The paper developed a hybrid framework that integrates a distributed process-based hydrological model and embedded neural networks (ENNs) for streamflow modeling in large alpine basins. The distributed EXP-Hydro model uses multiple mathematical equations to describe hydrological systems, including precipitation, snowmelt, runoff, and baseflow, which can be replaced by neural networks. The hybrid framework performs well in both gauged and ungauged basins across three large alpine basins. My major concerns are as follows:

Response: Thanks for your recognition and valuable suggestions. Please find our replies below.

Major comments:

1. I suggest the authors rewrite the abstract, as it is too long. Some sentences should be moved to the introduction or results sections of the manuscript.

Response: Thanks for your suggestion. We will rewrite the abstract in the revised manuscript.

"Alpine basins are important water sources for human life and reliable hydrological modeling can enhance the water resource management in alpine basins. Recently, hybrid hydrological models, coupling process-based models and deep learning, exhibit considerable promise in hydrological simulations. However, a notable limitation of existing hybrid models lies in their failure to incorporate spatial information within the basin and describe alpine hydrological processes, which restricts their applicability in hydrological modeling in large alpine basins. To address this issue, we develop a set of hybrid distributed hydrological models by employing a distributed process-based model as the backbone, and utilizing embedded neural networks (ENNs) to parameterize and replace different internal modules. The proposed models are tested on three large alpine basins on the Tibetan Plateau. A climate perturbation method is further used to test the applicability of the hybrid models to analyze the hydrological sensitivities to climate change in large alpine basins. Results indicate that proposed hybrid hydrological models can perform well in predicting runoff processes and simulating runoff component contributions in large alpine basins. The optimal hybrid model with Nash-Sutcliffe efficiency coefficients (*NSEs*) higher than 0.87 shows comparable performance to state-of-the-art DL models. The hybrid distributed model also exhibits remarkable capability in simulating hydrological processes at ungauged sites within the basin, markedly surpassing traditional distributed models. Besides, the results also show reasonable patterns in the analysis of the hydrological sensitivities to climate change. Overall, this study provides a high-performance tool enriched with explicit hydrological knowledge for hydrological prediction and improves our understanding about the hydrological sensitivities to climate change in large alpine basins."

2. The differences between the distributed models and the corresponding lumped models are unclear. From the manuscript, it appears that the only difference is that the lumped model simulates discharge for the entire basin, while the distributed model simulates discharge for each subbasin, and then summarizes the discharge for all the subbasins. Runoff routing is an important process in distributed hydrological models, which is also crucial for large basins. Please explain why river routing is missing.

Response: Thanks for your suggestion. In this study, we employ the distributed EXP-Hydro model as the backbone model. Compared with the lumped version, the distributed EXP-Hydro model first delineate the entire basin into many sub-basins, and all hydrological processes are calculated in each sub-basin. The final basin runoff is acquired by summing the runoff outputs from all basins. Besides, our hybrid models utilized ENNs to parameterize and replace internal modules. We used static basin variables as the inputs of ENNs to represent the spatial heterogeneity within different sub-basins. On the other hand, we agree with the reviewer that the routing method is important for hydrological modeling, especially in large basins. However, to achieve the coupling between physical models and neural networks and the simultaneous training of both the physical models and neural networks, all equations are formulated to be differentiable to ensure operating within the differential programming framework (DPF). The technical requirements of DPF limit the consideration of routing methods in our hybrid hydrological models. To compensate for the lack of consideration of the routing process, we calculate the river length from each sub-basin to the basin outlet and employ this static attribute as the inputs of ENNs to implicitly characterize the routing process within the basin. We will discuss this limitation in the revised manuscript.

3. Please demonstrate the importance of using subbasins in alpine basins due to the significant variability of precipitation and temperature in space. Additionally, the sensitivity of the area threshold for the subbasins is not discussed in the manuscript. While the authors may have experience defining the threshold in Tibetan basins, it is unclear how this applies to other basins

Response: Thanks for your suggestion. Many studies have demonstrated that our study basins exhibit significant spatial heterogeneity in precipitation and air temperature due to large topographical variations and complex weather systems (Ma et al. 2018, You et al. 2015). We will add this discussion in the revised manuscript. Besides, we used the green lines in Figure 2 to show the delineated river networks within three basins, which determines the shape and number of delineated sub-basins. Referring to the number of sub-basins divided by the THREW model, we delineated the Yellow, Yangtze, and Lancang into 83, 99, and 63 sub-basins. The detailed sub-basins information will be added in the revised manuscript.

4. The significance of model performance is not discussed in the manuscript. For example, DMθ-Q-T and DMθ-QSM-T have very close NSE values in the Yellow River and Lancang River. If the authors only trained the model once, it is unclear if the differences are statistically significant.

Response: We agree with the reviewer that a slight improvement in the NSE does not significantly demonstrate an enhancement of the model. In the revised manuscript, we will reassess the improvements of these models to enhance the credibility of the results.

5. The authors conducted a series of sensitivity tests of runoff to climate change. However, it is difficult to explain the internal structure of a neural network and how we can trust the extrapolated results. For example, the model was not trained on a 20% increase in precipitation, meaning the perturbed scenarios are extrapolations. It would be more accurate to refer to this as model sensitivity to dynamic inputs rather than concluding runoff sensitivities to climate change.

Response: Many studies demonstrated that the performance of deep learning in simulating data outside the training range is significantly lower than within the training range. In this study, we introduced certain physical mechanisms into the deep learning model to enhance the physical consistency of the simulation results. To evaluate the model's performance in simulating data outside the training range, we used the climate perturbation method to assess the sensitivity of runoff processes to changes in temperature and precipitation. Although we did not use the perturbed data for training, our results were compared with existing studies, demonstrating the reasonableness of our simulation results and the ability to analyze the sensitivity of runoff processes to climate change. Besides, numerous studies have employed similar methods, using physical hydrological models to evaluate the sensitivity of runoff processes to climate change (Cui et al. 2023). We will include additional explanations in the revised manuscript.

6. The improvement in streamflow estimation is important. However, it would be interesting to investigate when and where these improvements occur. Please analyze the spatial differences between the deep learning models and the EXP-Hydro model in simulated discharge

Response: Thanks for your suggestion. This study employed three metrics, including NSE, mNSE and PFAB, to evaluate the model improvement in different aspects. To further investigate when and where these improvements occur, we will add some analysis in the revised manuscript.

7. I found it hard to follow many sentences; please polish the language. Some examples are listed below.

Response: Thanks for your suggestion. We will polish the language in the full manuscript.

Minor comments:

1. Line 25: Alpine basins are important water sources, playing a crucial role in various aspects of human life and the environment, such as domestic water supply, irrigation, hydropower generation, and climate regulation. Please rewrite the sentence.

Response: Thanks for your suggestion. This sentence will be revised in the manuscript.

2. Line 26: The performance of a hydrological model can be accurate, to describe the model, use reliable could be better.

Response: The “accurate” has been revised as “reliable” in the revised manuscript.

3. Line 27: shorten the sentence and use ‘climate change and adaption’.

Response: This sentence is revised as “Developing reliable hydrological models is crucial for managing floods and improving water use efficiency under climate change.”.

4. Line 31: These models depend on physical laws and empirical knowledge.

Response: This sentence is revised as “These models depend on physical laws and empirical knowledge to describe physical processes and are grounded in well-defined physical mechanisms.”

5. Line 32-34: The sentence is too long. In addition, are these hydrological models sufficient to understand all hydrological processes?

Response: It will be revised as “They can be used to advance scientific understanding about the hydrological systems and provide the insight into the response of hydrological processes to climate changes”

6. Line 41: streamflow/discharge forecasting, snow water equivalent modeling, and groundwater level mapping. Please rewrite the sentence.

Response: We agree with the reviewer and the rewritten sentence is “They showcased exceptional model performance across diverse hydrological domains, including streamflow/discharge forecasting (Kratzert et al. 2018, Lees et al. 2021, Liu et al. 2021), snow water equivalent modeling (Duan and Ullrich 2021), and groundwater level mapping (Nourani et al. 2022, Solgi et al. 2021). ”

7. Figure 2. Please add some subplots to show the spatial variability of precipitation and temperature, which is the main reason for using the distributed schemes. Please show the subbasins and indicate the amount of subbasins.

Response: We agree with the revised manuscript and we will add some subplots to show the spatial variability of precipitation and air temperature and sub-basins.

8. Line 86: …the proposed models…

Response: Thanks for your suggestion and we will revise in the revised version.

9. Line 87-88: Can the ENNs produce optimal parameters?

Response: The differential programming framework ensures that the training parameters of hybrid models are similar to those of the deep learning model. By utilizing sufficient observed runoff data, although it cannot ensure obtaining the optimal parameters, it does ensure that the parameters are as fully trained as possible.

10. Line 203: The training period is 26 years and the evaluation/testing period is only 6 years. Is this setting reasonable? Why not set the same length for the training and testing? Please explain.

Response: The proposed hybrid models, similar to deep learning, have numerous parameters that need to be trained, requiring a large amount of observational data. Due to the limited availability of observed data, we set the training period to 26 years and the testing period to 6 years. To ensure a fair comparison, we set the calibration/training and validation periods for the comparison models, including the physical model and the deep learning model, to be the same as those for the hybrid models.

11. Line 247: I don't think an improvement of NSE from 0.06 to 0.09 is a substantial improvement. Please rewrite the sentence.

Response: We agree with the reviewer and the "substantial" and "noteworthy" have been revised as "slight" and "small".

Cui, T., Li, Y., Yang, L., et al. (2023). Non-monotonic changes in Asian Water Towers' streamflow at increasing warming levels. Nat Commun 14(1), 1176.

Duan, S. and Ullrich, P. (2021). A comprehensive investigation of machine learning models for estimating daily snow water equivalent over the Western US. Earth and Space Science Open Archive.

Kratzert, F., Klotz, D., Brenner, C., et al. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences 22(11), 6005-6022.

Lees, T., Buechel, M., Anderson, B., et al. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. Hydrology and Earth System Sciences 25(10), 5517-5534.

Liu, Y., Zhang, T., Kang, A., et al. (2021). Research on Runoff Simulations Using Deep-Learning

Methods. Sustainability 13(3), 1336.

Ma, Z., Xu, Y., Peng, J., et al. (2018). Spatial and temporal precipitation patterns characterized by TRMM TMPA over the Qinghai-Tibetan plateau and surroundings. International journal of remote sensing 39(12), 3891-3907.

Nourani, V., Khodkar, K. and Gebremichael, M. (2022). Uncertainty assessment of LSTM based groundwater level predictions. Hydrological Sciences Journal 67(5), 773-790.

Solgi, R., Loaiciga, H.A. and Kram, M. (2021). Long short-term memory neural network (LSTM-NN) for aquifer level time series forecasting using in-situ piezometric observations. Journal of Hydrology 601, 126800.

You, Q., Min, J., Zhang, W., et al. (2015). Comparison of multiple datasets with gridded precipitation observations over the Tibetan Plateau. Climate Dynamics 45, 791-806.