Referee 2: “The authors conducted a study to explore the best method for predicting river water temperature in the case of limited observations. They compared the performance of five different methods, i.e., RF, ANN, SVR, the hybrid Air2stream model, Multiple Regression. In general, they found the ML techniques have better prediction performance with appropriate hyperparameters. RF achieved best evaluation metrics in their study. Based on the RF model, they studied the importance of input variables and the connection between the watershed time of concentration and model performance. However, I doubt the novelty of the study, since all the methods used in this study have been employed in previous studies in predicting river water temperature. Also, some of them have compared the performance of some of the methods, e.g., Zhu et al. (2018), Rajesh and Rehana (2021) and Rehana (2019). Moreover, the overall presentation is not clear to me. My comments are listed below.”

RESPONSE: We sincerely appreciate all your valuable comments and suggestions, which helped us improve the quality of the manuscript. To facilitate the work of the reviewers and editor, we refer to the former manuscript indicating the line that was modified.

Referee 2: “I doubt the novelty of the study, since all the methods used in this study have been employed in previous studies in predicting river water temperature. Also, some of them have compared the performance of some of the methods, e.g.., Zhu et al. (2018), Rajesh and Rehana (2021) and Rehana (2019)”

RESPONSE: Thank you for pointing this out. We agree with the reviewer. The ML algorithms considered in this study have been previously considered to model river water temperature. Please note that this fact is shown in Table 1 where we present a list of reviewed publications on river WT modelling. In these studies, the number of sites modeled is quite small with one exception: the study conducted by DeWeber and Wagner (2013) which considered an ANN. The studies conducted by Moore et al. (2003) and Ducharne (2008) applied a multiple regression model. Additionally, the studies were conducted using large observed air temperature and water temperature datasets. Even so, the predicted RMSE considering all study results varies from 0.42ºC to 2.74ºC (μ=1.15ºC; σ=0.6ºC). This variability is mainly caused by:

i) water temperature measurement errors (e.g., sampling depth variability);
ii) the fact that river WT is not only affected by local environmental conditions but also by upstream conditions (Moore et al., 2005);
iii) watershed morphological differences across the different regions;
iv) differences in the type of models applied and in the model parameterization;
v) the consideration of different model predictors.

In our opinion, the development of model intercomparison studies is a useful way to evaluate the model’s performance under different forcing conditions. However, we also think that these model intercomparison studies must include a large number of modeled sites to reduce the degree of modeling uncertainty. The intercomparison studies carried out by Zhu et al. (2019), Rajesh and Rehana (2021) and Rehana (2019) considered 8, 1 and 1 modeling sites respectively.

In other words, in our opinion, the number of models intercomparison studies:

i) should increase;
ii) should cover different regions;
iii) the number of modeling sites per study should be higher;
iv) the studies should also focus on the modeling of river water temperature with limiting forcing data.

To clarify the reviewer's concern the following sentence was included in last line of the abstract:

Page 1 – Line 24: "Hopefully, the high number of modeled sections considered in this model intercomparison study and the specific model forcing conditions will help to reduce the overall WT modeling uncertainty."

Referee 2: "The objective of the study is not very clear in the abstract and introduction sections. The authors did not explicitly explain the reason why they selected regions with a lot of missing data as the study domain. What is the novelty of their study compared to other studies that compared model performance?"

RESPONSE: We agree with the reviewer. Reviewer 1 also pointed this out and the following sentences have, therefore, been included in the manuscript abstract and introduction:

Page 1 – Line 11: "Commonly, the WT observed in monitoring stations located near the downstream section of rivers are assumed to be the boundary condition of lake/reservoir water quality models. The main goal of this study is to identify a suitable WT modeling solution for these sections given the scarcity of the forcing datasets."

Page 1 – Line 24: "Hopefully, the high number of modeled sections considered in this model intercomparison study and the specific model forcing conditions will help to reduce the overall WT modeling uncertainty."

Page 3 – Line 91: "Hence, the main objective of this study is to identify a suitable WT modeling solution to improve the lake/reservoir water quality models' boundary condition. It is important to mention that, for this study, an absence of significant variation between the water quality station observed WT and the WT at the downstream portion of a river was assumed, which coincides with the lake/reservoir water quality model boundary condition."

Page 3 -Line 91: the following sentence:

"It is also important to note that the studies defined to evaluate the performance of different modeling approaches are normally restricted to a very small number of test sites and usually contain a reasonable amount of forcing data (Table 1). Hence, the vital importance of increasing the number of test sites and using a limited amount of forcing data to model river temperatures. This is the primary objective of this study and the methodological approach was, therefore, defined to attempt to answer the following questions:"

was replaced with:

"It is also important to note that the studies defined to evaluate the performance of different modeling approaches are normally restricted to a very small number of test sites and usually contain a reasonable amount of forcing data (Table 1). Hence, the vital importance of increasing the number of test sites and using a limited amount of forcing data to model river water temperatures. This is the primary and innovative objective of this study. The methodological approach was, therefore, defined to attempt to answer the following questions: "

The reviewer is right in their comments regarding the lack of clarity in relation to the method of station selection. We have considered all the stations with available datasets of water temperature and discharge available from the Portuguese Water Resources Information System (SNIRH). To clarify the reviewer's concern the following sentence:

Page 4 – Line 101: "To that end, 83 river sections with different geomorphological, meteorological and hydrological conditions were modeled using five different models, three of which use ML algorithms optimized with a sequential model-based optimization approach: Random Forest (RF); Artificial Neural Network (ANN) and Support Vector Regression (SVR)."

has been replaced with: "To that end, 83 river sections with different geomorphological, meteorological and hydrological conditions were modeled. These stations correspond to all the sections for which the Portuguese Water Resources

Information System (SNIRH) holds WT and discharge datasets, which are also, coincidentally, characterized by 98% missing data. The modeling ensemble includes five different models, three of which use ML algorithms optimized with a sequential model-based optimization approach: Random Forest (RF); Artificial Neural Network (ANN) and Support Vector Regression (SVR). “

Referee 2: I think the authors confused the concept of different data sets in ML and hydrology. In the paper, the validation set was the same as the test set and the validation set, and the training set was the same as the calibration set. Typically, in ML, the study data is divided into three sets (i.e., training, validation and test sets) or two sets (training and test sets). ML models are fitted on the training set to learn the relationship between input and output data. The validation set is used for hyperparameter tuning. The optimal hyperparameters are determined based on validation performance. The validation set is similar to the calibration set in hydrology. Finally, the test set is used to evaluate the ability of the trained ML model to handle previously unobserved data. It is similar to the validation set in hydrology.

RESPONSE: Thank you for pointing this out. We did indeed write ‘validation’ when we meant ‘testing’. Due to the size of the datasets we have only considered a training set and a test set. The optimization algorithm was applied to the training set. We have replaced the word ‘validation’ with ‘testing’ throughout the report.

To clarify the reviewer’s concern the following sentence was included on:

Page 7 – Line 164: “Due to size of the available datasets the validation phase was not considered.”

Referee 2: In the introduction part of Section 3, the authors described the workflow in text. In the study, they first trained ML models at 12 wells, and then apply the 12 models to each well, using the ensemble of the best results obtained across the 12 models per well as the final ML results. A figure showing the workflow would be helpful for understanding.

RESPONSE: We agree with the reviewer. We have included a schematic and simplified representation of the modeling process

Referee 2: In Section 3.5, the authors did not state which data set was used in hyperparameter tuning. How to calculate the algorithm score there?

RESPONSE: Thank you for pointing this out. The following sentences were included to clarify the reviewer concern:

Page 10 – Line 269: “The coefficient of determination (R^2) was considered as the algorithm score.”

Page 10 – Line 271: “The algorithm was applied to the training data set. Table 4 shows the model parameters and the optimization range.”

Referee 2: Too many big tables and too many plots in most figures, which are very distracting. Maybe only show the most important information in the main text, and move the reset to the appendix or supplementary materials.

RESPONSE: We agree with the reviewer and six tables have been transferred to the appendix.

Referee 2: Line 169: “L ≤ 50; 50> L ≤100; 100> L ≤200; L>200” should be “L ≤ 50; 50< L ≤100; 100< L ≤200; L>200”

RESPONSE: Thank you. We have inserted the correction.

Referee 2: Line 294: delete the extra “root”.

RESPONSE: Thank you. We have inserted the correction.

Referee 2: Equation 9: Please explain “r”.

RESPONSE: Thank you for pointing this out.

Page 12 – Line 308: The following sentence:

“values and σo the standard deviation of the observed values:”

, was replaced with: “values, σo the standard deviation of the observed values and r is the Pearson coefficient:”

Referee 2: Line 326: Please explain “with 3-par”. Similar phrases in Table 1.

RESPONSE: Thank you for pointing this out.

The following sentence was included:

Page 10 – Line 254: In this study five versions of this model were considered to model WT. The 3, 4, 5, 7 and 8 parameter versions. Please refer to Toffolon and Piccolroaz (2015) for a full description of each one of the models' parameterizations.

Additionally, in the text 3-par was replaced with 3-parameters.

The following note was included at the end of Table 1:

*The model can be applied with 3, 4, 5, 7 or 8 parameters (3-par; 4-par; 5-par; 7-par and 8-par)

References

DeWeber, J. T., and Wagner, T.: A regional neural network ensemble for predicting mean daily river water temperature, J. Hydrol, 517, 187–200, https://doi.org/10.1016/j.jhydrol.2014.05.035, 2014.

Moore, R.D., Nelitz, M., and Parkinson, E.: Empirical modelling of maximum weekly average stream temperature in British Columbia, Canada, to support assessment of fish habitat suitability, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 38(2), 135-147, https://doi.org/10.1080/07011784.2013.794992, 2013.

Ducharne, A.: Importance of stream temperature to climate change impact on water quality, Hydrol. Earth Syst. Sci., 12, 797– 810, https://doi.org/10.5194/hess-12-797-2008, 2008.

Rajesh, M., and Rehana, S.: Prediction of river water temperature using machine learning algorithms: a tropical river system of India, Journal of Hydroinformatics, 23 (3), 605–626, https://doi.org/10.2166/hydro.2021.121, 2021.

Rehana, S.: River water temperature modelling under climate change using support vector regression, in: Hydrology in a Changing World: Challenges in Modeling, edited by Singh, S. K., and Dhanya, C. T., World Springer, 171–183, https://doi.org/10.1007/978-3-030-02197-9_8., 2019.

Zhu, S., Nyarko, E. K., Hadzima-Nyarko, M., Heddam, S., and Wu, S.: Assessing the performance of a suite of machine learning models for daily river water temperature prediction, PeerJ, 7: e7065, https://doi.org/10.7717/peerj.7065, 2019.