***Reviewer #2***

The authors evaluate different transformer models under different scenarios (i.e., regression, data integration, and autoregression) in comparison with benchmark models, LSTM networks and DLinear, using various hydrologic datasets. In addition, they compare the performance of LSTM networks with pre-trained LLM in the zero-shot forecasting for autoregression tasks. They show that LSTM networks outperform transformers in regression and data integration tasks, and attention-based methods surpass LSTM networks in autoregression and zero-shot forecasting. The paper is well written, with deep discussion. I have minor comments below.

For boarder audiences with minor ML/DL backgrounds, it would be helpful to provide brief introductions of the DL models used in the manuscript. The information can be provided in the SI if there are limited spaces in the main text. Table 1 provides the main features of different variants of transformers. Since ML/DL has many unique terms, simply providing the names of features does not really help understand their differences. Maybe the authors can adapt the table with more general features, such as "Trained on time series".

We appreciate the reviewer's suggestion. To improve accessibility for broader audiences with limited ML/DL backgrounds, we will:

- Include brief and intuitive descriptions of each deep learning model in the Supplementary Information. These descriptions will explain the key principles, and features, helping readers clearly understand their differences.
- Revise Table 1 in the main manuscript to highlight more general and easily understandable characteristics, making distinctions between the models clearer for readers without extensive ML/DL expertise.

In Section 2.3-attention models, it is better to mention that the authors use pre-trained LLM for zero-shot forecasting. This information come out in Section 2.4.5. In addition, for zero-shot forecasting, the authors only compare DL model performance in autoregression tasks, and state that LSTM underperforms pre-trained LLMs in zero-shot forecasting. How about regression and data integration tasks? I would expect that the LLMs cannot really understand the relationship between input and output without fine-tuning.

We appreciate the reviewer's suggestion. To address this comment, we will clarify in Section 2.3 that we use pre-trained Large Language Models (LLMs) for zero-shot forecasting, referencing Section 2.4.5 for further details. Additionally, we agree with the reviewer about the limitations of pre-trained LLMs without domain-specific fine-tuning. We will mention this limitation and indicate that pre-trained LLMs primarily rely on contextual patterns rather than genuine hydrological understanding. Finally, we will conduct additional experiments on regression and data integration tasks to verify the reviewer's expectation that pre-trained LLMs struggle without fine-tuning.

Specific comments:

Some figures for equations are burr, like Equation 5.

We appreciate the reviewer's suggestion. We will retype Equation 5 and other affected equations to ensure clarity and readability.

Equation 6: stating that r is Pearson’s correlation coefficient.

We appreciate the reviewer’s suggestion. We will clarify Equation 6 by stating that r represents Pearson’s correlation coefficient between simulated and observed values.

Equation 8: I think the equation is wrong. It should be (RMSE^2-Bias^2)^0.5.

We appreciate the reviewer’s suggestion. Although our original formulation is mathematically equivalent, we agree that your proposed version is clearer and more intuitive. We will revise it accordingly.