Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC3
https://doi.org/10.5194/hess-2021-511-AC3, 2022

Responses to RC3 — addressed to Editor

Reyhaneh Hashemi et al.

Author comment on "How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models?" by Reyhaneh Hashemi et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-511-AC3, 2022

Dear Editor,

We believe that in the document provided in AC2 we have taken seriously into account every single point brought up in the first review (RC2) of Anonymous Referee #2 (R2).

If the editor finds it necessary, we will of course proceed to conduct further tests on hyperparameter tuning without having any problems. However, the elements given by R2 in their reviews do not identify what **exact tests** they expected and their attachment has left us perplexed. The latest literature admits also that hyperparameter tuning could be an endless job as, for instance, Klotz et al. (2021, under review) looks also for a "balance between computational resources and search depth".

Considering the hidden unit size hyperparameter, we took the choice concluded in Lees et al. (2021) since they had reported that they had systematically tested larger values for their entire sample — which was in a region close to ours — and they did not observe any performance improvement either.

In what follows, we elaborate on two points that we find important:

1. How hyperparameter tuning of LSTM networks is carried out in some recent related studies.

2. Our local training methodology and why our results are not in contradiction with what has been previously done in those studies.

We hope this will help you in making your decision.

Finally, we would like to thank you for taking charge of our paper as well as both referees for their time and reviews.

Best regards,

Authors of hess-2021-511

---

1) LSTM hyperparameter tuning in similar studies

We fully agree with R2 on the importance of hyperparameter tuning. We carefully read their comments but did not find an objective or — their — definition of the standards to achieve. We also carefully read the papers of the authors cited by R2. It appears to us that level of hyperparameter tuning is not uniform between studies and varies from study to study:

--Kratzert et al. (2018) report doing a "manual" hyperparameter tuning on "several" catchments from a region — Austria — different and far from the study region — the United States (US): "The specific design of the network architecture, i.e., the number of layers, cell/hidden state length, dropout rate and input sequence length were found through a number of experiments in several seasonal-influenced catchments in Austria". From this hyperparameter tuning — **conducted using a local LSTM and for catchments in Austria** — they conclude a 2 layer structure with 20 hidden units. They then **use these choices in their local — and regional — LSTMs** and **for 241 catchments located in the US** (CAMELS data set), "without further tuning" even if they acknowledge that it "is something to do in the future".

R2 highlights in their first review that "there is strong relationship between the dimension of the cell state and the sequence length". However, in this study (Kratzert et al. (2018)), the length of the input sequence — called lookback in our paper and AC2 — is not varied and it is fixed to 365 days "in order to capture at least the dynamics of a full annual cycle".

It turns out that, in this study, hyperparameter tuning is not performed in the same manner as required by R2.

-- Kratzert et al. (2019) performed a more elaborate hyperparameter tuning, described briefly in Annex A of their paper, but without presenting any detailed results. They considered the following variations for the following hyperparameters: "Hidden states: 64, 96, 128, 156, 196, 224, 256; Dropout rate: 0.0, 0.25, 0.4, 0.5; Length of input sequence: 90, 180, 270, 365; Number of stacked LSTM layer: 1, 2". No batch size variations are reported.

They finally chose a one layer structure, with 256 hidden states, a dropout rate of 0.4 and a length of input sequence of 270 [days]. This hyperparameter tuning is performed using only one performance metric and for only one regional LSTM (no local training in this

study). However, they then used "the same architecture (apart from the inclusion of a static input gate in the EA-LSTM), which found through hyperparameter optimization" to compare 3 different regional LSTM models for 2 different performance metrics.

It turns out that, in this study, hyperparameter tuning is not, either, done in the "for-each-model" fasion that R2 requires.

-- Gauch et al. (2021) conducted even a more complex hyperparameter tuning. In this recent study, three periods are defined, instead of the only two in previous studies (but with a k-fold cross validation using data of the first period). As mentioned by the authors, splitting the data into three periods (calibration, validation, test) is "a widespread calibration strategy for DL models". Gauch et al. (2021) compared four LSTM type models — Naive_daily, Naive_hourly, sMTS-LSTM and MTS-LSTM. It is crystal clear that the hyperparameters they tuned differ from model to model. Also, the sequence length and hidden unit size parameters are not varied in hyperparameter tuning of their Naive models and are set to a fixed value — contrary to what R2 believes about their strong inter connectedness. Please also note that Kratzert et al. (2019) had previously found a different (270 [days]) optimal value for lookback and although it was obtained in a different setting, the evidence for the importance of its variation was present for Gauch et al. (2021).

It turns out that, in this study, hyperparameter tuning is not, either, done in the "equivalently-for-each-model" fasion and for all hyperparameters that R2 requires.

Furthermore, for some unmentioned reason, the number of studied catchments in their study is not the same as previous studies of Kratzert, although the authors also used the same US CAMELS data set.

-- Klotz et al. (2021, under review) used three periods: "training, validation, and testing that are standard in the machine learning community". However, for some unmentioned reason, they did not choose the same dates and catchments as those taken in Gauch et al. (2021), although they had also used the US CAMELS data set. Apparently, the length of the input sequence (lookback) is not varied — this is not mentioned, but this is at least what one would understand from the preprint paper and the discussion available on HESSD on 2022-01-24.

Hyperparameter tuning is performed for 6 parameters (hidden states, number of components, noise, dropout, batch size and learning rate) and for four models (GMM, CMAL, UMAL and MCD). One LSTM model is added for comparison, but taking the hyperparameters obtained by Kratzert et al. (2019): "We, therefore, also compare a model with the same hyper-parameters as Kratzert et al. (2019), the latter model is labeled LSTMp".

It turns out that, in this study, hyperparameter tuning is not, either, done in the "equivalently-for-each-model" fasion that R2 requires.

Furthermore, the train/validation period of Kratzert et al. (2019) — 1999-10-01 to 2008-09-30 — overlaps the test period of Klotz et al. (2021, under review) — 1995-10-01 to 2005-09-01. There might be some unmentioned reason, but Deep Learning guidelines (Goodfellow et al., 2016) require choosing independent periods.

2) Local (versus regional) training

In the cited studies, authors seem to consider that local training is not useful and regional training using static attributes brings much better performances. According to their comments, this seems to be the opinion of R2 as well.

We would therefore like to highlight two points:

1- Contrary to what seems to be the prevailing view, it turns out that **local LSTMs have never been compared to regional LSTMs WITH static attributes** — probably, apart from the «unpublished» results mentioned by R2 in their first review. Indeed, Kratzert et al. (2018) compared local LSTM, regional LSTM WITHOUT static attributes and a third approach: "fine tuning the regional model for each catchment". The two first approaches

gave similar results, while the last approach clearly improved performance. Results of Kratzert et al. (2018) were obtained on a "subset" of the CAMEL set. Then, when moving to the entire US CAMEL set for their following studies (Kratzert et al. (2019), Gauch et al. (2021), Klotz et al. (2021, under review)), the authors abandoned both local training and fine tuning methods, focusing only on improving their regional LSTM models.

2- We tried to explain in AC2 that we used a different methodology for training our LSTM models. We used three independent sufficiently long intervals — training, validation, test. This made us possible to apply an early stopping criterion, for each catchment, individually. Therefore, **in our study, the number of epochs used to locally train LSTM differs from catchment to catchment**, depending on the loss obtained on the validation period.

Kratzert et al. (2018) used a different approach. They had two periods — one for "training-validation", the other to test their model. They carried out a preliminary test in which their "training-validation" period is divided into two parts: 14 years for local training, 1 year for validation. Based on the mean NSE calculated on the validation period, **they chose the same number of epochs for all catchments** and re-train the model locally on the whole "training-validation" period. This approach is fully understandable since the authors argue that "the goal of this study is therefore not to find the best per-catchment model, but rather to investigate the general potential of LSTMs for the task of rainfall–runoff modelling". Nevertheless, **this approach clearly penalizes local LSTMs**.

We must state that we find all cited peer reviewed papers excellent and we fully agree with the choices made by their authors. We only intended to stress that, in our opinion, the hyperparameter tuning tests presented in the preprint version of our paper are not in contradiction with what is reported in the existing literature of this space, contrary to what R2 suggests.

References

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff

prediction at multiple timescales with a single Long Short-Term Memory network, Hydrol. Earth Syst. Sci., 25, 2045–2062, https://doi.org/10.5194/hess-25-2045-2021, 2021.

Goodfellow, I., Bengio, Y., and Courville, A: Deep Learning, MIT, Press, available at: http://www.deeplearningbook.org, 2016.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty Estimation with Deep Learning for Rainfall–Runoff Modelling, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2021-154, in review, 2021.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, Hydrol. Earth Syst. Sci., 25, 5517–5534, https://doi.org/10.5194/hess-25-5517-2021, 2021.