Referee 1: Determining inflow temperatures of rivers or streams into receiving water bodies is a critical need. The authors are to be commended for addressing this topic. The overall theme of the paper was not crystal clear – was this about modeling river water temperatures themselves or modeling the inflows into receiving water models? Did the modeling include hydrology models predicting depth and flow? If so, then it was not clear that the models used were compared to flow and depth data which are critical for modeling river temperatures. Also, the overall RMSE error for the models was much higher than accepted river temperature models. This leads to the conclusion that there were underlying issues in the datasets used in the model. If the datasets were improved, would the conclusions of the study have been any different? An important aspect of this paper that was not evaluated was the advantages and disadvantages of using a physical based model rather than correlations and regressions to model water temperature.

RESPONSE: Thank you for the time spent and for the thoughtful comments and suggestions towards improving our manuscript. To facilitate the work of the reviewers and editor, we refer to the former manuscript indicating the line that was modified.

Referee 1: "The overall theme of the paper was not crystal clear – was this about modeling river water temperatures themselves or modeling the inflows into receiving water models?"

RESPONSE: Thank you for pointing this out. The methodological approach considered in this study was defined to improve the thermal characterization of the lake/reservoir water quality models' boundary condition. This was the reason for undertaking the study - a practical need to improve the water temperature (WT) characterization at these sections. We assume that there are no significant variations between the water quality station observed WT, and the WT at the downstream portion of a river, which normally coincides with the lake/reservoir water quality model boundary condition. We could have just stated that we wanted to simulate the water temperature of rivers, however, in our opinion, it was important to reflect about the rationale behind the objective of this study regardless of the assumptions made.

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

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."

The following sentences were included in the introduction section:

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 4 – Line 97 – "This modeling solution will be considered to improve the characterization of lake/reservoir WT boundary conditions (assuming that the observed WT of the downstream sections of the rivers are the boundary condition of lake/reservoir water quality models)."

The following sentence was also included in the methodology definition section:

Page 7 – Line 151: "iv) There are no significant variations between the water quality station observed WT, and the WT at the downstream portion of a river, which coincides with the lake/reservoir water quality model boundary condition."

Referee 1: "Did the modeling include hydrology models predicting depth and flow? If so, then it was not clear that the models used were compared to flow and depth data which are critical for modeling river temperatures."

RESPONSE: No, the modeling approach didn't include hydrology models predicting depth and flow. This fact was clarified by the inclusion of a table with the models' input/output (Page 7 – Line 163 – Table 3). A reservoir can have a high number of tributaries and the application of a physical based model is commonly limited by the availability of data required, not only for the river's conceptual representation but also for the calibration of the model (e.g., observed depth time series). It is also relevant to mention that, as shown by Toffolon and Piccolroaz (2015), the river's thermal response on a daily timescale does not strongly depend on the flow depth.

Referee 1: "Also, the overall RMSE error for the models was much higher than accepted river temperature models. This leads to the conclusion that there were underlying issues in the datasets used in the model."

RESPONSE:

The quantity and quality of the training and testing datasets was one of the primary reasons for undertaking this study. In practice these are the datasets that will be considered for the characterization of the boundary conditions of lake/reservoir water quality models for this region. The Air2stream model and the multiple regression approach were included in this study to define a benchmark. It is well established that both approaches perform well with regard to predicting water temperature values. Therefore, in our opinion the results obtained with these models should be considered as the reference point for the analysis of the ML (machine learning) algorithm performance. Additionally, the results of this study should be compared with studies that also have a significant amount of missing daily river water temperature values. Piccolroaz (2016), modeled the water temperature of two lakes located in the USA (Lake Erie and Lake Superior) and showed that when the length of the calibration period is one year and the percentage of missing data is in the range of 99%, the RMSE between observed and predicted lake water temperature is >3.5ºC.

We agree with the reviewer, it is common to obtain RMSE values in river water temperature modeling studies of less than 1ºC. In fact, the manuscript includes a list of reviewed publications on river water temperature modeling and the corresponding RMSE between observed and modelled water temperature values. It is important to mention that most of these simulations include well characterized rivers in relation to the forcing meteorology and water temperature observed values. Furthermore, they are limited to a small number of rivers. The RMSE values varied from 0.42ºC (Feigl et al., 2021) to 2.74ºC (Zhu et al., 2019). This last value is high and similar to the overall mean RMSE error obtained considering the models ensemble results of this study of 2.75ºC.

We also agree that there are underlying issues in the datasets, namely:

i) The significant amount of daily river WT values that are missing (96.9% to 99.9%) (μ= 98.8%; σ=0.68) (e.g., the minimum training dataset as 11 water temperature values). In such small datasets an outlier can have a significant effect on the RMSE, and the models' overfitting will represent an important drawback;
ii) The length of the simulation period: 1980-2020. The training/testing dataset spans a long period of time. Hence, the interannual variation of river entering fluxes (e.g. hydropower release) may have a significant effect on the quality of the training and testing dataset;
iii) The fact that the meteorological forcing was obtained from ERA5 reanalysis. McNicholl et al. (2021) found large biases between the satellite and land data for a temperate region (Dublin) and for a tropical region (Singapore). Furthermore, the revised version of the manuscript will include the comparison of eleven ERA5 daily air temperature values with observed air temperature values (we have considered all meteorological stations located within a 5 km radius). Results show that the annual RMSE obtained between the two datasets considering all stations varied from 0.02ºC to 4.15ºC (μ=1.43ºC; σ=0.55ºC). This interannual variability was also found in the water temperature modeling results, which suggests that the

meteorological forcing also has a significant impact on the overall RMSE obtained in this study;

iv) The effect of upstream conditions.

The outliers of the datasets could have been removed and synthetic samples could have been generated for some poorly represented ranges. The performance of the models would be significantly increased. However, in our opinion, this process must be driven by the lake/reservoir water quality model calibration.

We have included the following paragraph to clarify the reviewer's concern regarding the quality of the datasets:

Page 29 – Line 558: "It is also relevant to mention that the results of this study suggest that, besides the WT dataset gaps, the modeling results were also affected by the presence of a large number of WT outliers, by the uncertainty induced by the mean air temperature ERA5 reanalysis datasets and by upstream conditions, which increases with the watershed area. The results of this study considering the quality of the input datasets suggests that when the missing datasets reach 98%, a RMSE <3.0ºC is indicative of a good modeling performance. Also relevant is the fact that this error can be further decreased by the generation of synthetic samples to some poorly represented ranges within the datasets, by applying a model such as SMOGN (Branco et al. 2017)."

Referee 1: "If the datasets were improved, would the conclusions of the study have been any different?"

RESPONSE: The main conclusion of this study would be the same: from a practical modeling perspective, when the number of predictor variables and observed river WT values are limited, the application of all the models considered in this study, considering the hyperparameter optimization algorithm (hyperopt) is quite relevant. The machine learning model results obtained in this study are very similar, hence, the best model, Random Forest, can be replaced by the SVR or by the ANN.

It is also important to mention that by improving the datasets we assume that the reviewer means removing outliers and adding synthetic samples to the datasets by applying a model, such as SMOGN. A pre-processing approach for imbalanced regression tasks (Branco et al., 2017). In fact, the modeling results of this study could have been significantly improved with the implementation of a pre-processing approach, such as SMOGN. This algorithm was not implemented because the user needs to assign a greater degree of importance to the predictive performance obtained for some poorly represented ranges compared to other more frequent ranges. In our opinion, this process needs to be driven by the water quality model temperature calibration process. Hence, we have chosen to preserve the original datasets and to evaluate the model's performance over the raw datasets.

Referee 1: "An important aspect of this paper that was not evaluated was the advantages and disadvantages of using a physical based model rather than correlations and regressions to model water temperature."

RESPONSE: We agree with the reviewer. It would have been interesting to evaluate the advantages and disadvantages of using a physical based model. However, as previously mentioned, the application of a physical based model to 83 rivers is limited due to the significant amount of data required, including stream geometry, land use, meteorological conditions, and heat flux components, which are difficult to compute. However, the analysis does include a hybrid model characterized by a physical based structure associated with a stochastic calibration of the model parameters, Air2stream v1.0.0. (Toffolon and Piccolroaz, 2015) considering five different parametrizations. The results of Air2stream and multiple regression were included to define a benchmark for the overall modeling results.

Specific comments:

Referee 1: "Abstract: Line 16: define variables used for 'Multiple Regression' – this should be a separate sentence in the Abstract where the variables used for the different approaches are described."

RESPONSE: We agree with the reviewer. The following sentence was included:

Page 1 – Line 14: "With the exception of Air2stream, which was forced with mean daily air temperature and discharge, all other models were forced with: mean, maximum, and minimum daily air temperature, mean daily total radiation (shortwave), mean daily discharge, month of the year and day of the year."

Referee 1: "After reading the abstract, the reader is not left with a better understanding on how to fill in data gaps, other than just take more measurements!"

RESPONSE: To clarify this point the following sentence was included:

Page 1 – Line 20: "Therefore, the datasets gaps can be filled with the best model of the ensemble approach."

Referee 1: “Line 44-45: ‘common practice to average out sub-daily effects and to consider a daily discretization for modeling purposes’ – the impacts of this should be explored further since this can impact significantly the waterbody being modeled”

RESPONSE: We agree with the reviewer. This is a very important and complex challenge for water quality modelers: increase the sub-daily samples of flow and water temperature values for the cases when only a dataset of mean daily inflow values is available. We have included the following sentence to stress the importance of this limitation:

Page 2 – Line 46: “This assumption can have a significant impact on lake/reservoir water quality modeling results, namely when lake/reservoir inflows are large. The fall and spring turnover onset, stratification strength/length and the overall heat budget can be affected.”

Referee 1: “Line 45: ‘Air temperature approximates the equilibrium temperature of a river and is, therefore, frequently used as the independent variable;’ – these are indeed different even though the air temperature responds to the same atmospheric forcing as the equilibrium temperature of the waterbody. So, the text should read that air and equilibrium temperature correlate – but are not approximations for each other.”

RESPONSE: We agree with the reviewer. Thank you for pointing this out. This sentence was changed:

Page 2 – Line 46: “Air temperature correlates with the equilibrium temperature of a river and is, therefore…”

Referee 1: “Line 65-70: Almost as important as the approach used are the variables used in the regression and other models. Could a listing of the predictor variables be itemized? Or is it just air temperature?”

RESPONSE: We agree with the reviewer. The following table was included on the revised version of the manuscript:

Page 7 – Line 163:

Table 3: Model predictor variables

Model	Predictor variables	Output variable
RF	Mean, max., and min. daily air temperature (ºC) Mean daily total radiation (shortwave) (Jm^{-2}) Mean daily discharge ($m^3.s^{-1}$) MOY and DOY	Water temperature
ANN	Mean, max., and min. daily air temperature (ºC) Mean daily total radiation (shortwave) (Jm^{-2})	

	Mean daily discharge ($m^3.s^{-1}$) MOY and DOY	
SVR	Mean, max., and min. daily air temperature (ºC) Mean daily total radiation (shortwave) (Jm^{-2}) Mean daily discharge ($m^3.s^{-1}$) MOY and DOY	
Air2stream	Mean daily air temperature (ºC) Mean daily discharge ($m^3.s^{-1}$)	
ML	Mean, max., and min. daily air temperature (ºC) Mean daily total radiation (shortwave) (Jm^{-2}) Discharge ($m^3.s^{-1}$) MOY and DOY	

Referee 1: "Line 74: Why single out wind velocity? One also needs air temperature, dew point temperature (or relative humidity), cloud cover and short-wave solar radiation also to compute the heat balance."

RESPONSE: Because it is very difficult to have accurate shading and daily wind velocity and direction datasets for unmonitored (in regard to meteorology) water quality stations. However, we think that the sentence will be improved with the inclusion of the reviewer's suggestion.

Page 3 – Line 74: "...although they do require a large amount of forcing data, including stream geometry, air temperature, dew point temperature (or relative humidity), cloud cover and short-wave solar radiation, degree of shading and wind direction/ velocity."

Referee 1: "Line 76-77: It is not clear if the river modeling is using the predictors or if it is just the boundary conditions are being predicted for use in a river model."

RESPONSE: The reviewer is right. This sentence is not clear and was changed:

Page 3 – Line 77: "The number and type of predictor variables considered to force river WT models in several intercomparison studies is quite different."

Referee 1: "Table 1: I assume the error statistics are for the river WT – not the boundary conditions. So, I assumed in reading this that for each of these models, there was no explicit river model other than the correlations/stochastic models. But as a reader I am confused since I thought the intent of the paper was focused on boundary conditions and techniques to determine boundary conditions."

RESPONSE: The reviewer is right, the error statistics are for river WT, and not for the boundary conditions. We think that the sentences included in the abstract and on pages 3, 4 and 7 clarifies the main goal of the study.

Referee 1: “Line 86/87/105: Now the focus moves to lake or reservoir models. I thought the focus was on river models as the receiving water body – see Line 96. I agree that developing the boundary conditions would benefit lake/reservoir models, but the focus in the abstract and throughout the paper needs to be refocused to include any receiving water quality model, not just rivers.”

RESPONSE: The reviewer is right. The error statistics are for river WT and not for the boundary conditions. Also in this case, we think that the sentences included in the abstract and on pages 3, 4 and 7 clarifies the main goal of the study.

Referee 1: “Line 136/Table 2: I have no idea what the total number refers to in Table 2 nor the statistics. Are these ‘predictors’ or are these WT in the rivers? What are the units of mean, etc.? And please itemize clearly what the predictors are for training and validation. Or do they vary? This is critical to understanding if this approach can be used by others.”

RESPONSE: The total number shown in Table 2 is the sum of all WT values considering all stations. The predictors are the same for all models except for the Air2stream model. The inclusion of Table 3, in our opinion, clarifies this point.

Referee 1: “Line 143: ‘model a significant number of watersheds’ – does this mean just for WT and flow? Or does it include stage also?”

RESPONSE: Only WT was modeled. We think that the inclusion of Table 3 clarifies this point as it shows that the output variable for all models is WT.

Referee 1: “Line 148/149: Not having on-site meteorological data is large weakness of this study. We have found that the on-line estimates are often poor and significantly affect the model predictions. In the basin where you did your analysis, there must be some meteorological stations that could have been used for ground-truthing the ERA5 ‘data’. Doing that comparison would also help inform readers of the bias in using estimates when there is no on-site meteorological data.”

RESPONSE: We agree with the reviewer. A new section has been included in the manuscript to describe the evaluation of the mean daily ERA5 air temperature values considering all meteorological stations located within a 5 km radius of the studied stations (the following text, table and figure were included):

Page 7 – Line 166: The results section starts with the evaluation of the ERA5 mean daily air temperature datasets. These datasets were compared with ground measurements of mean daily air temperature considering all the meteorological datasets located within a 5 km radius of the stations considered in this study.”

Page 13 – Line 326: ”4.1 Air temperature - ERA5 versus ground observed datasets

In this analysis the observed air temperature datasets of a total of eleven meteorological stations were considered. These are all available air temperature datasets observed within a 5 km radius of the stations considered in this study. Results show that the mean RMSE obtained between the two datasets considering all stations varied from 1.26ºC to 2.05ºC (μ=1.54ºC; σ=0.24ºC) and that, according to the mean bias values, the ERA5 tends to overestimate the observed air temperature datasets at 91% of the stations. Overall, a mean value of 1.54ºC (σ=0.24ºC) and a mean NSE value of 0.90 (σ=0.07) is indicative of a good performance. However, as shown in Figure 2, there are some sporadic significant discrepancies between the two datasets (Fig. 2). Additionally, results show that the stations with a RMSE higher than 2ºC are scattered all over the country. In this context it is relevant to mention that McNicholl et al. (2021) also found large biases between the ERA5 daily air temperature datasets and land data for a temperate region (Dublin) and for a tropical region (Singapore). Generally, these results suggest that the consideration of the ERA5 air temperature datasets for WT modeling can, sporadically, induce some important discrepancies between the two datasets. The error can significantly increase if the model’s training/testing dataset is small.

Table 5: Evaluation of ERA5 daily air temperature datasets - MAE, RMSE, NSE, KGE, bias and R^2 (with standard deviation) between observed and ERA5 values

Station	N*	MAE	RMSE	NSE	KGE	bias	R^2
st4	80	1.10±0.26	1.39±0.28	0.91±0.04	0.94±0.05	0.74±0.47	0.94±0.02
st6	120	1.10±0.37	1.34±0.38	0.90±0.17	0.92±0.09	-0.15±0.79	0.90±0.11
st30	98	1.31±0.29	1.72±0.40	0.91±0.07	0.95±0.06	-0.48±0.70	0.92±0.05
st32	67	1.16±0.52	1.43±0.58	0.96±0.04	0.94±0.05	-0.75±0.90	0.97±0.02
st38	110	0.88±0.34	1.26±0.48	0.94±0.09	0.96±0.06	-0.46±0.57	0.95±0.04
st42	21	1.19±0.47	1.53±0.58	0.93±45.49	0.87±2.22	-0.42±0.75	0.94±0.03
st50	90	1.08±0.30	1.45±0.48	0.91±0.06	0.89±0.11	-0.14±0.39	0.92±0.04
st62	24	1.30±0.74	1.67±0.80	0.89±6.28	0.94±0.92	-0.17±1.36	0.90±0.04
st68	47	1.60±1.18	2.05±1.09	0.71±9.81	0.86±0.23	-1.47±1.24	0.88±0.2
st83	137	1.49±0.40	1.79±0.39	0.92±0.03	0.94±0.04	-0.60±0.88	0.93±0.02
st91	51	1.04±0.13	1.33±0.16	0.93±0.04	0.96±0.09	-0.46±0.47	0.94±0.04

*number of dataset values

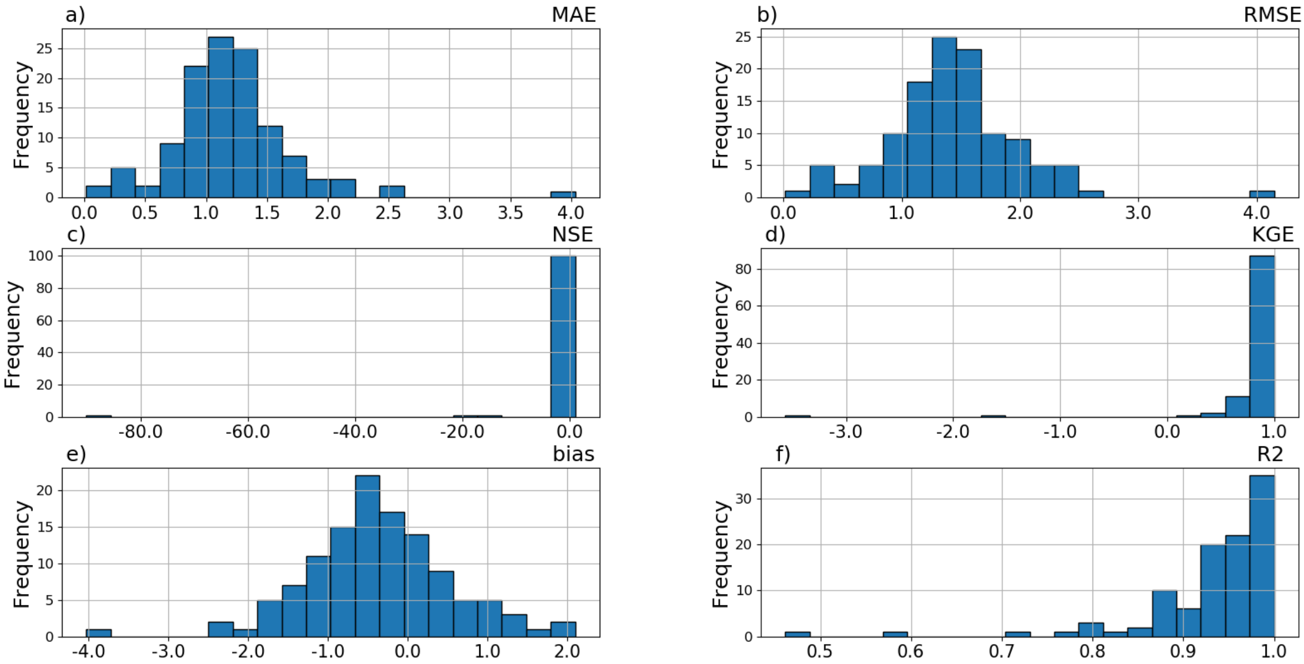

Figure 2: Metrics histograms of daily air temperature - ERA5 *versus* ground observed datasets

Additionally, the following sentence was included in the discussion section:

Page 13 – Line 558: "It is also relevant to mention that the results of this study suggest that, besides the WT dataset gaps, the modeling results were also affected by the presence of a large number of WT outliers, by the uncertainty induced by the mean air temperature ERA5 reanalysis datasets and by upstream conditions, which increases with the watershed area. The results of this study considering the quality of the input datasets suggests that when the missing datasets reach 98%, a RMSE <3.0 ºC is indicative of a good modeling performance. Also relevant is the fact that this error can be further decreased by the generation of synthetic samples to some poorly represented ranges within the datasets, by applying a model such as SMOGN (Branco et al. 2017)."

Referee 1: "Line 155: Why was this lapse rate chosen, -6.5oK/km? What are the impacts of assuming a fixed lapse rate over all your model domains?"

RESPONSE: Thank you for pointing this out. Initially this correction -6.0ºC /km (Faher and Harris, 2004) was applied to allow the comparison of the ERA5 air temperature reanalysis with ground measurements. The observed lapse rate is variable as a function of location and other variables, however, in our opinion, this correction is relevant. A constant lapse rate does not affect the modeling results.

The following sentence was corrected:

Page 7 – Line 156: … by considering a linear variation of air temperature with the altitude, dT/dz=-6.0°C/km.

Referee 1: “Line 163: Why is there a larger training dataset than a validation dataset? One would expect then with more training data that the results should be ‘better trained’ or more valid? What happens if the training and validation datasets were 80-20%?”

RESPONSE: If we are certain that the model correctly describes the process that we are modeling, then to determine the model parameters we can use all available data. However, in practice we are not sure that the model describes the phenomenon correctly. If we apply the model to the entire dataset the model will overfit the data and the algorithm, unfortunately, will not perform accurately against unseen data.

Hence, to avoid overfitting we divide the observations into training and testing datasets. Empirical analysis has shown that the best results are obtained if we consider 30%-20% for testing and the remaining percentage for training 70%-80% (Gholamy et al., 2018). Nguyen et al. (2021) shows that machine learning models are greatly affected by the training/testing ratios and concluded that the 30/70 ratio presented the best performance of the models. In this study we have chosen the 30/70 ratio because we have very small datasets for some stations and the consideration of a 20/80 ratio would determine a very small testing dataset.

Referee 1: “Line 196: ‘The results from the various models were evaluated with six metrics considering the observed and predicted annual, dry and wet season datasets for river WT.’ Does this imply that the predicted annual WT was used as a metric? Of what value is such a metric? Most river WT models are focused on maximum daily temperatures for fish habitat.”

RESPONSE: Thank you for pointing this out. No, the models were used to predict daily values of WT. This sentence was rewritten:

Page 8 – Line 197: The results from the various models were evaluated with six metrics considering the observed and predicted daily datasets of river WT. During the results evaluation three types of datasets were considered:

Annual datasets: All available daily averages of WT are compared to field data;

Wet season: Only the daily averages of WT corresponding to the wet season are compared to field data (October to March)

Dry season: Only the daily averages of WT corresponding to the dry season are compared to field data (April to September).

Referee 1: “Line 243 Eq (1): The equation has an error in the last term.”

RESPONSE: Thank you for pointing this out. The equation was corrected.

Referee 1: “Eq (1): This river equation assumes that the flow rate is based on steady-state flow with no dispersion. I assume this model runs on a daily time step – assuming a new steady-state distribution each day? This should be clarified. Also, the term H is a critical parameter in this model, how was it determined and what were the meteorological variables necessary for its computation?”

RESPONSE: No, the model converges to a single steady state distribution. The model considers a dimensionless discharge (theta=(discharge/mean discharge)^1/m) where m is related with the exponent of the rating curve. The model is fitted to the entire input dataset (air temperature, water temperature and discharge) and the value of m and the value of all the other model coefficients are estimated during the model optimization process (calibration).

The following sentence was included:

Page 10 – Line 251: “The parameter, a4 is related with the exponent of the rating curve. The model is fitted to the entire input dataset (air temperature, water temperature and discharge) and the value of a4 and the value of all other model parameters are estimated during the model optimization process (calibration phase).”

In fact, H is the net heat flux at the water-atmosphere interface. The model assumes that air temperature can be used as a proxy for all surface heat fluxes. A Taylor series expansion is used to include the overall effect of air temperature (Toffolon and Piccolroaz, 2015).

The following sentence was included to clarify the reviewer’s concern:

Page 10 – Line 247: “The model assumes that air temperature can be used as a proxy for all surface heat fluxes. A Taylor series expansion is used to include the overall effect of air temperature.”

Referee 1: “Eq (4): What precisely were the regression variables used in this model?”

RESPONSE: The model predictor variables were included in Table 3: Mean, max., and min. daily air temperature (ºC), mean daily total radiation (shortwave) (Jm^{-2}) discharge ($m^3.s^{-1}$), month of the year (MOY) and day of the year (DOY).

Referee 1: “Line 285: Was time of concentration only for the hydrology prediction? What is the hydrology model prediction equation?”

RESPONSE: No, the watershed time of concentration was estimated because it encapsulates some of the main watershed characteristics that affect the river water temperature. The main objective was to evaluate whether the modeling error was correlated with this variable.

Referee 1: “Line 317: With the best performing model to have a RMSE of over 3oC is not convincing. For river temperature models, this type of error in the river or in the boundary conditions is too high. There must be other issues with your approach that lends itself to such a poor predictor. In our experience, river models (and lakes and reservoirs) are often well below 1oC RMSE. I would not use any of these approaches if it had such a high RMSE. And if you fixed the underlying issues, the best approach may change.”

RESPONSE: In our opinion, the RMSE of the best performing value must be evaluated in the context of the study premises: 98% of missing data. This RMSE encapsulates the results of modeling a very large number of river sections (83) considering the datasets of a global climate reanalysis.

As water quality modelers we have very good results for lakes/reservoir and rivers WT predictions (RMSE < 1ºC), considering ML solutions and physical-based models, such as QUAL2E or CE-QUAL-W2. But we also have poor results. This balance depends primarily on the quantity and quality of the available datasets.

Referee 1: “Table 4 – provide units. By annual datasets – what does that mean? You are comparing annual averages or daily averages to field data?”

RESPONSE: Thank you for pointing this out. The units were included. We are comparing daily averages to field data, considering the entire dataset which includes dry and wet season.

The following paragraph was included to clarify this point:

Page 10 – Line 197 "The results from the various models were evaluated with six metrics considering the observed and predicted daily datasets of river WT considering the observed and predicted annual, dry- and wet season datasets for river WT. During the results evaluation three types of datasets are considered:

Annual datasets: All available daily averages of WT are compared to field data,

Wet season: Only the daily averages of WT corresponding to the wet season are compared to field data (October to March),

Dry season: Only the daily averages of WT corresponding to the dry season are compared to field data (April to September)."

Referee 1: "Table 5 – provide units. Explain dry season datasets – daily data during dry season, or averaged data over an entire season?"

RESPONSE: Thank you for pointing this out. The units were included. We are comparing daily averages to field data obtained for the dry and wet season.

Table 7 – provide units.

RESPONSE: Thank you for pointing this out. The units were included.

References

Branco, P., Ribeiro, R. P., Torgo, L., Krawczyk, B., Moniz, N.: Smogn: a pre-processing approach for imbalanced regression, Proceedings of Machine Learning Research 74, 36–50. 2017.

Fahrer, C., and Harris, D., 2004. LAMPPOST A Mnemonic device for teaching climate variables. Journal of Geography 103:86-90, https://doi.org/10.1080/00221340408978579, 2004.

Feigl, M., Lebiedzinski, K., Herrnegger, M., and Schulz, K.: Machine-learning methods for stream water temperature prediction, Hydrol. Earth Syst. Sci., 25, 2951–2977, https://doi.org/10.5194/hess-25-2951-2021, 2021.

Gholamy, A., Kreinovich, V., Kosheleva, O.: Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. Departmental Technical Reports (CS). 1209. https://scholarworks.utep.edu/cs_techrep/1209, 2018.

McNicholl, B., Lee, Y.H., Campbell, A.G., Dev, S.: Evaluating the Reliability of Air Temperature from ERA5 Reanalysis Data. IEEE Geosci. Remote Sens. Lett, 19, 1–5, https://doi.org/10.1109/lgrs.2021.3137643, 2021.

Nguyen, Q.H., Ly, H.-B., Ho, L.S., Al-Ansari, N., Van Le, H., Tran, V.Q., Prakash, I., Pham, B.T.: Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Math. Probl. Eng. 2021, 1–15, https://doi.org/10.1155/2021/4832864, 2021.

Piccolroaz, S.: Prediction of lake surface temperature using air 2stream model: guidelines, challenges, and future perspectives, Advances in Oceanography and Limnology, 7(1), 36-50, https://doi.org/10.4081/aiol.2016.5791, 2016.

Toffolon, M., and Piccolroaz, S.: A hybrid model for river water temperature as a function of air temperature and discharge, Environmental Research Letters, 10, 114011, https://doi.org/10.1088/1748-9326/10/11/114011, 2015.

Zhu, S., Hadzima-Nyarko, M., Gao, A., Wang, F., Wu, J., and Wu, S.: Two hybrid data-driven models for modeling water-air temperature relationship in rivers, Environ. Sci. Pollut. R., 26, 12622–12630, https://doi.org/10.1007/s11356-019-04716y, 2019.