RC1:

1.Comment: Abstract

It is recommended to avoid using unexplained acronyms, as this may hinder comprehension for the reader. Including a brief environmental and geological context of the studied basin would help justify the choice of the forecast horizon t + 6. This parameter is highly dependent on the characteristics of the river considered and may be excessive or not meaningful in other fluvial contexts. Consequently, the results cannot be generalized without appropriate caution.

Response: Thank you for your insightful comment. All acronyms have been explicitly defined upon their first appearance in the revised manuscript, including LSTM, FTA-LSTM, FTMA-LSTM, PHY-FTMA-LSTM, NSE, KGE and R^2.

In the revised manuscript, Page 1, Line 11-12:

To address these limitations, this study enhances the traditional **long short-term memory neural network (LSTM)** by introducing a **physics-guided feature-time-based multi-head attention mechanism LSTM(PHY-FTMA-LSTM)**.

In the revised manuscript, Page 1, Line 19-22:

The results demonstrate that the PHY-FTMA-LSTM in most cases outperforms the original LSTM, the **feature-time-based attention mechanism LSTM (FTA-LSTM)**, and the **feature-time-based multi-head attention mechanism LSTM (FTMA-LSTM**). For a lead time of t+1, the model achieves a **Nash-Sutcliffe efficiency coefficient (NSE)** of 0.988, with **Kling-Gupta efficiency (KGE)** and **coefficient of determination (R^2)** of 0.984 and 0.988.

The selection of the forecast period was determined through a comprehensive evaluation of multiple factors, including basin-specific geological and environmental characteristics (added in **Section 3.1 Line 295-304**). Statistical analysis indicated that the flood concentration time in the study basin typically ranges between 6-12 hours. Meanwhile, we referred to the 6-hour forecasting horizon following Wang et al. (2023), whose methodology demonstrated successful water-level forecasting in the Han River Basin (covering>30,000 km^2). Furthermore, extending the prediction

horizon was constrained by the inherent black-box nature of deep learning models, which exhibited significant performance degradation over longer periods.

In the revised manuscript, Page 12, Line 295-304:

Based on geological conditions and geomorphological features, the area can be divided into two dominant landform types: plateau and mountainous terrain. The plateau dominates the northern part of the basin, with elevations ranging from 1400 to 1600 meters and a gentle channel gradient averaging approximately 0.5‰. The remaining area comprises mountainous terrain, exhibiting complex topography shaped by prolonged denudation and erosion. This zone features steep mountains, densely distributed hills, and interspersed basins, with slope angles varying between 20° and 40°. In certain areas, rivers demonstrate intense downward cutting action, resulting in significantly steeper channel gradients — typically 2–6‰, while some medium and small tributaries exceed 20‰. Notably, flood wave propagation velocities reach 2.0-3.5 m/s due to these topographic conditions.

Manuscript Reference:

Wang, Y., Huang, Y., Xiao, M., Zhou, S., Xiong, B., Jin, Z., 2023. Medium-long-term prediction of water level based on an improved spatio-temporal attention mechanism for long short-term memory networks. J. Hydrol. 618(129163) .

2.Comment: Study area and data

It is unclear why the experiment was conducted in this particular watershed. What are its characteristics? Why is it relevant? How does it differ from others? Will the results obtained be valid only for this site, or are they generalizable? This should be clarified.

Response: Thank you for your insightful comment. The Luan River Basin is a large watershed (44880 km²) spanning Hebei, Inner Mongolia, and Liaoning provinces, holding critical geopolitical and economic significance. It serves as a vital ecological conservation zone and water source for the Beijing-Tianjin-Hebei region. However, as a seasonal river, it alternates between rapid flood peaks during rainy seasons and frequent dry-season flow interruptions. Channel encroachment and increased agricultural/industrial water withdrawals have exacerbated downstream flow breaks,

diminishing flood conveyance capacity and heightening disaster risks. Its complexity and representativeness establish it as a paradigm for flood forecasting research in similar basins.

Our modeling methods are universally applicable, while the parameter configurations exhibit distinctiveness that necessitates calibration based on local catchment characteristics for implementation in alternative basins.

3.Comment: Lines 297–306: The data sampling frequency is not specified, even though it is a fundamental parameter.

Response: Thank you for your insightful comment. The sampling frequency in our study complies with China's National Hydrological Data Compilation Standards, which require dynamic rather than fixed-interval sampling. For flood hydrographs, additional 2-3 measurements are taken before the rising limb and after recession stabilization to facilitate baseflow separation, with critical points captured during rising/falling limb transitions and peak periods (minimum 3-5 measurements around peak discharge). Precipitation monitoring also adopts intensified logging frequency during heavy rainfall events to ensure data accuracy under extreme conditions. This adaptive sampling protocol ensures comprehensive hydrological process documentation while meeting technical requirements for flood forecasting analysis. So the data were processed as 1h time step according to Line 308-313.

4.Comment: Figure 2: it is recommended to change the colors, as the triangles and the star are not clearly visible.

Response: Thank you for your insightful comment.We have made the hydrological and rainfall stations clearer in the revised manuscript by changing the colors and sizes.

In the revised manuscript, Page 13, Fig.2 :

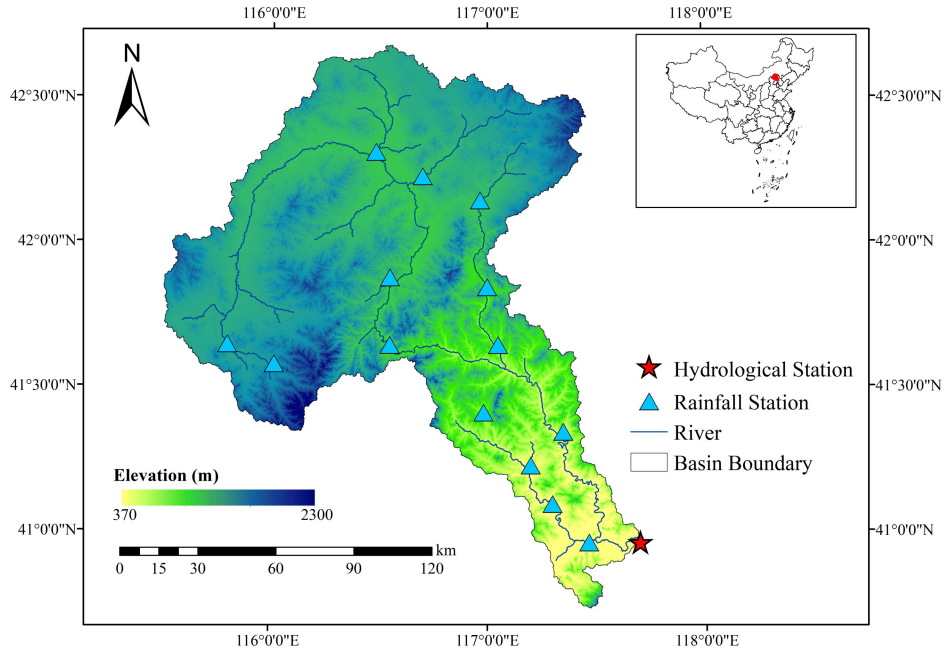


Fig.2. Geographical location of the study area and hydrological and rainfall stations.

5.Comment: The dataset split into training and validation sets appears to be the main critical issue of the study. What rule was followed? Currently, the most accepted strategy is to divide the dataset into three parts (training, validation, and test), using the validation set during batch steps. Why wasn't this approach followed? Is it due to the limited number of available cases? An explanation is required.

What happens if the events that compose the three subsets are changed? Does the predictive performance of the models vary? Using techniques like cross-validation or bootstrapping would allow for the analysis of error distributions. How stable is a model trained multiple times on the same initial dataset? Answering these questions would strengthen the scientific approach of the paper, moving it beyond a simple application. The results presented seem fragile as they might depend on the initial, arbitrary assignment of events to the training, validation, and test phases.

Response: Thank you for highlighting datasets concerns. Our data partitioning strategy was rigorously based on two primary considerations: First, with only 30 historical floods, and most of them had been short-lived, resulting in a limited sample

size, dividing the three datasets would leave ≤ 6 events to test, which was insufficient to capture spatiotemporal heterogeneity. We will also add more floods in the future if they become available. Second, this approach followed established precedents by researchers including Lv et al., Read et al., Xie et al., and Jiang et al., who used dual dataset partitioning and not only for flood forecasting.

As can be seen from Table 1, our dataset partitioning had taken into consideration key flood characteristics including temporal occurrence, peak discharge, and flood duration. This intentional design ensured that both datasets comprehensively covered a wide spectrum of flood features, thereby enabling thorough training and rigorous testing of the model. Furthermore, we had experimented with exchanging some floods within the training and validation sets. While preserving the diversity profiles of flood characteristics in both datasets, the model prediction performance changed little .

Importantly, we implemented a sliding window mechanism with a 12-timestep window length and 1-timestep stride for sample construction. This configuration ensured:

(a) Continuous temporal coverage by advancing the window progressively at each computational step.

(b) Maximized data utilization through 92% overlap between consecutive windows.

(c) Effective capture of hydrological process evolution patterns characteristic of flood events.

While cross-validation wasn't explicitly used, sliding windows inherently achieved dynamic CV (used by Gao et al., Kao et al. And Ding et al. for flood forecasting). Each timestep participated in multiple windows, akin to data recycling in CV.

Manuscript Reference:

Lv, N., Liang, X., Chen, C., Zhou, Y., Li, J., Wei, H., Wang, H., 2020. A long

Short-Term memory cyclic model with mutual information for hydrology forecasting: A Case study in the xixian basin. Adv. Water Resour.

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., Karpatne, A., Hansen, G. J. A., Hanson, P. C., Watkins, W., Steinbach, M., Kumar, V., 2019. Process-Guided Deep Learning Predictions of Lake Water Temperature. Water Resour. Res. 55(11), 9173-9190.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C., 2021. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. J. Hydrol. 603, 127043.

Jiang, S., Zheng, Y., Solomatine, D., 2020. Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning. Geophys. Res. Lett. 47(e2020GL08822913).

Gao et al., 2020. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. J. Hydrol., 589 (2020), Article 125188.

Kao et al., 2020. Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting.J. Hydrol., 583 (2020), Article 124631.

Ding, Y., Zhu, Y., Feng, J., Zhang, P., Cheng, Z., 2020. Interpretable spatio-temporal attention LSTM model for flood forecasting. Neurocomputing. 403, 348-359.

6.Comment: Are 30 events sufficient to train deep learning models? The size of the original dataset and the derived datasets is not clear. I suggest conducting a distributional analysis of the events. If the analysis focuses on a limited number of cases, they should be hydrologically analyzed and shown to be statistically representative of the hydrology of the basin under study.

Response: Thank you for highlighting datasets concerns. The hydrological records for the Luan River Basin are inherently limited, with available data spanning discontinuous periods (1964-1989, 1991, and 2006-2017), amounting to 39 years of

flood records. When selecting specific floods, we had checked the rainfall runoff data of each flood in order to ensure the reliability and representativeness of the hydrological data, and finally selected 30 floods under the premise of guaranteeing the inclusion of three types of flood magnitudes, namely large, medium and small, and covering the single-peak and multi-peak flooding processes. In addition, the sliding window mechanism (12-timestep window, stride=1) generated 4025 temporally correlated training samples from the 30 events, effectively. We believe that the number of flood events and samples could basically reflect the watershed situation and support the model training.

7.Comment: Line 323: Indicate the version of the TensorFlow library used.

Response: Thank you for your insightful comment. We used TensorFlow 2.9.1, which we have added in the revised manuscript.

In the revised manuscript, Page 15, Line 344:

.....are constructed using the Keras library in TensorFlow 2.9.1.

8.Comment: Line 327: Provide a citation for the activation functions employed.

Response: Thank you for your insightful comment. We have added reference to the activation functions used.

In the revised manuscript, Page 35, Line 785-786:

Nair, V., Hinton, G. E., 2010. Rectified linear units improve restricted boltzmann machines vinod nair. Omnipress.

9.Comment: Line 330: it is unclear how overfitting is being mitigated by early stopping. It must be demonstrated that the models are not affected by overfitting. Furthermore, splitting the dataset into three sets is a fundamental first step to prevent both overfitting and underfitting.

Response: Thank you for your insightful comment. We employed early stopping to monitor changes in the loss function (Mean Squared Error, MSE) and terminated

process if the loss showed no improvement for 20 consecutive epochs. We have also added this to the revised manuscript.

Additionally, although not mentioned in the article, during each run we visualized the loss curve to observe its trend, thereby assessing model convergence and potential overfitting. Overfitting typically manifests as strong performance on the training set but poor generalization on the validation set. However, as demonstrated by the metrics we provided, the model exhibited good generalization capability on the validation set. Regarding the dataset, as previously explained, the limited number of flood events constrained the dataset splitting.

In the revised manuscript, Page 16, Line 352-354:

To avoid overfitting, all models employ early stopping based on the mean squared error (MSE) loss function, with a maximum iteration limit of 200 epochs. The training process automatically terminates if no improvement in loss is observed for 20 consecutive epochs.

10.Comment: Results

Tables 3 and 4: It is advisable to replace the tables (which can be included as supplementary material to ensure transparency of the raw data) with plots showing the metrics as a function of lead time for each model. This would help reveal potential trends and the presence or absence of overfitting. Additionally, the reported results may lack statistical validity and could be coincidental. It is necessary to repeat the training procedures, as mentioned above, to assess the robustness of the outcomes.

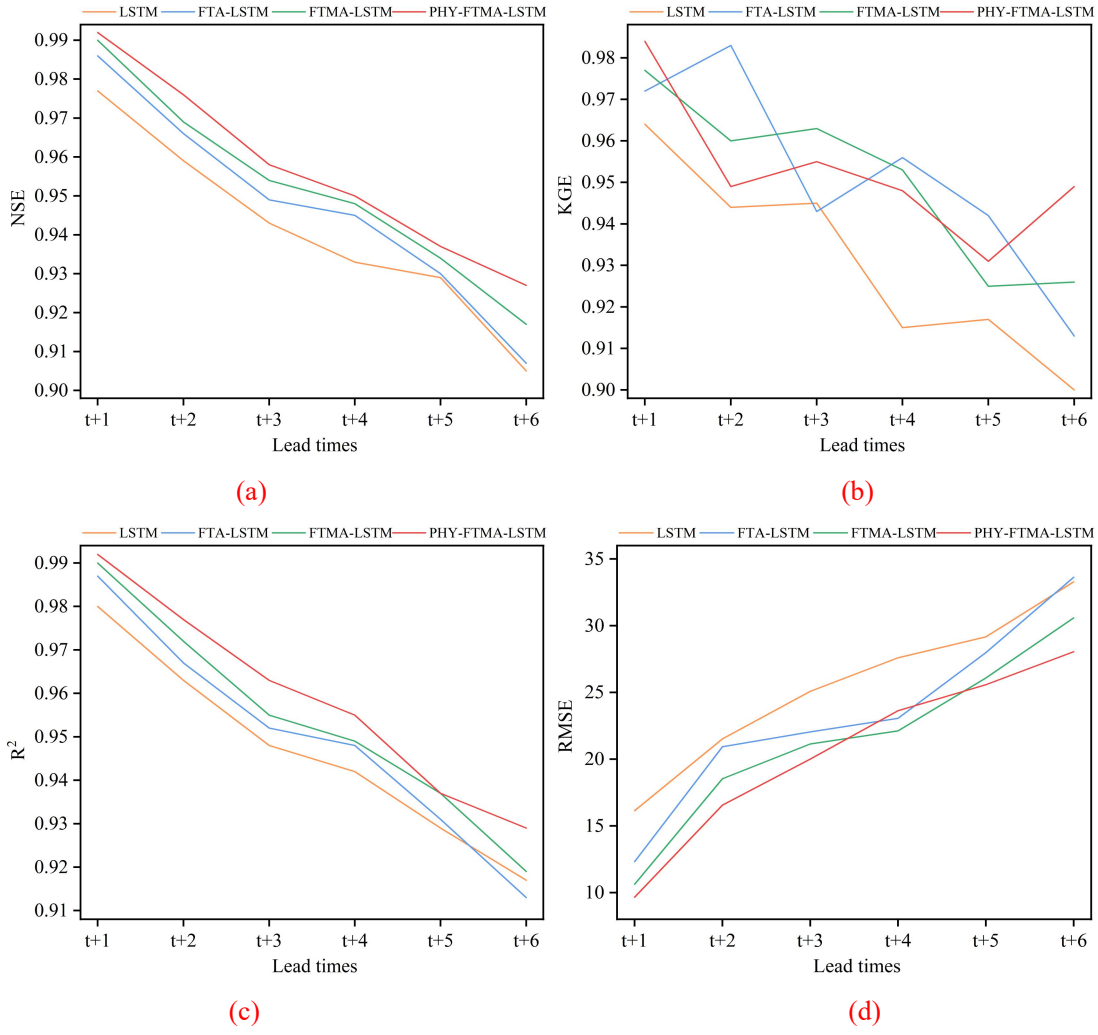
What if the error metrics were computed only for data exceeding a certain threshold (statistical or physical)? Focusing on peak flood events, would the metrics change? Would more patterns emerge?

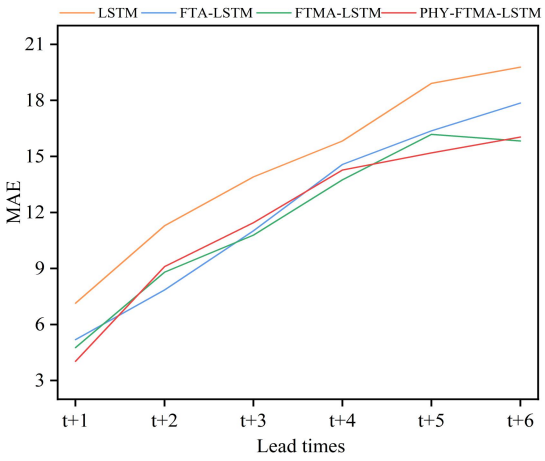
Response: Thank you for your insightful comment. We have replaced the tables with plots showing the metrics as a function of lead time for each model (**Fig.4. and Fig.5.**). As detailed in Line 362-364, all four models were repeated for five runs at

each lead time to assess stability. It is confirmed that there was little difference in performance between runs, and the best performing implementation was selected for final analysis to ensure that it would be ready for use in the event of a flood forecast.

In addition, we initially employed boxplots of peak discharge relative errors but peer reviewers noted both the absence of significant inter-model differences and insufficient representation of process dynamics, such as rising/falling limb error. Thus, we transitioned to observed vs. predicted scatterplots (Fig. 6), which could reveal full-process error patterns, identify extreme-event outliers (e.g., 19740723 flood), and visualize model-specific strengths.

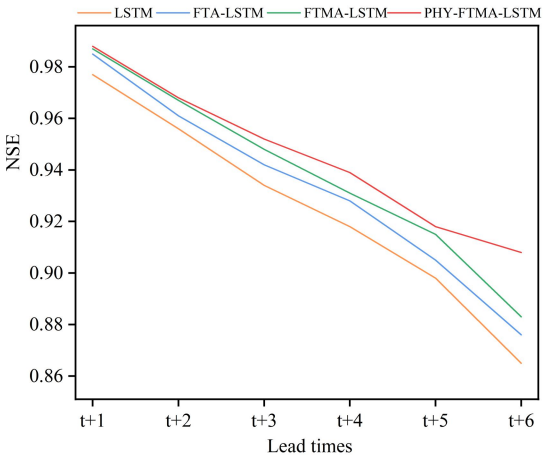
In the revised manuscript, Page 20-21, Line 460-475:



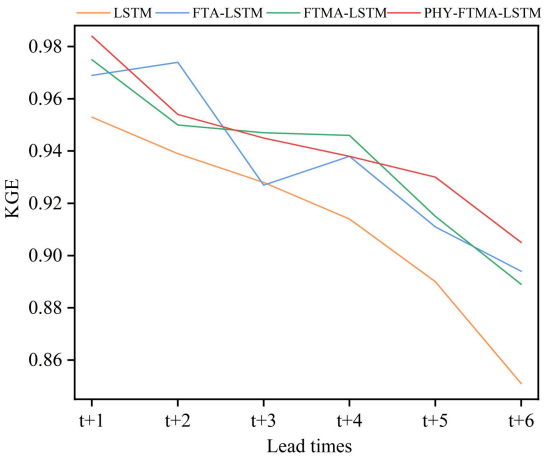


(e)

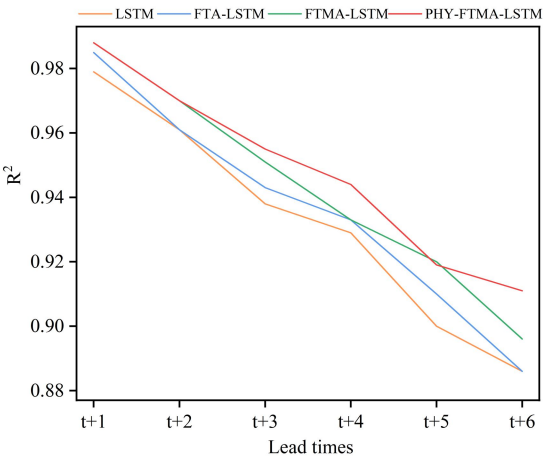
Fig.4. Performance of the four models for flood forecasting at different lead times for training (a) NSE, (b) KGE, (c) R2, (d) RMSE and (e) MAE.



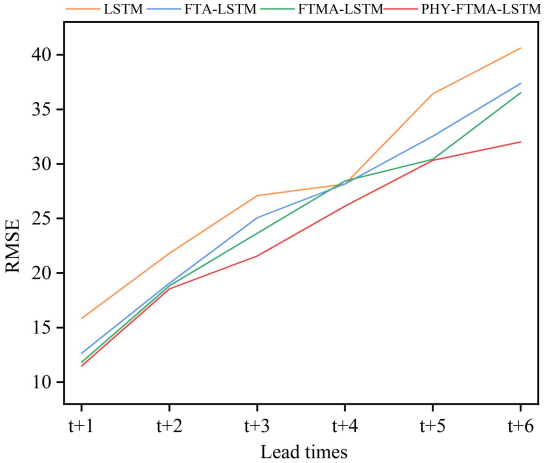
(a)



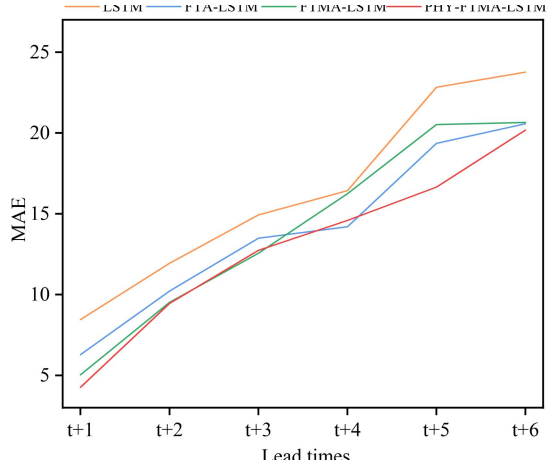
(b)



(c)



(d)



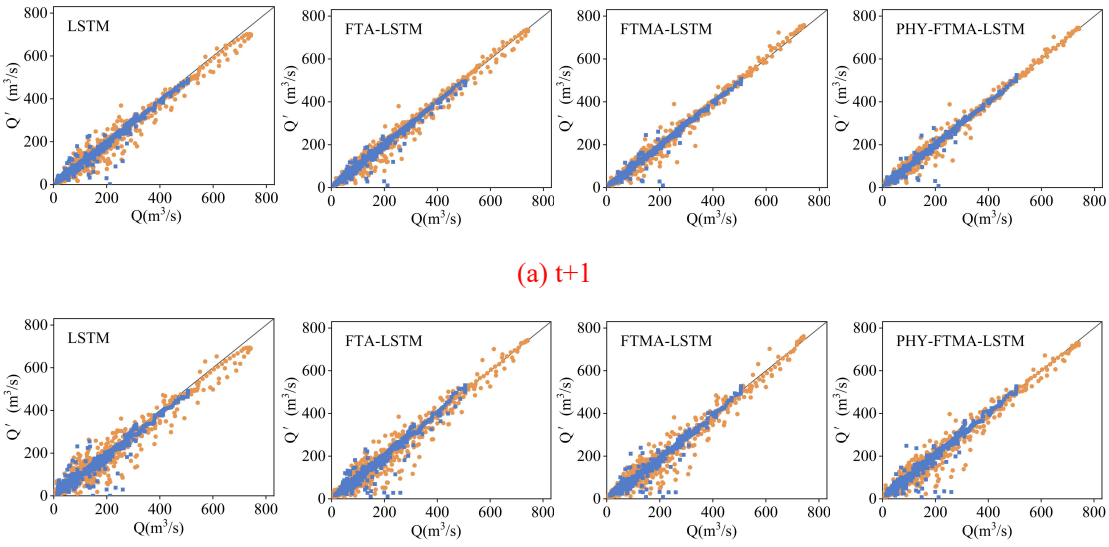
(e)

Fig.5. Performance of the four models for flood forecasting at different lead times for validation (a) NSE, (b) KGE, (c) R2, (d) RMSE and (e) MAE.

11.Comment: Figure 4: The axis labels are not legible.

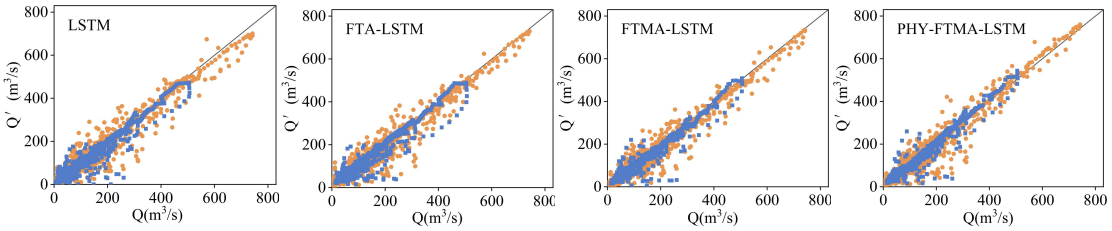
Response: Thank you for your insightful comment.We have revised the font size to ensure that the labeling in Figure 6 (formerly Figure 4) are clear and appropriately sized.

In the revised manuscript, Page 23, Line 495-506:

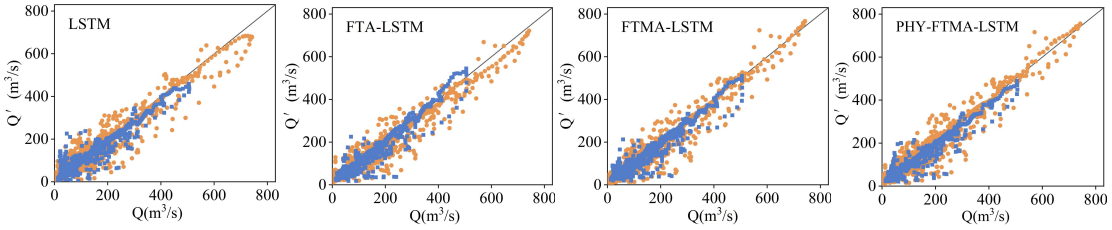


(a) t+1

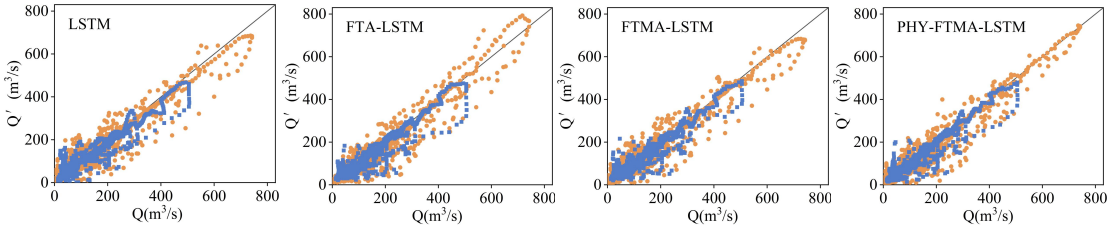
(b) t+2



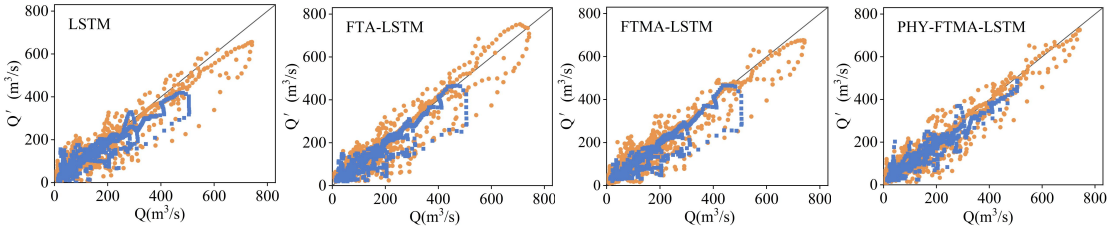
(c) t+3



(d) t+4



(e) t+5



(f) t+6

Fig.6. Scatter plots of observed and predicted discharges in the training and validation periods, in which yellow represents the training period and blue represents the validation period.

12.Comment: This observation applies to all time horizons, but is particularly evident for t + 5 and t + 6: for observed discharges above approximately 300 m³/s, an anomalous behavior appears in the scatterplot points, forming a curve. In my experience, these points likely correspond to a specific event that the model fails to simulate correctly, tending to underestimate the flows. Suppose this hypothesis is confirmed by the authors. In that case, it should be discussed, as it would reveal an interesting phenomenon: the model is unable to overestimate flow in advance and

instead tends to underestimate it as lead time increases.

These models seem to suffer from a common limitation: the inability to anticipate runoff before the onset of precipitation. This limitation may be understandable given the lack of meteorological forecast input to the model. Nonetheless, this observation opens up interesting research avenues that the authors are encouraged to explore in the discussion and conclusions.

Finally, if the hypothesis that those outlier points belong to a single event holds true, the most significant errors in predicting large events should be analyzed in detail. All these aspects could serve as input for a revision of the discussion and conclusions, enhancing the scientific impact of the paper, which in its current form lacks significant novelty.

Response: Thank you for your insightful analysis of the systematic flow underestimation at high discharges (>300 m^3/s). We confirm and deeply appreciate your hypothesis. The outliers indeed cluster within the 19740723 flood event (validation set peak). Our analysis have been added to the revised manuscript.
Furthermore, we acknowledge this constraint in our current modeling framework - the absence of meteorological forecast inputs restricts runoff anticipation capability prior to precipitation events. In both the Discussion and Conclusion sections, we have added content highlighting the issues of current research lacking weather forecasting inputs and structural constraints in models.
All of the constructive critiques have profoundly shaped our research trajectory, and we thank you for elevating the practical relevance of this work.

Discussion section

In the revised manuscript, Page 30, Line 644-651:

Notably, across all forecast periods—particularly at t+5 and t+6—scatterplot points

(Fig.6.) exhibit deviant behavior forming curve patterns for discharge values exceeding approximately 300m^3/s. The analysis reveals that the outliers primarily cluster during the 19740723 flood event, mainly attributable to training dataset limitations. This extreme event featured both an exceptionally prolonged duration and high peak discharge - characteristics absent from the training data. Consequently, the model demonstrates insufficient capacity to simulate such threshold-exceeding events, yielding suboptimal performance. However, as this represents an extreme scenario, model accuracy is expected to improve with expanded data accumulation.

In the revised manuscript, Page 31, Line 656-668:

While our framework currently caps at 6-hour predictions, extending this horizon requires confronting two fundamental constraints: (1) Input deficiency: The absence of real-time meteorological forecasts prevents runoff anticipation prior to precipitation; (2) Structural saturation: Memory decay in recurrent units limits long-range dependency capture. To address current limitations, future research will pursue a dual-track improvement strategy: Near-term efforts will focus on implementing error correction techniques (including KNN and BP algorithms) coupled with advanced data assimilation methods (such as Ensemble Kalman and Particle filters) to enhance real-time forecasting accuracy, while more fundamental enhancements will involve the strategic integration of numerical weather prediction inputs (particularly ECMWF or CMA-GFS datasets) to enable pre-rainfall runoff anticipation and systematically extend predictive lead times beyond the current 6-hour threshold. Thereby addressing both immediate performance gaps and long-term capability requirements in flood forecasting.

Conclusions section

In the revised manuscript, Page 32, Line 693-704:

In this study, we have successfully incorporated both the attention mechanism and physical mechanism into a DL model to improve the accuracy of flood prediction while ensuring interpretability and physical consistency. While our current framework

demonstrates strong performance within 6-hour predictions, we recognize two key constraints for extending this horizon: the input deficiency due to missing real-time meteorological forecasts and the structural saturation caused by memory decay in recurrent units.To address these limitations, future research will provide improvements through error correction techniques and data assimilation, as well as fundamental enhancements through the integration of ECMWF/CMA-GFS numerical weather prediction inputs to enable pre-rainfall runoff prediction and extend the forecast period beyond 6 hours. Additionally, we suggest exploring other interpretation techniques to deepen understanding of the model's decision-making, while expanding the physical-DL integration through more detailed basin subsurface information and novel combination methods.