Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2020-670-RC2, 2021

Comment on hess-2020-670

Adrien Michel (Referee)

Referee comment on "Machine-learning methods for stream water temperature prediction" by Moritz Feigl et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-670-RC2, 2021

Review of « Machine learning methods for stream water temperature prediction

Dear authors, dear editor,

The paper submitted discusses the usage of machine learning (ML) models to simulate water temperature on 10 catchments located in Austria. The results obtained from the ML models are compared to a linear regression model and to the model *air2stream*. The authors show that ML models achieve better results for water temperature simulations than the 2 benchmark models used. They also show that the choice of the hyperparameters of the ML models has an important role in the performance of the models, and they present a method to reduce the computational time required to optimize the values of these parameters. Finally the choice of the input variables forcing the ML models and its impact on the model's performances is discussed. This work is part of a recent ongoing effort to apply ML models in hydrology and it brings some new interesting insights. The comparison with two benchmark models really allows to correctly asses the performances of the ML models.

In general, the paper is clear and well written and shows clean figures. In addition, all the source code is provided with clear instructions and in-code documentation. As I detail below, there are some important points to be addressed regarding the amount of information provided in order to allow reproducibility, the clarity of some parts of the methods, and finally, the possible application discussed (short terms prediction and climate change impact studies). I have no doubt that these points can be clarified and/or enhanced by the authors and that a reviewed version will fit for a publication in HESS. Indeed, water temperatures has historically received less attention than discharge in modelling, while becoming a more and more important variable with ongoing and foreseen climate change. The contribution brought by this article is therefor really valuable.

I have to mention here that my expertise is on the water temperature side and not on the machine learning side, and that the editor might want a review by an expert from the ML community.

Do no hesitate to contact me for further discussions.

Best regards,

Adrien Michel (adrien.michel@epfl.ch)

Major comments:

Impact of snow/glacier cover and catchment size

The authors discuss the importance of snow/glacier cover in the perspective of climate change (CC) application (I discuss CC below). However, this is not discussed for the application done in the paper. First, I suggest to add in Table 1 the mean catchment elevation and percentage of glacier cover in the catchments in order to allow a quick overview of the contribution of glacier and snow melt we can expect on each catchment. I would expect the TQ experiments to perform significantly better than the TP ones in catchment where snow plays an important role. Indeed, snow melt dynamic is captured in Q while I doubt TP experiments will be able to get it. This is difficult to see from Fig 4c and A1, so I would suggest to add further information about TP vs TQ comparison in high elevation catchment. Especially since the authors mention TP performance as an argument for usage in CC impact studies (lines 532 to 537).

Catchment's size seems to have a clear influence on the results. Indeed, if we neglect the Danube catchment, Figures 4c and A1 show a reduction of RMSE and MAE with increasing catchment's size. This is not surprising since I would expect local scale effects, harder to capture in models, to be smoothed out when increasing catchments size, leading to an increase of the model performance. It could be interesting to replace catchment with catchment size in the linear regression for test RMSE, or use both, in order to really asses it (in any case this would mean a regression with both discrete and continuous variables). This size effect is currently not discussed (except for the Danube) while I think it should definitely be mentioned.

ML models details

As mentioned above, I have no expertise in the ML domain beyond basic knowledge. As a novice, I found the Section 2.4 quite technical (especially Section 2.4.5). Having in mind the target audience of HESS, I would suggest to keep in the main text an overview of the different ML models used, and to move the most technical parts in Supplementary Material along with the details requested below.

This would allow the authors to present more details about the reproducibility of the study which are not yet presented. Indeed, Appendix A only shows the hyperparameters bounds used for the Bayesian optimization. The final set of hyperparameters should be provided (along with the parameters of the two benchmark models). It is not completely clear for me if the Bayesian optimization is done in general or per catchment. This should be stated. Also, is the optimization run for each separate experiment, or only once? And in this case for which experiments? If the Bayesian optimization is done per catchment separately, is not it a risk of overfitting? In summary, some details and clarity are missing about how the Bayesian optimization is done (which catchments, experiments and time periods), and how the models' training is done. Note that the calibration procedure of *air2stream* should also be presented.

The computational cost seems to be a major concern using ML models and is mentioned multiple times throughout the paper. It would be interesting to have indications about the hardware used and the total time needed for the Bayesian optimization, the learning phase, and the running phase along with the time needed to calibrate and run the two benchmark models. This would help the reader to apprehend the computational implication of using ML models.

Finally, the training is done on really different time period lengths. The results seem to suggest that there is no correlation between length of the training period and RMSE. This difference in period length is not really discussed in the paper. In general, having similar training periods would be beneficial to really compare the models' performances across catchments. Indeed, with the data provided, we do not know if the differences observed across catchments are due to catchments' characteristics or to the training time period length. I imagine that the heavy computational time forbids to re-run all catchments using similar datas. However, a re-run on a single catchment with > 30 yrs of data, but using only 10 years as for the Enns, could be interesting to assess the impact of the length of the training time series on the results. Note that this question of length of the time series available for training is an important point for application perspective. Indeed, water temperature measurements network are usually quite recent (few decades), compared to time series available for discharge.

Models evaluation and application

The Section 3.3 focuses on the catchment obtaining the second lowest RMSE. It would be interesting to perform a similar analysis for a catchment obtaining higher RMSE. For example, analyzing the Kleine Muehl, where the median of ML RMSE is close to *air2stream*, would be interesting. In any case, some plots of the whole time series for all catchments should be presented in Supplementary in order to allow the reader to look at

the real outputs and not only have access to RMSE boxplots (I mention here that I do particularly appreciate Figure 5 where all outputs are always shown in grey in the background).

Two main applications are mentioned: short term predictions (especially high summer water temperature peak) and climate change (CC). For the first application, further discussion could be added in Section 3.3. Indeed, In Figures 5, we do see that many short timed high temperature events are not captured during the summer, which would be problematic for predictive application. Metrics do exist to assess the quality of models to capture such events (see e.g. the two-alternative forced choice score used in Greissinger et al. (2019)). While this paper of course does not pretend to deliver operational models – I rather see the paper as a demonstration of current capabilities of ML models for water temperature predictions along with pros and cons of different models – a discussion of the performance of ML models regarding short term high temperature event would be a great addition.

The author choose an interesting approach to assess the ability of the model to cope with non-stationarity in time series by using the CC signal present in the measurements timeseries. They show that ML models still obtain good performance in the warm year 2015. This could be enhanced by showing the whole time series in order to see if the error grows with time (and to better compare with benchmark models). Using one catchment with >30 yrs time series and train it only with the first 10 years (as suggested above to assess the effect of training time period length) could also be interesting in this regard. Indeed, the temperature increase between the 80's and 2015 will be more important than the one in the time series of the Inn catchment used in the paper to do this validation. Note that the years 2003, which shows an important water temperature anomaly, is also an interesting benchmark year.

While these tests are important and showing that models are able to correctly predict water temperature when out of the training range is a really good point for ML models, the increase of water temperature expected with CC is far above the range tested here. As a consequence, I do not think that it is really possible to assess ML models' ability to correctly predict CC impacts on water temperature with historical data. A comparison with physically based models could be an approach, but is beyond the scope of this work. Consequently, I would suggest to revise the lines 535 and 536 in the discussion.

Minor comments:

Figure 1: Please specify where you obtained the catchments delineation. Also, the figure mention "Danube catchment" while Donau is used elsewhere in the paper. Maybe the English name Danube should be used, it would be more accessible to international readers.

Table 1: In addition to what is already mentioned above, add calibration and testing

periods to the table.

Line 48: date -> data

Lines 67-69: *"Another main concern is that parametric statistical models showed higher prediction performances on weekly, monthly or seasonal time scales in the past (Caissie, 2006) leading to a loss of temporal variation (DeWeber and Wagner, 2014)."*

Higher prediction performances compared to what (models and/or temporal resolution)?

Line 177: moths -> months

Line 255: Misplaced parenthesis and inversed sum bounds in eq (5)

Section 2.4.5: There is an overall inconsistency between parts of the text and equation in the usage of bold font for vector and matrix terms. E.g. in the paragraph at lines 308-310 they are in bold, while in the following paragraph they are in italic. I would suggest the usage of bold fonts everywhere.

Line 329: Hyperparameters meaning is never really defined in the paper

Line 356: The difference between validation and testing periods is not really clear. I understand the validation is used to choose the best models, and then test period is used to compare the set of best models (lines 377-380). This should be clarified. In addition, it is not clear if validation is done for the hyperparameters selection (the 5 setups mentioned at line 372) or between different trained version of the model using the same hyperparameters set. Also, is there first a phase to select the hyperparameters (which require to train and test the model), and then a new training phase, or are they both done at once?

Lines 366-367: Do you mean 60% -> training, 20% -> validation? Please clarify.

Lines 368-369: what is the "standard way of training neural networks…", the 50 times training or your approach?

Lines 395-397: Please add some citation to the statements made here.

Line 398-399: I do not understand this sentence.

Lines 438-439: Do you have any explanation regarding the difference of performance observed?

Lines 450-455: I'm not completely convinced by the significance of this regression. Indeed, simply from boxplots we do see that the difference between catchments is by far the most important predictor.

Line 487: What is the number of time steps and optimal time step here?

Lines 489-490: Total time should also be provided in order to see how this ~2h decrease is important. Also, how is the *p-value* obtained here?

Line 536: The claim about short term predictions and CC is too ambitious here. What is shown is the increase in performance compared to benchmark models, which is already a really important step. For short term prediction and CC, see my longer comments above, but I think more work and discussion are needed to really assess the ability of the models for these applications.

Lines 546-547: The improvement here can be explained by the Bayesian optimization method used?

Lines 548-550: what do you mean by "spatial information at different scale"? Indeed, ML models do not provide any spatially distributed output (which can be achieved with distributed physical models), but only point informations.

References

Griessinger, M. Schirmer, N. Helbig, A. Winstral, A. Michel, T. Jonas, Implications of observation-enhanced energy-balance snowmelt simulations for runoff modeling of Alpine catchments, Advances in Water Resources, Volume 133, 2019, https://doi.org/10.1016/j.advwatres.2019.103410.