Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/hess-2020-670-AC1, 2021

Authors answers to RC1

Moritz Feigl et al.

Author 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-AC1, 2021

Dear Reviewer,

We thank you for your encouraging and positive feedback and sincerely thank you for your insightful comments and suggestions. Please find our answers to your comments below.

Review:

The comparisons of models results with in situ measured data using only errors metrics is insufficient and does not help in providing robust conclusions regarding models accuracies, robustness and fitting capabilities. Specifically, using several kinds of goodness-of-fit indicators should be more useful: the coefficient of determination (R^2), the Nash-Sutcliffe efficiency (NSE), and the index of agreement d, are highly recommended for hydrological models evaluation (Legates and McCabe 1999; Moriasi et al. 2007; Harmel and Smith 2007; Gupta 1998, 2008; Krause et al. 2005).

Answer:

We agree that by choosing a variety of metrics, a more concise picture of model performance can be shown. However, we noticed that some metrics are not sensitive enough to compare the results of this study. The NSE values of the presented models were all >0.9 and usually around 0.98. We noticed that while we could still see differences in model performance in RMSE and MAE, there were no or hardly any differences in the first three decimals of the NSE. Similar observations were made for the coefficient of determination and the index of agreement. Therefore, we think that adding these metrics would decrease readability while not adding a lot of information, but we will add them to the appendix to allow for future comparisons.

The following table contains an overview of all metrics for the best ML models and the two benchmark models per catchment:

	Best ML model results						LM					Air2stream				
Catchment	Model	RMSE	MAE	NSE	d	R2	RMSE	MAE	NSE	d	R2	RMSE	MAE	NSE	d	R2
Kleine Mühl	XGBoost	0.740	0.578	0.982	0.995	0.983	1.744	1.377	0.899	0.971	0.903	0.908	0.715	0.973	0.993	0.974
Aschach	XGBoost	0.815	0.675	0.983	0.996	0.983	1.777	1.408	0.920	0.978	0.924	1.147	0.898	0.969	0.992	0.970
Erlauf	XGBoost	0.530	0.419	0.985	0.996	0.986	1.345	1.057	0.884	0.968	0.900	0.911	0.717	0.959	0.989	0.960
Traisen	FNN	0.526	0.392	0.985	0.996	0.985	1.254	0.970	0.912	0.977	0.915	0.948	0.757	0.951	0.988	0.955
Ybbs	RF	0.576	0.454	0.989	0.997	0.989	1.787	1.415	0.889	0.971	0.890	0.948	0.744	0.968	0.992	0.969
Saalach	XGBoost	0.527	0.420	0.977	0.994	0.979	1.297	1.062	0.864	0.961	0.883	0.802	0.633	0.951	0.988	0.955
Enns	FNN	0.454	0.347	0.984	0.996	0.985	1.425	1.166	0.834	0.951	0.840	0.835	0.671	0.946	0.986	0.952
Inn	FNN	0.422	0.329	0.984	0.996	0.984	1.376	1.098	0.829	0.949	0.830	1.170	0.952	0.882	0.968	0.882
Salzach	FNN	0.430	0.338	0.986	0.996	0.986	1.327	1.077	0.862	0.961	0.864	0.743	0.589	0.957	0.989	0.963
Donau	RNN-LSTM	0.521	0.415	0.986	0.996	0.989	2.145	1.721	0.842	0.955	0.843	1.099	0.911	0.961	0.990	0.968

We used the index of agreement by Willmott, 1981.

References:

Willmott, C. J. (1981). On the validation of models, Physical Geography, 2(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

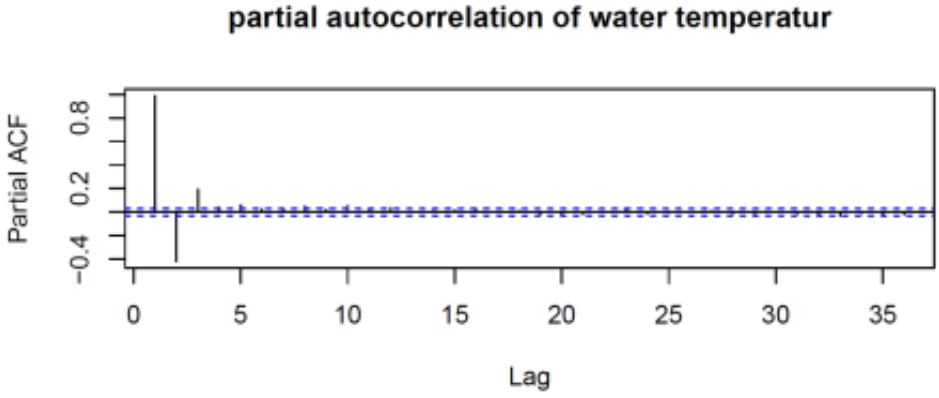
Review:

Models structures need to be clarified. In Lines 173-175, the authors argued that including the lag of all variables for 4 previous days can help in improving models accuracies according to Webb et al. (2003). First, using only 4 previous lag should be justified, on which basis it was selected (i.e., cross-correlation analysis can be helpful for answering this question)? Second, according to Webb et al. (2003), adopting the previous lag as input variables can be useful on only hourly data scenario. Therefore, a comparison between models with and without lag data may be a good option.

Answer:

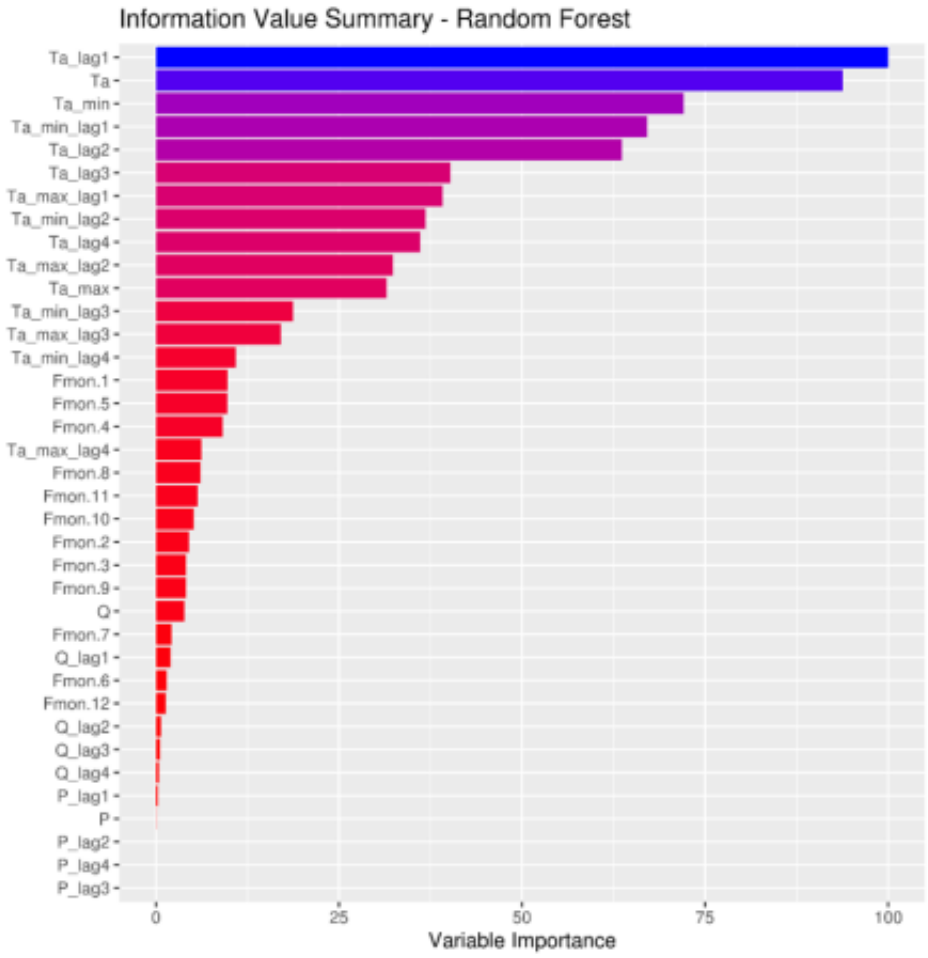
This is indeed an important point and we agree that we should explain our reasoning. Our decision on the number of lags was based on the results of an explorative data analysis of daily data we carried out before starting to work on the model structures. This included three major parts that guided our decision:

- Assessing partial autocorrelation plots of water temperatures: They showed a significant importance usually up to the 4th lag, as is illustrated in the following plot showing the partial autocorrelation of stream water temperatures of the Ybbs catchment.



- Assessing variable significance in a linear regression model using multiple lags: This also pointed to the fact that 4 lags are an adequate range for the given basins.
- Assessing variable importance in a simple Random Forest (RF) Model, which also

pointed to the fact that 4 lags are a reasonable choice, as is visible from the following plot showing the RF variable importance for predicting water temperatures in the Ybbs catchment. Variable importance refers to how much a given model "uses" that variable to make accurate predictions.



Our initial analysis pointed to the fact that lags are important and that 4 is a good choice for our set of basins, thus resulted in our decision for this study. Nevertheless, this is a relevant initial decision. Our results let us assume that this might be quite different for different basins and maybe a more dynamic approach might be valuable. This could for example be including the choice of lag time depending on the mean concentration time of a catchment, or the catchment size and should be explored in future studies.

We will add a short summary of these initial analyses after Line 173 to explain our choice of numbers of lags:

"The lag period of 4 days was chosen based on an initial data analysis that included (i) assessing partial autocorrelation plots of water temperatures, (ii) testing for significance of lags in linear regression models, and (iii) checking variable importance of lags in a random forest model."

Regarding the study of Webb et al. (2003): While they only showed the importance of lagged variables for hourly data, they also summarized findings regarding lagged air temperature for daily data (Grant, 1977; Jeppesen & Iversen, 1987; Stefan & Preud'homme, 1993). These previous findings, together with our initial analyses (especially the RF variable importance), made it clear that daily lags are relevant inputs

and necessary for a high model performance.

References:

Grant, P. J. (1977). Water temperatures of the Ngaruroro river at three stations. Journal of Hydrology, 16(2), 148–157. https://www.jstor.org/stable/43944413?seq=1

Jeppesen, E., & Iversen, T. M. (1987). Two Simple Models for Estimating Daily Mean Water Temperatures and Diel Variations in a Danish Low Gradient Stream. Oikos, 49(2), 149. https://doi.org/10.2307/3566020

Stefan, H. G., & Preud'homme, E. B. (1993). Stream temperature estimation from air temperature. JAWRA Journal of the American Water Resources Association, 29(1), 27–45. https://doi.org/10.1111/j.1752-1688.1993.tb01502.x

Willmott, C. J. (1981). On the validation of models. Physical Geography, 2(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

Review:

The introduction is not deeply written and in some cases need improvement. Specifically, the proposed ML reported in the literature should be presented, discussed, and the strength and weakness of each one would be more useful and effective if they are highlighted. Using lumped references do not help in understanding the mains contribution of the work.

Answer:

Thank you for pointing this out, comparing the different models instead of only giving an overview of past applications will make this section more informative. For the revised manuscript, we propose to include a more general overview of the different model approaches and their strengths and weaknesses.

Review:

Research gap. What are the mains contributions of the present study in comparison to what is already done? What does it add to existing literature?

Answer:

We agree, this might not be stated clearly enough in the manuscript yet. We think that the summary of the important contributions of this study in your general comments are very much on point. We will include it in the last paragraph of the introduction.

Review:

Lines 47 to 50, from Austria to characteristics. To our opinion this paragraph is more suitable to be moved to section 2.1.

Answer:

Agreed, we will change it according to your suggestion.

Review:

Line 79: ''To the author's knowledge, RF has not been applied for river water temperature prediction yet''. This statement is incorrect. The RF was recently reported as a powerful tool for predicting river water temperature (Heddam et al. 2020).

Answer:

Thank you for pointing this out. This is indeed an interesting and relevant publication. We propose to change the sentence in line 79 to the following:

"Up to date, only one previous study by Heddam et al. (2020) already applied RF for predicting lake surface temperatures."

Review:

Models comparison using cross-station scenarios can help in providing more conclusions, and a clear idea about models capabilities outside of their own catchment area: models calibration using data from on station and validated for other stations (i.e., see Zhu and Heddam 2019).

Answer:

We do agree that the application outside the initial trained catchment is an important type of application. However, we found it necessary to only focus on the model prediction capabilities in single catchments, to derive the general applicability of different model types and data inputs. These should be used as a foundation to derive transferable models and modelling approaches. The transferability of these models in this study cannot be adequately tested, as there is no information provided to the tested models to conduct this transfer. In our opinion, this transfer would need additional basin characteristics as inputs and consequently a larger number of basins for training and testing (multi-basin training). We certainly do see this as an important next step, but would refrain of applying the single basin trained models for this task. We thank the reviewer for this thought and propose to add this topic in the conclusions regarding future research fields.