Response to Reviewers for Manuscript: Hydrological Concept Formation inside Long Short-Term Memory (LSTM) networks

Comments/Text of Referee posted in **black**, our text in **purple**.

Reviewer 2: Anonymous

Lees et al. adapted a novel method used in Natural Language Processing, “probe”, to examine the internal function of the Long Short Term Memory (LSTM) model in rainfall-runoff predictions. Their results over 669 catchments in Great Britain show a good correlation between the LSTM internal states with re-analysis and independent soil moisture and snow cover products.

I agree with the authors that this paper could be a stepping stone to a myriad of interesting explorations in the field of hydrology. I also appreciate the author's effort in providing additional analysis in the appendices. However, I have some minor comments about some parts of the manuscript, mostly about the clarity and the tone toward traditional hydrologic models.

We thank the reviewer for their summary of the paper, and are glad that our message has been correctly interpreted! By addressing the reviewers suggestions, we have revised our manuscript, and we believe our manuscript has significantly improved

I feel the structure of the Introduction is a bit difficult and redundant for me to follow. I could not get the logical flow here. I found the main objective was stated in both the beginning and the end of the introduction. Why do we need a separate and long paragraph about interpreting machine learning from other fields? This paragraph disrupts my focus on LSTM interpretability.

We apologise for any confusion caused. We clarify our intentions below:

1) The objective outlined at the start of the introduction describes WHAT we hope to achieve. It gives the direction of the paper, describing the research as a hypothesis and a set of questions to answer. The objective outlined at the end of the introduction describes HOW we set about addressing that hypothesis and those questions.

2) The reason for the paragraph outlining the interpretation of ML in other fields is that we are aware that explainable AI and interpretable machine learning is a growing field with much to offer us in Hydrology. We ourselves are using a tool developed elsewhere and applying it to LSTMs trained for rainfall runoff modelling. We think it is important to honestly state that the presented methodology in this manuscript is not our invention and that it is a known technique in other fields of science. Acknowledging this fact, we think it

is important to give context about the state of ML interpretability outside of the field of hydrology.

I think the authors don't have to state that LSTM is the best rainfall-runoff model multiple times in the paper (Introduction and Conclusion). While this statement is still debatable, in my opinion, each rainfall-runoff model has its place in the modeling world. LSTM is increasing its popularity because of its robustness, computational efficiency and accuracy. Period. There is no need for bashing one over another.

We never intended to "bash" one model structure over the other, and have updated several sentences to avoid unfair criticism. We do share a pluralist view on the modelling landscape, where different models should be used for different tasks if they excel at them. Thus, we have tried to make our language as accurate and scientific as possible. To us, this also implies that the strong predictive performance of LSTM-based approaches for discharge modelling should not be played down, while also recognizing that other modelling approaches (such as physically based models) often provide very useful insights into the system being modelled and that system's dynamics. The purpose of this study was to bridge this gap and try to develop a methodology for providing a useful understanding of the LSTM-based approaches.

We believe the following four sentences, might be the cause of the reviewers concern:

1. Introduction: *"LSTMs have demonstrated state-of-the-art performance for rainfall-runoff modelling for a variety of locations and tasks [Frame et al., 2021a; Kratzert et al., 2018, 2019e; Lees et al., 2021b; Ma et al., 2020]*"
2. Conclusion: "*LSTM-based rainfall-runoff models offer good hydrological performance*"

We believe that these are important sentences because they motivate the manuscript, since there is a need to understand what these models are doing *because* of the fact that they offer highly accurate simulations.

3. L268: "*Firstly, the LSTM produces more accurate discharge simulations than any other hydrological model, and so we might expect that the intermediate variables are also better represented by the LSTM.,*"

We believe that this sentence is important because it outlines the logical steps in our thinking:

1) The LSTM is accurate at discharge simulation given meteorological inputs
2) Therefore, we expect the LSTM to have accurate intermediate representations of hydrological stores

In order to address the reviewer's concern, we will update the sentence to read:
"*Firstly, previous studies have demonstrated that the LSTM produces more accurate discharge simulations when benchmarked against other hydrological models, and so we might expect that the intermediate variables and processes are also better represented by the LSTM.*"

L282: "*Since the LSTM is often the best performing rainfall-runoff model for discharge (Kratzert et al., 2018, 2019c; Gauch et al., 2021; Frame et al., 2021; Gauch et al., 2020) it makes sense to explore the soil moisture that the LSTM associates with a simulated level of discharge.* "
This sentence adds another logical step to our argument outlined in the previous point:

3) Therefore, we can extract accurate soil moisture simulations that might be useful for other end users

In order to address the reviewer's concern, we will update this to the following:
"*Since the LSTM is often the most accurate rainfall-runoff model for discharge (Kratzert et al., 2018, 2019c; Gauch et al., 2021; Frame et al., 2021; Gauch et al., 2020) it makes sense to explore the soil moisture that the LSTM associates with a simulated level of discharge.* "

Section 2.3 ERA5-Land Data: there is an imbalance between the descriptions of soil moisture and snow depth. I would expect to see more information about snow depth and its accuracy over GB.
We will update this section with a more complete discussion of the snow depth variable and its usefulness over GB, which now includes the following sentences:

"ERA5-Land data has demonstrated improved representation of snow depth, in part due to the increased spatial resolution which adds value in complex mountainous terrain due to improved representation of the orography, and therefore, a better representation of the surface air temperature (Munoz 2021). To our knowledge there are no direct verifications of the ERA5-Land snow product over the area in Northern Scotland where our snow-depth experiments are conducted. However, there is evidence to suggest that in mid-altitude ranges, the improved spatial resolution of ERA5-Land is a dominant factor in improving snow depth estimates, at least over the ERA5 product (Munoz 2021)."

Figure 2: no y label
We will update this as proposed. (Thanks!)

Figure 5: no y label
We will update this as proposed.

Line 249: I thought there are only two meteorological drivers (temperature and precipitation (line 6))?
Apologies, this is the model version with potential evapotranspiration, it's the same model setup as the 2021 benchmarking study (Lees et al 2021). We will update the associated lines describing model inputs.

Line 268: See the second opinion
"*Firstly, the LSTM produces more accurate discharge simulations than **any other** hydrological model, and so we might expect that the intermediate variables are also better represented by the LSTM.,*"
We believe that this sentence is important because it outlines the logical steps in our thinking:

4) The LSTM is accurate at discharge simulation given meteorological inputs

5) Therefore, we expect the LSTM to have accurate intermediate representations of hydrological stores

We will update the sentence to read:
"*Firstly, previous studies have demonstrated that the LSTM produces more accurate discharge simulations when benchmarked against other hydrological models, and so we might expect that the intermediate variables are also better represented by the LSTM.*"

Line 282: See the second opinion
"*Since the LSTM is often the best performing rainfall-runoff model for discharge (Kratzert et al., 2018, 2019c; Gauch et al., 2021; Frame et al., 2021; Gauch et al., 2020) it makes sense to explore the soil moisture that the LSTM associates with a simulated level of discharge.* "
This sentence adds another logical step to our argument outlined in the previous point:

6) Therefore, we can extract accurate soil moisture simulations that might be useful for other end users

We will update this to the following:
"*Since the LSTM is often the most accurate rainfall-runoff model for discharge (Kratzert et al., 2018, 2019c; Gauch et al., 2021; Frame et al., 2021; Gauch et al., 2020) it makes sense to explore the soil moisture that the LSTM associates with a simulated level of discharge.* "

Line 319: I found a recent paper (Tran et al. Development of a Deep Learning Emulator for a Distributed Groundwater–Surface Water Model: ParFlow-ML. Water. 2021) in which the spatial information is included in the LSTM architecture. Do the authors think the probing technique could be used in this architecture? Can the probing technique map between predicted and observed spatially-distributed soil moisture?

This is a very interesting point, thank you for sharing this paper with us.

The analysis presented in this manuscript has the potential to apply to any matrix that can be mapped to a vector. The approach would be directly applicable if the size of the cell state is the same as the target variable (i.e. if there is a 20x20 pixel grid of soil moisture estimates, the cell state size would need to be a 20x20x{hidden_size} pixel grid of cell-state estimates). Then you can imagine mapping that 20x20x{hs} tensor to the 20x20 pixel grid of soil moisture data to capture the amount of information stored in all of the cell state. We are therefore using the regression to collapse the final dimension (hs) into the 20x20 grid of soil moisture estimates. This is a specific example describing how the method would work so long as you can imagine using a linear regression with the input features mapping to a target variable of interest.

In future work we intend to explore spatially explicit LSTM architectures, like the one you have shown here, and extending our method to these approaches is certainly an interesting area for future work.