Major Comments:

1. The structure of the manuscript may need to be optimized to make it easier to read. For instance, Section 2.3 is titled “Differences with the Hydrologic Reservoir”, but I am hard to get the differences between LSTM and the Hydrologic Reservoir; some terms, such as optimal lag memory, significant values for weights and so on, are not well defined in the manuscript; I am confused about how Experiment 2 helped with the topic.

   Response: First, we describe the similarities between both representations and the slight differences within them. In section 2.3, we outline what we consider the main two differences: state evolution and gating behavior. The first one describes how LSTM undergoes continuous warming up at each time step to approximate the state value. The second one describes the information used to infer the gating behavior. Since section 2.3 is fundamental for understanding the proposed structure, we will add more details and clarifications in this section.

2. The manuscript uses a long length to compare linear reservoir model and LSTM. However, the linear reservoir model is not used in the case study. Is it possible to use the linear reservoir model as the benchmark model to relate the parameters of the linear reservoir model and LSTM in order to discuss the physical meaning of LSTM parameters.

   Response: The analogy between the linear reservoir and LSTM is employed solely to elucidate the functioning of the representation. The concept of utilizing a conceptual lumped model as a benchmark in terms of performance and certain global behaviors has been employed in prior research (De la Fuente et al., 2023). Nevertheless, this comparison serves only to comprehend the advantages and shortcomings of each representation. Extracting the knowledge encoded within an ML model necessitates additional steps, which are currently being explored in the paper.

   The relationship between the parameters used by a linear reservoir and LSTM has not been explored. However, the differences between the two representations in terms of states (water vs. information), state tracking (continuous vs. warming up), and gate behavior (constant vs. dynamic) do not guarantee a correct and unique relationship between them.

3. My primary concern is that the HydroLSTM/LSTM is not well validated. From Figure 4, the performance of HydroLSTM/LSTM is not very good. Only 5 out of 10 basins have a KGE value larger than 0.7. In some previous studies, most runoff simulations with a local LSTM can obtain a KGE greater than 0.7. I am concerned about whether

the parameters and structure of the ML model are well set. Further, for a ML model with poor performance, the interpretability of the model does not seem to be very meaningful. Also, I wonder why the authors did not use more catchments to verify the reliability of the model. In my opinion, the summary of 588 catchments makes Figure 4 more credible.

Response: Local models have the potential to achieve overall good performance when the inputs are carefully selected. However, in our case, we are utilizing a parsimonious representation that only includes precipitation and temperature as inputs. This limitation restricts the maximum performance regardless of where the representation is applied.

In Figure 4, our objective was to demonstrate that under these simplistic conditions, LSTM and HydroLSTM exhibit similar performance. Furthermore, the selection of catchments in our study encompasses representative regions with traditionally good performance (wet catchments) and poor performance (arid catchments), which is consistent with the range of results shown in Figure 4b.

Additional cases could be included in Figure 4; however, the overall situation is unlikely to change significantly since HydroLSTM is merely a modification of the LSTM representation. Therefore, in terms of performance, they are expected to perform similarly, resulting in more data points clustered around the 1:1 line. Another limitation is the computational time required, as each catchment involves 20 runs multiplied by 8 lags (2, 4, 8, 16, 32, 64, 128, and 256 days) multiplied by 6 cells (1, 2, 3, 4, 8, 16 cells) multiplied by 2 representations, resulting in a total of 1920 models. This computational constraint restricts the analysis to a subset such as the entire dataset.

4. From Figure 5, the uncertainty of the model parameters seems large. The large uncertainty of parameters may make the model less interpretable. I think it is necessary to explain the effect of parameter uncertainty on the model.

Response: We acknowledge that the uncertainty in the weight values is substantial, indicating a high degree of freedom and equifinality issues. However, the overall pattern per catchment remains consistent (Fig. C1b), which is a novel finding. This is noteworthy because weight values in machine learning models are typically considered random and non-interpretable. We consider this as one of the key outcomes of our study since it signifies the potential for extracting knowledge and enhancing interpretability from ML models. We will provide further explanations regarding the uncertainty in our upcoming revisions, adding more clarity to this aspect.

5. Why does Figure 5 use the logarithmic horizontal axis? If using a regular coordinate axis, I think it is hard to distinguish the fluctuation of precip weights between 0-10 days and after 10 days in the ID11473900 catchment. The fluctuation of Pot. Evapot. Weights in a regular coordinate axis seem to be a periodic variation in the ID9035900 catchment. The logarithmic horizontal axis may mislead readers into thinking that there is a trend from 0-1 days.

   Response: We opted for a logarithmic scale because the highest weight values are typically found within the 0-10 day range (Figure C1), and the relative importance of past information tends to diminish for longer lags. However, we observed some periodic behavior for longer lags, indicating that weight distributions carry informative signals about the relationship utilized by the ML model, and these distributions are specific to each catchment. To prevent any potential misinterpretation, we will include figures on a regular scale in the supplementary material, along with further explanations regarding the use of a logarithmic scale.

6. The manuscript analyses the physical meaning of the output gate. I'm wondering if the forget gate and input gate have a corresponding interpretation.

   Response: Effectively, the forget and input gates do possess some level of interpretability. However, their interpretation is more closely tied to the state variable and the nature of the input employed. Consequently, interpreting these gates directly is not feasible without appropriate regularization. For instance, the state variable may be storing diverse forms of pertinent information, making it difficult to determine the exact extent to which the model should remember or forget. Currently, there are ongoing works that concentrate on imposing constraints on the storage of specific entities such as volume and energy. In such cases, it may be possible to derive meaningful interpretations.

7. Why does Experiment 2 classify the catchments with Aridity rather than catchment dominance factors in Experiment 1?

   Response: The criteria presented in Table 1 were indeed developed using 160 catchments of the MOPEX dataset. Since this research used 588 catchments of CAMELS dataset, with only a small overlap between them, a direct comparison between the two datasets is not feasible. However, to address this challenge, we are exploring the possibility of incorporating clustering techniques for the catchments under study. By employing clustering, we aim to establish a meaningful comparison between the evaluated catchments and the criteria presented in Table 1, despite the differences between the datasets.

8. How is the optimal number of lagged days obtained in Experiment 2? From Figure 8, I think the optimal numbers of lagged days of the catchments with AI<0.6 and AI>1.0 are also 128 days. There needs to be more discussion about the relationship between required memory time scales and aridity.

   Response: The description of the best lag is based on the median value (represented by the red line in the boxplot). However, it is important to note that this value serves as an overall summary of the trend where increasing lag corresponds to higher aridity levels. It is crucial to recognize that catchment memory is influenced by various factors beyond just aridity. Therefore, it is not possible to strictly define the "best" lag for a specific level of aridity. Rather, we can only observe that a relationship exists between aridity and lag. We will provide additional discussion on this topic to further elaborate on the complexities and limitations involved.

Minor Comments:

1. We usually use the Hydrologic Reservoir model rather than the “Hydrologic Reservoir”. Just a suggestion.

   Response: We will incorporate the suggestion.

2. Table 2. It is necessary to explain the difference between “Recent” and “Historical”.

   Response: We will include a summary in terms of the classification mentioned in the paper.

3. Figure 4. How to choose the “red *”?

   Response: The "red*" indicates the best performance achieved by Hydro-LSTM when using one or two cells, where at least one "*" is present in the LSTM parameter sets. We employ this approach to demonstrate that Hydro-LSTM exhibits a parsimonious state representation while achieving similar performance compared to LSTM.

De la Fuente, L. A., Gupta, H. V., & Condon, L. E. (2023). Toward a multi-representational approach to prediction and understanding, in support of discovery in hydrology. *Water Resources Research*, 59, e2021WR031548. https://doi.org/10.1029/2021WR031548