Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-540-RC1, 2020

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.

Jens Kiesel (Referee)

kiesel@igb-berlin.de

Received and published: 14 December 2020

GENERAL COMMENTS

The manuscript "Rainfall–Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network" (LSTM) by Martin Gauch et al. presents an extension of LSTM hydrological models to sub-daily time steps. In previous publications, LSTMs as hydrological models were used on a daily time step. The authors explore multiple approaches to achieve a ’multi-timescale’ model, of which three (naive LSTM, sMTS-LSTM, MTS-LSTM) are evaluated in more detail and less promising experiments are briefly explained in an Annexe. Similar to previous applications of LSTMs, the models

Printer-friendly version

Discussion paper

are applied at the CAMELS dataset, encompassing 516 basins across the contiguous USA where hourly data is available. Results are compared to the NOAA National Water Model (NWM) and show that all LSTMs architectures outperform the NWM. The authors suggest that the MTS-LSTM provides most flexbility for future use.

The manuscript is generally well written and structured, figures and tables support the results. Having a more process-based hydrological background, I nevertheless read the paper with interest and believe it fits well in the scope of HESS. I see the work as highly relevant, especially in the field of flood modelling and (eventually) forecasting, but also generally in the application of LSTMs at different temporal resolutions. However, especially regarding the latter, I think the authors should invest more work to improve the usefulness of the paper. Please find below more detailed comments, questions and suggestions that hopefully initiate a fruitful discussion and help in improving the paper.

SPECIFIC COMMENTS

ABSTRACT: I suggest to mention the difficulties and challenges applying the models (parameter estimation) and discuss the work still to be done regarding different time scales (e.g. generalization of parameters)

INTRODUCTION I think you are missing a research gap in your introduction which is important to apply the LSTM for different time steps, since there seems to be a time-step dependency of model parameters / hyperparameters (e.g. hidden size, sequence length, batch size, forget gate bias, learning rate?, others?). Due to the computationally expensive training of LSTMs, knowing which ones need to be adjusted, in about which range and identifying ideal values is essential. I would like to see this topic included in the "contributions" you list at the end of the introduction (and therefore also more prominently in the respective chapters).

p.2 l.29-41: I think this section is difficult to understand for a reader without firm neural networks background. Particularly phrases like: "partitions a recurrent neural network into layers with individual clock speeds", "process irregularly sampled inputs by means of a time gate that only attends to the input at steps of a learned frequency", "the approach depends on a binary decision that is only differentiable through a workaround". I acknowledge that your paper cannot serve as an introduction to the topic. I have no clear suggestion other than making this paragraph more accessible to readers with a hydrological background through using less specialized jargon, if possible.

p.2 l.45 and 47: You write that Araya et al predicted wind speed at "multiple timescales". Then you mention that your objective is "multiple outputs, one for each target timescale". I don't understand the difference between that.

p.2 l.54ff: I see the capability to process input data in irregular intervals as an advantage. Think of satellite products that have different data gap length (e.g. soil moisture or altimetry products combining multiple sensors). You can discuss this further, but at least I suggest to write on p.3 l.77: "...LSTM can ingest individual and multiple sets of forcings each having regular time intervals for each target timescale. This closely resembles..."

p.3 l.70-72, 74-75: I suggest not to mention the results of your study in the introduction

p.3 l.64-78: These three paragraphs reveal that your introduction could be structured a bit better, ideally introducing the reader to these three problems/research gaps that need to be solved for "Rainfall–Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network". You have motivated the first paragraph, but the second and third 'contribution' that you list appears a bit unexpected since your previous introduction does not resemble that structure. For instance, instead of refering to sections later in the paper, I believe it would be better to introduce the reader to the problem of inconsistencies. You briefly mention this on p.2 l.27-28 for conventional hydrological models, but this can be extended, especially targeted on machine learning.

DATA AND METHODS

p.4 l.92-94: The distinction into training, validation and test is not fully clear to me. You use the validation period to evaluate different architectures and to select model hyperparameters. Could you elaborate on the reason why the evaluation of architecture and the hyperparameter selection cannot/should not be done during the training period?

p.4 l.101ff: Can you describe the datasets used in NWM and the basic characteristics (e.g. spatial application range, calibration strategy and performance) of the v2 reanalysis product?

p.5 Fig1: please also mention what "x" and "+" represent

p.6 l.127-130: I am particularly interested in how you tuned these parameters and how you decided which parameters to adjust and which ones not. As you mention, the LSTM application is computationally expensive and parameter selection and ranges are therefore important. Therefore, I would rather want to see Appendix D in the main text, and include information why certain parameters are time step dependent and others not. Also, In Table D1, it seems you ended up with 336 hrs sequence length for both architectures. Would an even longer sequence length lead to better results? What is the tradeoff between higher sequence lengths and computational costs?

p.6 l. 146-156: Could you explain why these two different LSTM architectures were developed? What are the expected advantages/disadvantages? The last sentence is crucial for the undestanding of the differences, I believe "weights of the sMTS-LSTM are shared across all per-timescale branches and its state transfer layers are identity operations." What is an identity operation?

p7. Figure 2: I understood from the text that both the sMTS-LSTM amd MTS-LSTM are branching out at each day into hourly predictions. The MTS-LSTM predicts 24 hours, using 72hrs sequence length. Is this the same for the sMTS-LSTM? The difference between sMTS-LSTM and MTS-LSTM is difficult to understand from just the figure caption. I think it would help to construct the illustration for both architectures to visualize the differences, if possible including the different weights for the MTS-LSTM and the similar weights for the sMTS-LSTM in the diagram.

p.7 l.158: I don't understand why the MTS-LSTM is more flexible in terms of input data than the sMTS-LSTM. In the sMTS-LSTM section you write (p.6 l.139): "we....ingest the hourly input sequence of length TH to generate 24 hourly predictions that correspond to the last daily prediction." Looking at Fig 2, to me this is similar in the MTS-LSTM, where the daily forcings have an effect until the hourly branch starts and then no update using the daily forcings/predictions seems to be made in the hourly branch. Therefore, effectively, you use the daily data until the model branches out and then you use the hourly forcings only? Again, I think it would help to show both architectures in Fig 2.

p.8 l.170-184: If I understand it correctly, adding the term into the loss function 'encourages' the model to minimize the difference between daily and sub-daily simulation. But similar to the NSE, this ideal value may not be reached, ending up with a model that is not consistent - even if you put an exceptionally high weight on the mean squared difference? Is there a reason why you don't 'force' consistency across timescales? E.g. when looking at Figure 2 I imagine you could add a function (e.g. simple multiplication of a term) that scales either the daily or the sub-daily prediction (or the average between the two) so that both match the consistency criteria (I now notice that may be similar to what you did in "B1 Delta Prediction")?

p.9 Table 2: it is a bit confusing to have these different sequence lengths. In the previous section it is 72hrs, here 168hrs, in Table D1 it is 336hrs. Can you harmonize this or explain why there are these differences?

RESULTS:

p.9 l.210: that means running ten seeds based on the parameterization in bold in Table D1? If so, I'd add this here

p.9 l.219: I find this particularly interesting when thinking about hydrological processes. The model parameter values (hidden and cell states) of the last coarse time step (Td - Th/24) are basically your boundary condition/initial state for the hourly model. It seems a bit counterintuitive that the sMTS-LSTM performs better than the naively trained full hourly LSTM. So the ’error’ you introduce through the daily average initial state must be insignificant (due to a sufficiently long sequence length?). Particularly in small basins and for flood peak prediction, this may not always be the case. A plot showing the spatial differences in performance between the naively trained LSTM, the sMTS-LSTM and MTS-LSTM (e.g. similar to Fig 4) could reveal if/where these differences exist. I’d however not be surprised if this plot will show no pattern due to input data uncertainty and randomness in the LSTM and the small performance difference between the LSTM types.

p.14 l.237-250: Interestingly, the Naive LSTM deviates most - probably because the sMTS-LSTM and the MTS-LSTM use recent states from the daily model and are therefore ’closer’ to the daily models flow (volume) prediction? The beneficial influence on the NSE could arise because you are introducing a ’physically plausible’ constraint in the model which ’helps’ adapting the network to the processes? (see also my comment to p8. l.170-184). That is an interesting prospect and if true, could mean adding more of such physical constraints (e.g. global water balance closure) could improve the LSTM even further?

CONCLUSIONS:

p.16 l.292: it depends on how the NWM was calibrated and what the main purpose is (see also comment to p.4 l.101ff)

p.17 l.293: I understand and agree. But given that LSTMs perform so well for hydrological modelling, efforts should be made to generalize the hyperparameter values for different time steps. I believe you were not sufficiently confident with your tests to deduce general rules for the hyperparameter settings (and that may be a reason why this analysis ended up in the Annexe). But I think it would help the future application of LSTMs if you could give a summary of your experience: e.g. which parameters are time-step dependent, should a parameter increase or decrease with increasing/decreasing time steps, what if someone applies an even coarser time step (monthly)?

p.17 l.296-298: I know the differences are not statistically significant, but can you speculate on why the models are ranked in that order? Somehow the naive hourly LSTM seems not to be able to use this additional information content, or the half year sequence length is not sufficent to depict all states (e.g. groundwater storages may need longer sequence length in some catchments)?

p.17 l.299-305: Can you speculate why the daily forcings to the hourly MTS-LSTM improve the performance?

I believe there is more research to be done that you can mention here? E.g. a thorough investigation of time step-dependency of hyperparameters, find measures to use physical constraints in the LSTM (e.g. the regularization)

TECHNICAL CORRECTIONS

once introduced, you can stick to the abbreviations (e.g. NWM, MTS-LSTM)

p.1 l.14: LSTMs can predict hydrological processes in multiple...

p.3 l.58-60: I think you can refer to Appendix C here

p.5 l.118: ...half a year...

p.8 l.191-192: it is uncommon to mention results in the methods

p.8 l.199: this link is supplied here for the third time. Not sure if this is how HESS wants to have references to URLs.

p.9 l.215: 'even the naive ones' - the naive LSTM acts as a benchmark, so it is expected it performs better than (s)MTS?

p.9 l.216: I think it is fair to add that this worse performance on hourly is much more visible at the NWM

p.14 l.255: which parameterization and number of basins is meant here? I can't imagine you mean all basins, 10 seeds, 30 epochs?

p.17 l.311-312: I find this first sentence difficult to understand. If possible, split in two

p.18 l.334: I like the documentation of the failed approaches and where appropriate, I suggest to reference these in the main text

---

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