Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-540-RC2, 2021

Hydrology and Earth System Sciences
Discussions
Open Access

Interactive comment on "Rainfall–Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network" *by* Martin Gauch et al.

Thomas Lees (Referee)

thomas.lees@chch.ox.ac.uk

Received and published: 3 January 2021

1 OVERVIEW

This paper seeks to answer the question: "Can a single LSTM model be used to produce accurate and consistent discharge simulations at daily timescales and sub-daily timescales?".

The major finding was that yes, you can use a single LSTM to produce daily and hourly predictions. Furthermore, compared with more traditional hydrological models, the MTS-LSTM shows a much smaller performance deterioration when comparing daily simulations (better) to hourly simulations (only slightly worse).

The novel contributions of this paper are threefold:

1. The development of a new multi-timescale LSTM (MTS-LSTM) that produces discharge simulations at both daily and sub-daily timescales (including the flexibility to include arbitrary timescales).

2. The manipulation of the loss function to explicitly account for prior knowledge about the translation between daily and sub-daily timescales. Related to the "hierarchical" nature of these timescales.

3. The benchmarking of a suite of LSTM-based models against the operationally used NOAA National Water Model (NWM).

Research into LSTM based rainfall-runoff modelling has, thus far, mainly focused on simulations at daily timescales. This paper provides a welcome addition to the literature, since sub-daily trends can be important for flood impacts and for water resource managers. The authors focus on producing discharge simulations at a daily timescale and an hourly timescale, although they also show results for 3-hourly and 6-hourly timescales (see Table 2 p9, Table 7 p16).

In order to explore the LSTM architectures that can produce discharge simulations at multiple timescales, the authors suggest three possible avenues (more are included in Appendix B):

- Multiple LSTMs with different timescales, an hourly LSTM and a daily LSTM (naive).

- A "shared" multiple-timescale LSTM (sMTS-LSTM), which overcomes the problems of overly long input sequences, causing long training and inference times for the naive model.

- The MTS-LSTM, which overcomes the problems of the sMTS-LSTM being unable to include different input data for the different timescales.

Both the sMTS-LSTM and the MTS-LSTM are novel contributions to both hydrological modelling, and as far as I am aware, machine learning more generally. My main comment about the paper is that the difference between the sMTS-LSTM and the MTS-LSTM could be made clearer.

The authors describe four experiments to demonstrate the usefulness of their newly developed models:

1. Benchmark the MTS models (sMTS-LSTM & MTS-LSTM) against traditional hydrological models (NOAA NWM) and the naive LSTM (which although "naive" is still the most difficult benchmark to compete with). This comparison is thorough and explores accuracy across the hydrograph (see Table 4, Figure 3, Figure 4).
2. Explore the consistency of the MTS models hourly discharge predictions when aggregated to the models daily discharge predictions. The regularisation of the loss function improved the consistences of the sMTS-LSTM
3. Compare the computational efficiency of the 3 LSTM-based models. The MTS-LSTM was the most computationally efficient.
4. Test whether including the same information from different timescales improves model accuracy. The extra information improved forecast accuracy over a range of performance metrics (Table 6).

Overall, these experiments are well thought through and they meet the aims of HESS. The research advances hydrological modelling by:

- benchmarking data-driven models (LSTMs) on an hourly timescale

- developing novel model architectures that show state-of-the-art performance
- demonstrate a next step for LSTM-based models to be used in operational forecasting settings
- demonstrate the flexibility of manipulating the loss function in data-driven models to meet different requirements (e.g. timescale consistency).

Furthermore, the availability of the code via the neuralhydrology repository, with an accompanying notebook makes it possible to view the authors assumptions and reproduce the figures in the paper.

2 SPECIFIC COMMENTS

I was grateful for the following:

- Figure 2 (P7) is extremely helpful and very professionally made. This is extremely helpful when trying to parse the novel model architecture (MTS-LSTM) proposed by the authors.
- The overt structure outlined on P3 L64-78 is a very helpful signpost to the reader.
- The regularization used to ensure timescale consistency (Sect 2.3.2) is novel and interesting for the target audience of HESS, hydrologists and earth scientists. It confirms the view that the loss function offers huge flexibility to modellers to improve their models for specific use-cases.
- Equation 1 (P8 L180), the annotations to this equation are extremely helpful.

- Table 3 and Table 4 demonstrate an extremely thorough comparison of the models for various metrics and hydrological signatures. This could be used as an example for future benchmarking experiments as an extremely thorough inter-comparison, exploring the various facets of the hydrograph.

- Appendix B is a very worthwhile addition, since these negative results can help the field from repeating these results, especially because they turned out to work less well than the model architectures included in the main text. It also outlines the thoroughness of the authors experiments.

- The inclusion of the data and a Jupyter Notebook for readers to reproduce the results is to be applauded. The notebook is well written and the community will be grateful for the time and effort that the authors have put into making their code available and their experiments reproducible. Thank you.

I have a series of comments:

- **P3 L80-87** Are you still using the CAMELS observed discharge or do you now exclusively use the USGS Water Information System REST API values for both hourly and daily evaluation?

- **P3 L81-82** Just to confirm, this is still a “predict timestep including all input data up to time t” rather than a forecast. This is confirmed on P17 L306 but might be worth also including that information here.

- **P4 L101-104** In Section 2.2.1 you describe that you use the NWM v2 Reanalysis product. You describe that this is an hourly product. Do you therefore calculate a daily average of these results to compare against the daily simulations?

- **P6 L131-155** I am still not fully clear on the difference between the sMTS-LSTM and the MTS-LSTM. Can we work to make this slightly clearer in Section 2.3.

- Do the sMTS-LSTM and the MTS-LSTM receive the same input data?
- Do both the sMTS-LSTM and MTS-LSTM require two forward passes (L140)
- It seems that the MTS-LSTM "splits the LSTM into two branches" (L148), which is described as unique to the MTS-LSTM, but then Figure 2 suggests that the sMTS-LSTM also does this splitting but the fully connected layers (FC_c, FC_h) are simply identity functions.
- Does the one hot encoding (L141) mean that the LSTM weights are copied in both branches but then zeroed if we are looking at either the hourly or the daily data? If so then why can we not use different input datasets in the sMTS-LSTM as we can in the MTS-LSTM?

There are various solutions. One could: include a table explaining the differences explicitly; include the sMTS-LSTM as its own diagram in Figure 2; or spend more time in Section 2.3 clearly outlining the differences between the two architectures.

- **P6 L154-156** "This architecture makes it clear why we call the other variant "shared" MTS-LSTM: Effectively, the sMTS-LSTM is an ablation of the MTS-LSTM. Both variants have the same architecture, but the weights of the sMTS-LSTM are shared across all per-timescale branches and its state transfer layers are identity operations." I am not what this sentence means. I think it could potentially be clearer for a hydrological audience. My understanding is that an "ablation" means that the sMTS-LSTM is missing something that the MTS-LSTM has, but if they have the same architecture then I am not certain what is missing? From reading and re-reading the difference is something to do with the fully connected layer but I am just a little bit confused about the difference between these two models.

- **P7 L158-169** Related to the misunderstanding of the difference between the sMTS-LSTM and the MTS-LSTM, I am not certain what it means to include multiple datasets and why this could not be done for the sMTS-LTSM. I know that in the paper: "A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff modeling", some of the authors have shown that the LSTM produces more accurate discharge simulations with multiple sources of rainfall information. Is that what is being done in this experiment? Furthermore, if the sMTS-LSTM has the same architecture as the MTS-LSTM (as outlined in the caption to Figure 2), then why can't the sMTS also include new information to the hourly branch? (I am assuming here that the only difference is that the FC_h and FC_c are identity functions rather than linear functions as in the MTS).

- **P7** Related to the comment above. "In the other, we additionally ingested the corresponding day's Daymet and Maurer forcings at each hour. ". Is this data at a daily resolution? If so, does this mean that you are copying the daily inputs 24 times as input for each hour? So if we have hourly NLDAS, You are including Daymet for Day 1 24 times? NLDAS1 + Daymet 1, . . ., NLDAS24 + Daymet1. Apologies if I have misunderstood.

- **P7 L167** "... In the other, we additionally ..." I think it would make sense to explicitly write that you are using the NLDAS forcings AND the Daymet/Maurer forcings. Perhaps something like: "In the other, we ingest the NLDAS forcings as well as the corresponding day's Daymet ..."

- **P13 Table 4**: You write in the Table caption "Bold values highlight results that are not significantly different from the best model in the respective metric or signature (α = 0.001)". I am sure I have misunderstood, but when I look at the Hydrologic Signatures, for example Daily, Q mean. Both the Naive (0.986) and NWM (0.972) results are highlighted. However, Both the sMTS-LSTM (0.985) and the MTS-LSTM (0.984) have values closer to the best model. Is this an artefact of the aggregation? Where the mean is hiding the distribution of Pearson Correlation scores across multiple basins? If so that is fine I just wanted to ensure that this was not a mistake.

- **P14 Table 5** Why do we only see results for the sMTS-LSTM. I believe you have written that it is the “best benchmark model”, but is there any other reason to include/exclude the MTS-LSTM? If the experiment was already run it might be an idea to include it, but it is not necessary.

3 FORMATTING

I am not certain of the procedure here but I am drawing to your attention in case it is useful.

- **P2 L33**: "... (e.g., Schmidhuber (1991), Mozer (1991))." to "... (e.g., Schmidhuber 1991, Mozer 1991)"
- **P8 L173-174**: "... (e.g., computer vision, Zamir et al. (2020))" to "... (e.g., computer vision, Zamir et al. 2020)"

4 SUGGESTIONS

I believe that you are using the terms "look-back window" and "input sequence" interchangeably. Is it perhaps worth using one term consistently through the paper?

- P5 L115 "... look-back windows of 365 days ..."
- P6 L128 "... input sequence of 4320 hours (180 days) ..."
- P6 L137 "... input sequence of T_D time-steps ..."
- P6 L143 "... has access to a large look-back window ..."

- P8 L190 "... achieve a sufficiently long look-back window ..."

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-540, 2020.