1. Scope

The scope of the paper is well suited for HESS.

2. Summary

The authors based their study on the hybrid model proposed by Li et al. (2023b), in which different modules of a conceptual hydrological model are substituted by Neural Networks. The authors then proposed a distributed version of the model, in which they subdivide the basins of interest into smaller subbasins and apply the hybrid models to each of these subbasins.

The authors then compare the performance of the lumped and distributed hybrid models against purely data-driven techniques (LSTM and CNN-LSTM) and show all models achieved similar performance. They also test the performance of the hybrid models to predict discharges in some of the predefined subbasins (untrained gauges). In the last sections, the authors run some experiments looking at the behaviour of the models when boundary conditions are modified (changes in precipitation and temperature).

3. Evaluation

Overall, the manuscript has the potential to be a good contribution, however, there are certain aspects mentioned in the questions below that should be taken into account before moving on to the next steps.

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

3.1 Major comments:

- **The code is not published, and the authors indicate that it will be opened once the manuscript is accepted. I strongly recommend the editor to ask for an open code during the review process, as it increases the transparency of the study. I also tried to look for the code of the previous study (Li et al., 2023b) however I was not able to find it.**

  Response: Thanks for your suggestions. We will open the source code in the revised manuscript.

- **They also indicate that the discharge information is not publicly available due**

to privacy reasons. Even though this reason is valid and outside of the capabilities of the authors, it automatically makes the study non-reproducible, which is especially important when machine learning methods are being proposed.

Response: Thanks for your suggestions. To enhance the transparency and reproducible of our study, we will provide the simulated results in Yellow in the revised manuscript.

- **The printing quality of all figures should be improved. When I zoom in, I cannot see the details. I suggest the authors print the figures in 300 dpi.**

Response: We will improve the printing quality in the revised manuscript.

- **The authors do not show the subbasins they used to create the distributed model. I encourage them to include this information. Also, I was not able to find information of the amount of subbasins they used.**

Response: 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 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.

- **One major concern is that they are not considering any routing method. Consequently, even with a distributed model they just sum up the discharge coming from each subbasin. Moreover, the authors are working with large basins (over 90 000 km2 according to the manuscript) in which the routing processes can become highly relevant. Is there a reason why no routing is being used?**

Response: We agree with the reviewer that the routing method is important for hydrological modeling, especially in large basins. In this study, 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.

- **The good performance of the distributed models can also be attributed to the fact that one has a more flexible model. More flexible models can get a better fit to the data, but this is not directly related to having a distributed version. One way to test this hypothesis is to use the same number of models you used in the distributed version but let them receive the same data (similar to Feng, 2022). If your models performed better, then you can say that the distributed nature of the models is beneficial, and the improvement is not just because of the increase in flexibility. If the performance is the same, then it would mean that the distributed version (especially without any routing) does not give an advantage.**

Response: Feng et al. (2022) conceptualized each basin as being composed of multiple parallel components to represent spatial heterogeneity. The inputs of all components are assumed to be same and the parameters in different components are acquired independently. So theses models only enhanced model flexibility by setting more parameters but did not represent the spatial heterogeneity in the real world. Different from Feng et al. (2022), we developed hybrid models in a distributed manner similar with traditional distributed hydrological models. All parameters of physical models are calculated based on the static attributes of sub-basins. And all ENNs for replacement of different internal modules utilized static attributes and dynamic driving variables as inputs. Our models can better represent spatial heterogeneity and are less flexible than models of Feng et al. (2022). And if the model inputs of all sub-basins are set the same, parameters in all sub-basins will be same, which means that all outputs are also same and the final results are same or similar with lumped models. Besides, Patil and Stieglitz (2015) demonstrated that our backbone model, the distributed EXP-Hydro model, outperform the lumped version by capturing spatial heterogeneity. Therefore, compared with lumped models, our distributed hybrid models are enhanced due to the consideration of spatial heterogeneity but not the improvement of model flexibility.

- **In Section 3.2 the authors present the results of the evaluation on untrained gauges. One can see that for the MT subbasin the performance of three models completely drops when compared to the performance in the original basin (TNH). Why? In the text it was mentioned that "the models show comparatively poorer performance in runoff modeling at the MT station" but it was never explained why. Also, the discharge range (Figure 5) of the subbasins is similar, so is there a reason for the big differences in performance?**

Response: Thanks for your suggestion. The upper reaches of the source region of the Yellow (the basin above the MT Hydrological Station) are characterized by high altitude and low temperatures. The alpine hydrological processes, including soil freeze-thaw, snow, and ice processes, significantly impact runoff processes. These influences are more pronounced compared to the entire source region of the Yellow, making hydrological modeling more challenging. The distributer EXP-Hydro, lacks the capability to accurately depict these alpine hydrological processes, resulting in relatively low simulation accuracy. In contrast, the ENNs, which represent these processes, are better at capturing them, demonstrating superior accuracy. This also proves that the coupled model constructed in this study can effectively learn and represent these hydrological processes. We will add detailed explain in the revised manuscript.

3.2 Minor comments:

- **Line 25: Is the word 'almost' a typo?**

Response: "almost" has been deleted in the revised version.

- **Line 32: Saying that process-based models can be used to understand the entire hydrological system including all internal processes is an overstatement, especially if you are referring to conceptual models. Conceptual models are mostly based on parameterized (empirical) relationships that somehow account for our understanding of the system, however the physics behind them is not much.**

Response: We agree with the reviewer that it is an overstatement. This sentence is 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".

- **Line 122: It is not clear what the authors mean. Are the parameters the same for all basins or do they change?**

Response: We want to show that the parameters of all sub-basins within the basin are assumed to be the same in distributed EXP-Hydro model. It has been revised as "The calibration parameters of all sub-basins within the basin are assumed to be the same in distributed EXP-Hydro model, while many of them related to sub-basin attributes should be different".

- **Line 145: It would be better to create a table to specify the characteristics of each model.**

Response: Thanks for your suggestion. We will create a table to specify the characteristics of each model.

Table 1. Design details of different hybrid models. "√" represents that the model employs the corresponding ENNs while "×" means not.

Model	ENN_{θ}	ENN_{S}	ENN_{M}	ENN_{Q}	Temperature is the input of ENN_{Q}
DM	×	×	×	×	×
DM_{θ}	√	×	×	×	×
$DM_{\theta-Q}$	√	×	×	√	×
$DM_{\theta-Q-T}$	√	×	×	√	√
$DM_{\theta-QSM}$	√	√	√	√	×
$DM_{\theta-QSM-T}$	√	√	√	√	√

- **Line 158: Hochreiter and Schmidhuber, 1997 created the LSTM architecture, but in this line, it sounds like they proposed the architecture for hydrological modelling. This paper should of course be cited, but not mixed with the other papers of hydrological applications.**

Response: We have revised this citation in the next version.

- **Figure 2. How was the spatial discretization of the basins? This should be included in the paper.**

Response: Similar with the fourth comment and response in the major comment, we will add the detailed description about spatial discretization of the basin.

- **Line 195: What does suites of experiments mean? Shouldn´t it be set of experiments?**

Response: We will revised as "set of experiments" in the revised manuscript.

- **Line 242-244: The PFAB values show a difference, but the changes of 0.01 in NSE are not significant. This can be just because of the initialization of the model. I do not agree that there is evidence to support that one model has an augmented ability to simulate overall runoff process.**

Response: We agree with the reviewer that the DM_θ model cannot be demonstrated to achieve an augmented ability to simulate overall runoff process. It has been revised as "Specifically, the comparison results show that DM_θ model exhibits a closed but slightly better performance than the DM model in overall runoff modeling, with an increase NSE and $mNSE$ of 0.01-0.03 in all three basins. Additionally, lower $PFAB$ results imply that the DM_θ model contributes to an improved performance in peak runoff modeling. The incorporation of ENNs to represent spatial heterogeneity of calibration parameters can reduce the peak simulation biases and slightly improve the overall performance."

- **Figure 3: The size of the figure should be increased, right now it is hard to see the details of the hydrographs. Also, the authors should use the same line width for all models. Right now, some lines look thicker than others, which gives a bias to the figure. The last hydrograph does not have values on the y-axis.**

Response: We will increase the size of all figures and revise the line width in Figure 3 in the revised manuscript.

- **Line 263: The authors say that when one includes air temperature in the ENN there is an evident enhancement of the model performance. The PFAB values vary a bit more, but the differences in NSE values are extremely small (0.01 for 4 cases, 0.02 for 1 case and 0.03 for another). This is just one metric summarizing more than 5 years of data. I do not agree that NSE values show an evident enhancement of the model performance.**

Response: We agree with the reviewer. This statement has been revised as "Results indicate that $DM_{\theta-QSM-T}$ and $DM_{\theta-Q-T}$ models exhibit improved performance in peaking runoff modeling compared to the $DM_{\theta-QSM}$ and $DM_{\theta-Q}$ models, respectively. This enhancement is evident through closed NSE and $mNSE$ and lower $PFAB$ in all three basins".

- **Figure 5e: How are the attributes being normalized? I do not understand how the area of the subbasins is comparable with the area of the entire basin. Also, why is the figure showing a range when referring to static attributes? This figure is not explained in the text.**

Response: Thanks for your comment. Figure 5e shows 9 static attributes, except river length in Table A1, of all sub-basins within the basin. These static attributes are additional inputs of ENNs to capture the spatial heterogeneity among sub-basins among basins. Four lines represent static attribute ranges of the four basins with the four hydrological stations (TNH, JG, MQ, and MT) as the outlet. Among them, the basin with the TNH as the outlet is the largest basins and contains other three basins. These attributes of all sub-basins in four basins are normalized based on Equation 1. The normalized attributes are final inputs of ENNs. Figure 5 are intended to show the meteorological and hydrological difference among four basins. We will add some description about Figure 5 in the revised manuscript.

$$x_N = \frac{x - x_{min-TNH}}{x_{max-TNH} - x_{min-TNH}} \quad (1)$$

x and x_N represent the initial and normalized attributes, respectively. $x_{min-TNH}$ and $x_{max-TNH}$ represent the minimum and maximum variables among sub-basins within the basin with the TNH as the outlet.

- **Figure 7. The figure title indicates that the grey and yellow shading indicate annual and monthly responses. However, there are no shadings in the figure.**

Response: We want to illustrate that annual responses are showed in the grey background while monthly responses in the yellow background. The caption of Figure 7 is revised as "Relative change of annual (grey background) and monthly(yellow background) runoff response to the perturbed precipitation (a-c) and air temperature (d-

f) in Yellow, Yangtze, and Lancang, respectively".

- **Figure 8. I suggest the authors use a proper name for the figure and not just refer to another figure.**

Response: It will be revised as "**Figure 8**. Runoff responses to altered temperature in the Yellow, Yangtze, and Lancang basins (a-c for annual; d-f for monthly). The error bars in panels a-c and the shaded areas in panels d-f denote the range of simulated runoff"

Feng, D., Liu, J., Lawson, K., et al. (2022). Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. Water Resources Research 58(10).

Patil, S.D. and Stieglitz, M. (2015). Comparing spatial and temporal transferability of hydrological model parameters. Journal of Hydrology 525, 409-417.