Dear Tadd,

Thank you for your kind comments. We will incorporate your suggestions and clarify parts that were not clear enough.

Best regards,

De la Fuente et al.

Hello,

Thank you for the lovely preprint. I enjoyed reading your work and offer the following suggestions below. I believe the paper should be reconsidered for HESS, with major revisions, and look forward to reading the next submission.

Best,

Tadd Bindas

Editorial questions:

1. Does the paper address relevant scientific questions within the scope of HESS?
    1. Yes.
2. Does the paper present novel concepts, ideas, tools, or data?
    1. The concept proposed by their HydroLSTM model is novel. The authors are looking to add more interpretability to the LSTM architecture and get similar results with fewer cell-states using the HydroLSTM code they developed.
3. Are substantial conclusions reached?
    1. I'm not sure. As a summary of my understanding of the paper: the results obtained from their first experiment showcase that a simplified LSTM framework (similar to our understanding of a reservoir) can use one cell to learn a relationship between inputted forcings. The second experiment shows how their model performs when compared to 588 CAMELS basin observations.

        Response: We agree. We would like to add that the second experiment explores how lag memory is another hydrological characteristic that is encoded in the weight pattern. This finding reinforces the idea that catchment attributes play a role distinct from meteorological forcing.

    2. My confusion arises with how the authors train their HydroLSTM and LSTM in experiments 5 and 6. From what I've read, and understood from talks at conferences, LSTM models should be trained using all basin data, then tested at individual sites using either a PUB, PUR approach, or median NSE/KGE metric for all catchments. I do not believe the authors are doing this, thus, I am curious if training their HydroLSTM and LSTM models on all catchments would show the same results.

       Response: The approach mentioned is commonly used when the goal is to demonstrate the temporal or spatial consistency of a global-scale model. In our case, we are focused on predicting at a single catchment scale and aiming to determine the minimum complexity required to achieve similar performance, adhering to the principle of parsimony. Our findings indicate that incorporating contextual information in the gates simplifies the number of cell states and allows the weight patterns to encode hydrological knowledge. However, it is important to note that the behavior of total lag and weight patterns is specific to each catchment. Therefore, it becomes necessary to introduce appropriate regularization techniques to enable knowledge transfer between catchments. We are currently studying this step and plan to incorporate it in future applications of a global HydroLSTM representation.

4. Are the scientific methods and assumptions valid and clearly outlined?
    1. Yes. Table 1 does a good job of showing similarities between Storage and LSTM equations.
5. Are the results sufficient to support the interpretations and conclusions?
    1. I believe more work needs to be done to validate the conclusion that HydroLSTM provides comparable performance with LSTM, but with added interpretability. A PUR or PUB experiment to see how a HydroLSTM trained on all CAMELS basins performs would be appreciated.

       Response: Adding a comparison with LSTM in terms of PUR and PUB would indeed provide valuable information if the sole purpose were to evaluate performance. However, such a step would require further research into the regionalization of weights within the gates, which is currently under development. Therefore, we will modify the text to explicitly state that our conclusions are only applicable at a single catchment scale.

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?
    1. Almost. I still need clarification on the model training procedure.

Response: We will incorporate more details about the training procedure and provide the reasons for adopting this approach.

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution?
    1. Yes
8. Does the title clearly reflect the contents of the paper?
    1. Yes
9. Does the abstract provide a concise and complete summary?
    1. Yes
10. Is the overall presentation well-structured and clear?
    1. Yes
11. Is the language fluent and precise?
    1. Yes
12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?
    1. It would be appreciated to italicize all equations when in-line. It was hard to read/locate them amongst the text. There are also some repeated variable names (See the comments for an example).

        Response: The suggestion will be incorporated in the next version of the paper.
13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?
    1. The model training could be a little clearer (Similar to the above comment).

        Response: We will include additional details about the training procedure and elaborate on the reasons for adopting this specific approach in the next version of the paper.
14. Are the number and quality of references appropriate?
    1. Yes
15. Is the amount and quality of supplementary material appropriate?
    1. Yes

Major Comments:

- Can you italicize all in-line variables and equations? It’s hard to determine which parts of the text describe equations/LSTM properties. In some cases, I’ve had to reread a paragraph multiple times to search for an equation I missed.

  Response: The suggestion will be incorporated in the next version of the paper.

- (Lines 245) Are any static attributes used in model training?

  Response: Static attributes are not utilized in the current approach since each catchment is trained locally. Including static attributes would essentially introduce a biased term that is already accounted for in each gate. Hence, for the sake of consistency and avoiding redundancy, static attributes are not incorporated into the training process.

- (Lines 252-259) I suggest swapping Calibration, Selection, and Evaluation periods with the training, validation, and testing periods within the parentheses. It looks like you are using the train, validation, and test verbiage throughout the paper, and only referring to calibration, selection, and evaluation periods once (Line 427) after being defined.

  Response: We will review the paper using only the terminology of Calibration, Selection, and Evaluation.

- (Section 5.1) Would it be possible to include a PUR analysis rather than a 10-basin (PUB) holdout? So, rather than having two basins from each region, you would test on all gages within a snowmelt-dominant or Recent rainfall-dominant region. I believe this study would benefit from comparing how each LSTM performs on regions not included in the training set. This analysis would strengthen the claim that HydroLSTM has similar model performance to LSTM, but with heightened interpretability.

  Response: We acknowledge the value of PUR and PUB analysis in evaluating model performance. However, it is important to note that the current HydroLSTM representation is not designed to handle multiple catchments simultaneously. This limitation arises from the weight pattern within the gates. The lag and maximum weight values of one catchment cannot be directly transferred to another catchment solely through catchment attributes, as the attributes need to modify the weight pattern. Therefore, in order to transfer the knowledge learned from one catchment to another with similar catchment attributes, appropriate regularization techniques (such as regionalization) must be incorporated into the HydroLSTM architecture.

- (Line 310) How many total catchments were included in the training period? It is mentioned in Section 6.1, but not in 5.1. Is it just one catchment?

Response: Line 310 and Table 2 describes the 10 catchments used in the calibration (training) period. Experiment 1 (section 5) and Experiment 2 (section 6) are calibrated using one model per catchment but with a different total number of catchments in the analysis (10 and 588 catchments respectively).

- (Line 428) From my understanding of the literature, the best-performing LSTM models are using forcings, and attributes, from all basins in their inputs. For example, if there are 588 catchments, all catchments would be included in the training set. Then, testing would be done on all catchments, to determine a median KGE. Is training HydroLSTM on all catchments, or using basin attributes, something you have explored? Is the optimal lag memory hyperparameter the reason against having an entire CAMELS-trained LSTM? More explanation would be appreciated.

  Response: The method described is commonly used to assess the performance of different architectures, aiming to identify the best-performing option. However, in our study, our focus is primarily on the interpretability of what is learned by the architecture at each step. As a result, the analysis conducted at a single-level catchment can be considered an initial step before proceeding to train at a global scale.

  Based on this analysis, we have reached the conclusion that important hydrological properties, such as total lag memory and weight patterns, are encoded in the representation. This behavior is desirable for achieving interpretability. However, it also highlights a structural distinction in how meteorological forcing and catchment attributes should be processed.

  Due to this observed difference, we have not presented a global model in this particular paper.

- (Section 6) Is it possible to add a comparison against an LSTM applied to a large sample of catchments?

  Response: Given the current limitations of HydroLSTM, which can only be tested at a single catchment scale, while LSTM has already been tested at a global scale, a direct comparison between the two representations may not yield informative conclusions. However, we are currently working on incorporating a specific regularization technique that would enable the comparison of two global models.

- (Section 6) Is it possible to add a PUB comparison to this section?

  Response: As mentioned in the previous comments, a comparison under the current conditions is not possible.

Minor Comments:

- (Affiliations) The s in the United States is cut off

  Response: We will correct that mistype.

- (Line 58) Is Expected Gradient supposed to be capitalized?

  Response: The capitalization will be modified.

- (Line 107, Line 120) The symbol for the output gate, and the time-constant value, are both o. This could lead to some confusion.

  Response: We use the same letter to represent that both have the same behavior however we agree on extra differentiation should be added to avoid confusion.

- (Line 130-135) Physical state and informational state don't need to be italicized.

  Response: Those words were italicized because we considered them a simplified (even colloquial) characterization of the state. We will evaluate to change them in terms of the comments of other reviewers.

- (Table 1) Are the brackets supposed to be facing outward? (ex: *o =]0,1[*)

  Response: We used outward brackets to emphasize the asymptotic behavior of sigmoid and hyperbolic tangent functions. In the case of the linear reservoir, we can use inward brackets to show that they can take on zero or one value, despite being extreme cases of the linear reservoir.

- (Line 212) Typo. There needs to be a space inside Wand

  Response: We will fix this typo.

- (Line 253) I believe you mean "Commonly referred to as Training, Validation, and Testing." You used evaluation twice in this part.

  Response: We will fix this typo

- (Line 286) You didn't establish what a testing period is (see earlier comment for Line 253). Testing should be replaced with "Evaluation."

  Response: We will fix this typo

- (Line 304) The header “5 Experiment 1” reads weird. Maybe change to “5 First Experiment?”

  Response: We will replace the title with “First Experiment”

- (Figure 3) It may be clearer to the reader that rows are the Catchment Studied if you put the gage number on the row’s y-axis in bold above “Cells.”

  Response: We will modify Figure 3 following the suggestion.

- (Line 417) Same as the above comment. Maybe replace this with 6 Second Experiment. The section title reads weird.

  Response: We will replace the title with “Second Experiment”

- (Line 439) There is an unnecessary space before “However”

  Response: We will fix this typo.