RC2:

1. Comments: Clarification on training data size: The manuscript states that only 20 flooding events are used for training, with each event lasting less than 10 days. Could the authors specify the total number of training samples (e.g., input-output pairs or sequences) generated from these events? This information is important for evaluating the robustness and generalizability of the model.

Response: Thank you for raising this important clarification. The training samples were generated through a sliding window approach with a 12-timestep window length and 1-timestep stride for sample construction. This resulted in 2859 unique training samples (input-output sequence pairs). Data size have been added to the revised manuscript.

In the revised manuscript, Page 16, Line 383-384:

The samples are constructed through a sliding window, resulting in the generation of 2859 training samples and 1166 validation samples.

2. Comments: Physics-based loss in PHY-FTMA-LSTM (Line 224-251): Further clarification is needed regarding the implementation of the physics-based loss. Specifically, how are the perturbations δe, δs and δt, introduced during training? Are fixed values pre-specified and added to the input variables? If so, what are the chosen values, and how are they justified? Explicit details on this setup would greatly improve the reproducibility and interpretability of the method.

Response: Thank you for this critical technical inquiry. We didn't introduce fixed perturbation values. As specified in Line 212-218 of our implementation, while keeping other input variables unchanged, we employed the random.uniform function (a random number generator producing values from a uniform distribution within specified bounds) to apply random minor increments within the range [0, 0.1) to the temporal sequences of precipitation, evaporation, and initial watershed soil moisture. This process generated new perturbed temporal sequences for these variables, which were then combined with the unchanged variables' sequences to form modified input datasets.

The difference between the runoff simulation values derived from these perturbed inputs and those from the original inputs was calculated and subsequently transformed into specific loss values via the ReLU function (ensuring non-negative loss). These computed loss values were then incorporated into the overall loss function for model optimization.

We've made minor changes to this section.

In the revised manuscript, Page 10, Line 220-231:

Under the assumption that all other input variables remain unchanged, a new time series of rainfall, evapotranspiration, and initial soil moisture is generated respectively **by applying random minor increments within the range [0, 0.1) using the random.uniform function**. These new time series are then combined with the unchanged time series to form new input data. The difference between the predicted values corresponding to the new data and the predicted values corresponding to the original input data is calculated. This difference is then converted into a specific loss value using the ReLU function and added to the loss function.

3.Comments: Terminology clarification (Line 120): The term "dot product" is typically reserved for operations between vectors, whereas matrix operations such as the one described are more commonly referred to as element-wise multiplication or Hadamard product. Based on the following context, it appears that the authors intended to apply an element-wise product rather than a dot product. I recommend revising the terminology to avoid confusion and ensure mathematical accuracy.

Response: We sincerely appreciate this precise technical correction. You are absolutely correct that the operation described in Line 120 constitutes an element-wise multiplication (Hadamard product) rather than a dot product. We have revised this terminology throughout the manuscript to ensure mathematical accuracy.

In the revised manuscript, Page 5, Line 125:

By taking the element-wise product of these two matrices, the model generates the feature-time-based attention matrix.

4.Comments: Undefined abbreviations (Line 181): The abbreviations FA and TA are introduced without prior definition. For clarity, all abbreviations should be clearly defined at first mention to ensure readability for a broad academic audience.

Response: We sincerely appreciate this careful observation. You are absolutely correct that the abbreviations "FA" and "TA" are inadvertently undefined at first mention in Line 181. We have implemented the revisions to ensure clarity.

In the revised manuscript, Page 7, Line 188-189:

where FA represents feature-based attention weight matrix, TA represents time-based attention weight matrix.

5.Comments: Figure 1 clarity: Figure 1, particularly subplot (b), is difficult to interpret. The label "head_m" appears to encompass multiple attention mechanisms, including feature attention, time attention, and feature-time attention — not solely multi-head attention as the label may imply. I suggest renaming the label in subplot (b) to more accurately reflect its composite structure and enhance reader comprehension.

Response: We appreciate this feedback. To enhance clarity, we have revised the label for subplot (b) to: Per-head Feature-Time Attention. This represents the formation of feature-based attention and time-based attention matrices from inputs followed by element-wise product.

In the revised manuscript, Page 9, Line 221:

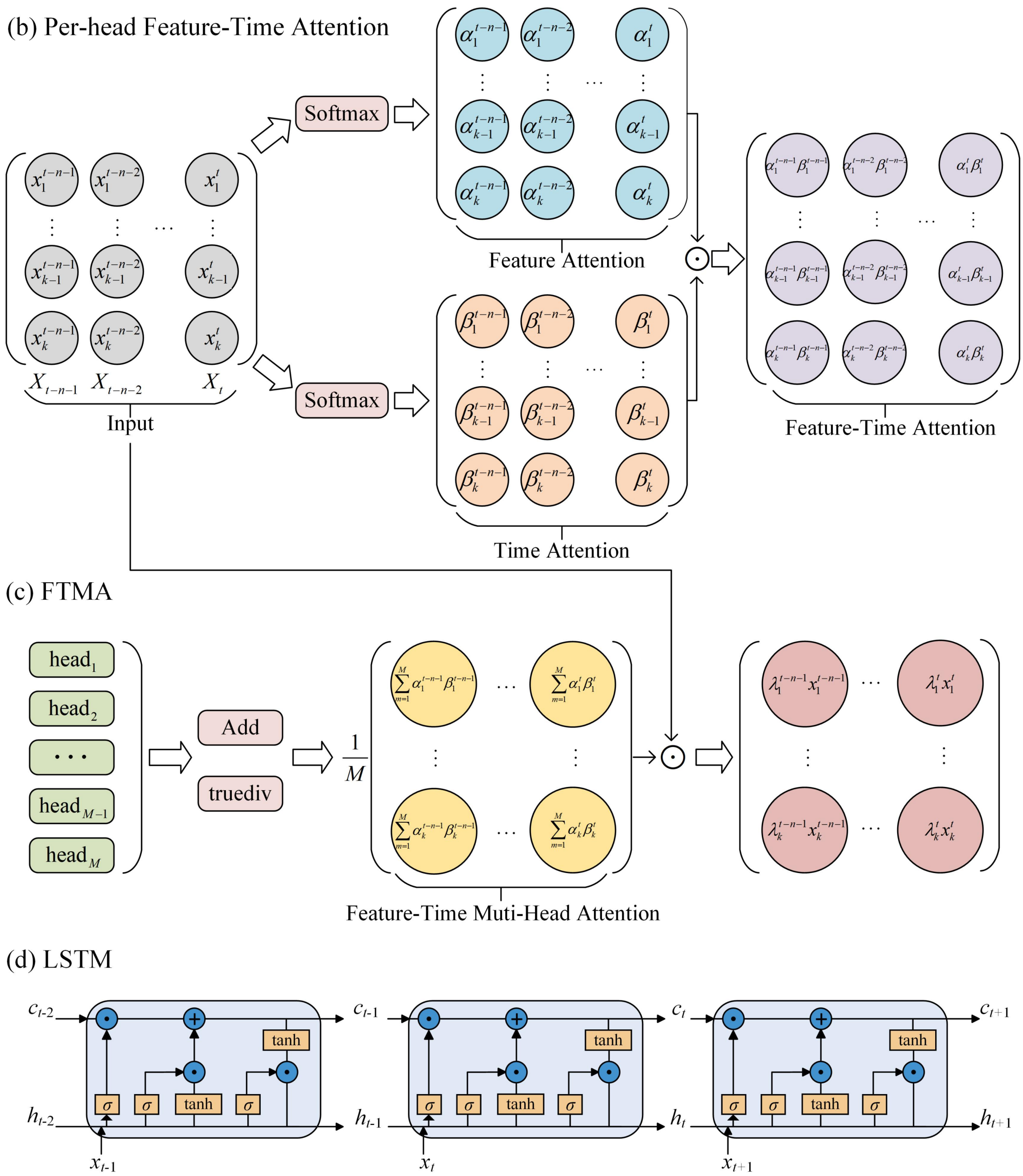

Fig. 1. (a) The PHY-FTMA-LSTM model architecture. (b) **Feature-time-based attention matrix generation process for each attention head**. (c) Feature-time-based multi-head attention workflow. (d) The internals of LSTM cells.

6.Comments: Labeling in Figure 5: In Figure 5, it would be more intuitive to label the x-axis using calendar dates (e.g., MM-DD-HH) rather than elapsed time in hours. Using time in hours may be easily confused with forecast lead times, potentially causing misinterpretation. I recommend updating the x-axis to calendar dates to improve clarity and reader understanding.

Response: Thank you for this constructive suggestion. We agree that using calendar dates on the x-axis of Figure 5 would enhance temporal clarity and avoid potential confusion with forecast lead times. We have updated x-axis labels to MM-DD-HH format.

In the revised manuscript, Page 25-27, Line 548-564:

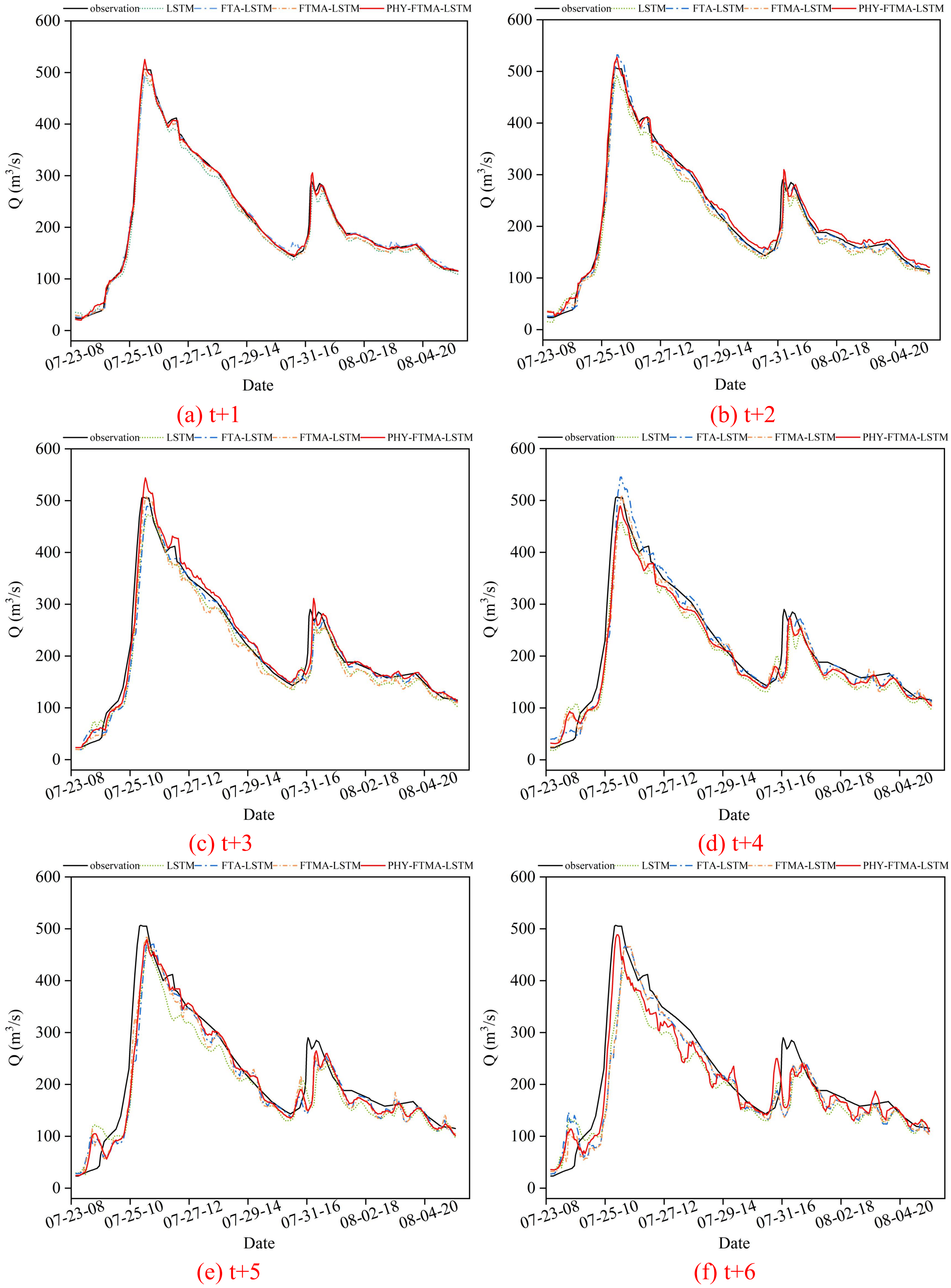

Fig.7. Comparison of observed and predicted values of the 19740723 flood event by the four models.(The x-axis displays dates in MM-DD-HH format, representing month, day, and hour

respectively)

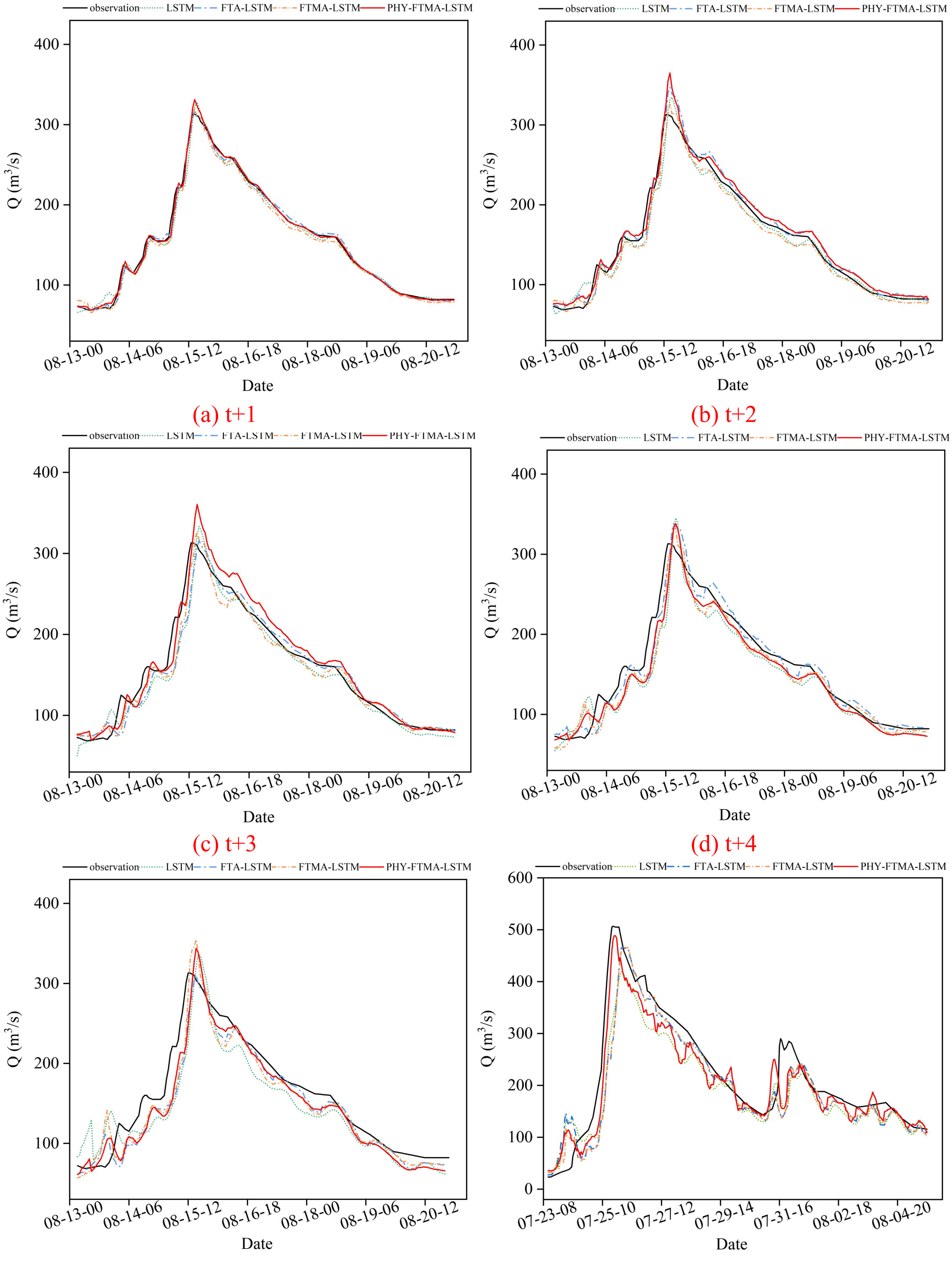

Fig.8. Comparison of observed and predicted values of the 19790813 flood event by the four models.