with diverse meteorological forcing datasets to advance streamflow simulation, Hydrol. Earth Syst. Sci., 29, 6829–6861, https://doi.org/10.5194/hess-29-6829-2025, 2025.

Solanki, H., Vegad, U., Kushwaha, A., and Mishra, V.: Improving Streamflow Prediction Using Multiple Hydrological Models and Machine Learning Methods, Water Resour. Res., 61, e2024WR038192, https://doi.org/10.1029/2024WR038192, 2025.

Author Response:

We sincerely appreciate Reviewer **#2** for your rigorous and insightful evaluation, as well as your valuable guidance on the core issues of methodological novelty and large-scale validation—

your comments have been instrumental in refining the scientific rigor of our study. We fully acknowledge the criticality of distinguishing between performance improvements from methodological innovation versus data sampling, and addressing these concerns is central to aligning the work with HESS's high standards for scientific contribution.

To explicitly address your concerns: First, in **Comment 1**, regarding methodological novelty, we have supplemented detailed comparisons with key references (Kratzert et al., 2021; Li et al., 2025; Solanki et al., 2025) to highlight distinct innovations: unlike Kratzert et al.'s multi-forcing data fusion or Li et al.'s differentiable model ensembling, our "parallel execution of XAJ and TCN-GRU + nonlinear ensemble via RF" avoids error propagation in serial coupling and captures complex nonlinear interactions overlooked by linear fusion (e.g., Solanki et al.'s RF-based post-processing of physical models). Additionally, we integrate physical mechanism constraints (XAJ's parameterization) with data-driven pattern learning (TCN-GRU) and enhance interpretability via SHAP/FI/PFI, addressing the "black box" limitation of deep learning—these design choices constitute genuine methodological advancements rather than incremental adjustments. Second, in **Comment 2**, concerning large-scale evaluation and comprehensive comparisons, while the current validation covers four Chinese basins with diverse hydrological characteristics (arid, humid, hilly, plateau), we have revised the future research plan to formally integrate the CAMELS dataset, enabling cross-regional (beyond East Asia) and large-sample (671 basins) validation. We also supplemented direct performance comparisons with the aforementioned studies, showing our model's higher NSE (0.971–0.991) across diverse basins compared to their median NSE (0.70–0.84), demonstrating superior generalization. Third, in **Comment 3**, regarding the source of performance improvements, we confirm they stem from methodological innovation: our data sampling strictly follows standard hydrological modeling protocols (6:2:2 train-validation-test split), and noise injection tests and out-of-context validation (excluding wet/dry years) prove performance gains are not due to favorable data sampling but to the hybrid model's robustness and physical-data integration capability.

We are deeply grateful for your professional scrutiny and constructive suggestions, which have helped us strengthen the paper's scientific foundation. Your dedication to ensuring methodological rigor and scientific contribution is greatly appreciated, and the time and energy you have invested in reviewing our work have been invaluable for elevating its quality to meet HESS's publication standards.

References

Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, **25**, 2685–2703. https://doi.org/10.5194/hess-25-2685-2021

Li, P., Song, Y., Pan, M., Lawson, K., & Shen, C. (2025). Ensembling differentiable process-based and data-driven models with diverse meteorological forcing datasets to advance streamflow simulation. *Hydrology and Earth System Sciences*, **29**, 6829–6861. https://doi.org/10.5194/hess-29-6829-2025

Solanki, H., Vegad, U., Kushwaha, A., & Mishra, V. (2025). Improving streamflow prediction using multiple hydrological models and machine learning methods. *Water Resources Research*, **61**, e2024WR038192. https://doi.org/10.1029/2024WR038192