We sincerely thank the editor and reviewers for taking the time to review our manuscript and providing constructive feedback to improve it.

RESPONSE TO REFEREE 2

Referee 2: Thanks for addressing my comments and adding a figure to explain their modeling process. The manuscript is easier to read than the previous version. However, its presentation quality still needs to be significantly improved before it is eligible for publication. Here are my comments.
Response: Thank you for your comments. We have improved the manuscript presentation quality.

Referee 2: Some long sentences need to be rephrased in order to be more understandable. The authors may split them into two shorter ones.
Response: Thank you for your comment. The manuscript was reviewed to further improve the clarity of the text.

Referee 2: I appreciate it that the authors added Figure 2 to explain their modeling process, but I feel the modeling process is not very clear. The authors may add more information to the caption of Figure 2 to explain their method.
Response: We agree with the reviewer. Fig.2 was modified, and the caption was replaced. The modeling process description was also modified as follows:

Line 217 "Given the large number of input datasets and the fact that the optimization process can be very time consuming, the following approach was implemented (Fig. 2):

i) The 83 stations were ordered as a function of the number of samples (lowest to highest) and were divided into four different classes ($L \leq 50$; $50 < L \leq 100$; $100 < L \leq 200$; $L > 200$). Three stations were selected within each class: 1) the station with the least samples; 2) the station with the most samples; 3) the station with the number of samples that most closely corresponded to the average sample number for each class. The 12 datasets selected corresponded to Stations 1, 7, 12, 13, 22, 29, 30, 46, 59, 60, 73 and 83;

ii) The ML and TPE algorithms were applied to the 12 datasets. At this stage there were 12 optimized model structures computed with the TPE algorithm for each ML model;

iii) The 12 optimized models obtained for each ML were subsequently applied to the 83 datasets and the best performing model at each station was calculated on the basis of the computed root mean square value (RMSE). Hence, the ensemble

of the best results obtained across the 12 different models for the 83 stations defines the overall ML results."

Referee 2: In the result section, the authors mentioned "ensembles" several times, such as ANN ensembles, but did not explain them anywhere. The authors may state that they evaluated the performance of the ensemble of the 12 models and optimal individual models.
Response: Thank you. The word "ensembles" was not correct and was replaced with "ensemble". This variable was already described in the manuscript.

Line 227: "Hence, the ensemble of the best results obtained across the 12 different models per station defines the overall ML results."

Referee 2: The quality of the figures still needs to be improved.
Response: Actioned. All figures have been improved.

Referee 2: In the discussion section, the authors mentioned that all models outperformed the others under some criteria. The authors may consider to provide a guideline on the selection of the models.
Response: Thank you. The following text was added to the discussion section to clarify the guideline that was considered for the selection of the models: "considering all metrics"

Line: 642 "That said, the RF model ensemble, considering all the metrics, produced the best results for the annual datasets and was the model that provided the greatest contribution in relation to overall ensemble results

Line: 663 "This explains the fact that the best results, considering all the metrics, were obtained with the 3-parameter model"

Specific comments:
Referee 2: Line 25-28: The sentence "when the number of predictor variables … (NSE): 0.56 ± 0.48)" is not clear to me. I am not sure what the authors mean here.
Response: Thank you. To clarify the reviewer's concern the following sentence has been modified:

This sentence:
"In general terms, the results of the study demonstrate the vital importance of hyperparameter optimization and suggest that, 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 is relevant (models ensemble mean annual – Root mean square error (RMSE): 2.75 ºC ± 1.00; Nash-Sutcliffe efficiency (NSE): 0.56 ± 0.48). Therefore, the datasets gaps can be filled with the best model of the ensemble approach."

Was replaced with: "In general terms, the results of the study demonstrate the vital importance of hyperparameter optimization and suggest that, from a practical modeling perspective, when the number of predictor variables and observed river water temperature values are limited, the application of all the models considered in this study is crucial. Basically, all the models tested proved to be the best for at least one station. The Root Mean Square Error (RMSE) and the Nash-Sutcliffe efficiency (NSE) values obtained for the ensemble of all model results was 2.75ºC ± 1.00 and 0.56ºC ± 0.48, respectively."

Referee 2: Line 97-98: It seems that the sentence "Hence, the vital importance …river temperatures" misses the verb.
Response: We have reviewed the sentence and we think that it is correct.

Referee 2: Line 155: What do "sections" mean here? Do they mean stations?
Response: The reviewer is right. Thank you. "sections" was replaced with "stations".

Referee 2: Line 182: Should it be "were computed"?
Response: The reviewer is right. Thank you.

Line 231. "where computed" was replaced with " were computed".

Referee 2: Table 3: Replace "ML" to "MR". Why does Air2stream only have 2 input variables?
Response: Corrected.

Referee 2: Line 207: Could you explain which 12 datasets were used for optimization?
Response: Yes. The following sentence has been included in the manuscript.

Line 217: "). Given the large number of input datasets and the fact that the optimization process can be very time consuming, the following approach was implemented (Fig. 2):
i) The 83 stations were ordered as a function of the number of samples (lowest to highest) and were divided into four different classes (L ≤ 50; 50< L

≤100; 100< L ≤200; L>200). Three stations were selected within each class: 1) the station with the least samples; 2) the station with the most samples; 3) the station with the number of samples that most closely corresponded to the average sample number for each class. The 12 datasets selected corresponded to Stations 1, 7, 12, 13, 22, 29, 30, 46, 59, 60, 73 and 83;"

Referee 2: Line 261: I think here should be testing dataset.
Response: This sentence is correct. "The model has no capacity to provide information on energy flux mechanisms within the river and has a tendency to overfit the training dataset, thereby considerably diminishing the model's ability to generalize the features or patterns present in the training dataset (Srivastava et al., 2014)."

Referee 2: Line 306: Please specify which disadvantages of the Gaussian Process the TPE algorithm fixes.
Response: 2. Thank you. The following sentence has been included:

"It can be difficult to select the right hyperparameters for GP with EI due to the many different Kernel types associated with this process. TPE uses simpler Kernels as a building block, which facilitates hyperparameter selection. Furthermore, TPE is faster than GP with EI when the number of hyperparameters increases."

Referee 2: Line 313: Table 3 does not provide the corresponding optimization range.
Response: The reference in the text was not correct. The optimization range of the models' parameters is provided in Table A1.

"Table A1 shows the models parameters and the corresponding optimization range."

Referee 2: Table 4: Please provide unit to the metrics.
Response: Corrected. Thank you.

Referee 2: Figure 3: What does the x axis mean?
Response: Thank you for pointing this out. It means the number of bins.
The figure has been corrected.

Referee 2: Figure 7: I believe here is MAE in the figure titles.
Response: Thank you. The figure has been corrected.

Referee 2: Figure 9: Please explain Ct in the caption.
Response: Corrected. Thank you.

RESPONSE TO REFEREE 3

The manuscript presents a study that employs five models to predict water temperatures for 83 low order rivers with limited water temperature datasets. The models used include three machine-learning algorithms, the hybrid Air2stream model with all available parameterizations, and a Multiple Regression model. Overall, the authors found the ML techniques exhibited superior predictive performance when coupled with appropriate hyperparameters. Among the models tested, RF demonstrated the best evaluation metrics. The authors suggest that the high number of modelled sections and specific model forcing conditions help to reduce the overall uncertainty in river water temperature modelling. The scientific approach and methods applied in this study are sound, and the results obtained are reliable and reproducible. However, the abstract and introduction are not well-organized and do not clearly state the significance, research gap, and novelty of this study. In addition, some parts of the discussion repeat information presented in the introduction, which should be revised accordingly. I think the present work will be a valuable paper in the field of modelling river temperature with a significant amount of missing values.
Response: We thank the reviewer again for the positive feedback. We agree with the reviewer. The abstract, introduction and discussion sections were improved.

Major comments:
Referee 3: 1. The scope of the main target of this study is formulated in a way that is too narrow. This manuscript presents a study aimed at filling gaps in observed river water temperature datasets to improve boundary conditions for lake/reservoir water quality models. However, this goes far beyond just being the boundary conditions for lake/reservoir models. The authors should consider this more, for instance, it can also be valuable for analyzing seasonal/diurnal trends and biogeochemical processes in rivers based on observation datasets. Furthermore, even for modelling lakes/reservoirs physics/hydrodynamics, inflow temperature is also critical. Therefore, it may be beneficial to at least delete the phrase 'water quality' from the sentence.
Response: We agree with the reviewer. The following sentences have been included in the manuscript:

Line 10 – “Water temperature datasets for low order rivers are often in short supply, leaving environmental modelers with the challenge of extracting as much information as possible from existing datasets, usually without the use of physically based models, due to the significant amount of data required (e.g., river morphology, degree of shading, wind velocity).”

Line 94 – “Therefore, the main objective of this study is to identify a suitable WT modeling solution for rivers with limiting forcing data. Improving this type of solution would deliver potential benefits for a wide range of environmental modeling applications, such as the analysis of seasonal/diurnal trends and biogeochemical processes in rivers based on observation datasets and the improvement of lake/reservoir water quality model boundary conditions.”

Referee 3: 2. Besides model performance, the number of the input parameters should also be taken into account when comparing models, for example, in AIC/BIC, model with fewer parameters has a better score. Given the results, I have the feeling that air2stream would be the most favorable alternative, as all the RMSEs were relatively high and similar (all over 3 oC), and some of the machine learning techniques exhibited signs of overfitting the datasets. However, air2stream needs fewer input predictor variables in comparison to other methods. Moreover, air2stream considers the physical process, which will be more robust.

Response: Thank you for your comment. As the reviewer will be aware, machine learning algorithms do not allow the computation of an AIC or BIC. The majority of these methods are neither likelihood-based, nor can one readily account for model complexity, because the number of parameters does not reflect the effective degrees of freedom (Hauenstein, 2017). We agree with the reviewer that considering the AIC/BIC definition the Air2stream model would produce a better score. We also think that a simple linear regression would also have produced a better score considering the penalization that the AIC/BIC equations give to the number of models parameters and the complexity of ML models. However, we do not see this as a fair comparison. One of the main conclusions drawn from this study is that, from a practical perspective, all models should be applied, as all models performed best for at least one station. The Air2stream considers the physical process, however, it also has some relevant simplifications. Overall, by considering a metric that is easily interpretable, such as the MAE, and 83 testing sites, the Air2stream performance was not better than the ML models.

From a practical perspective this is also a measure of model robustness. When describing nonlinear correlations, ML models have frequently performed as well or better than physical-based models with less input data (Virro et al., 2022).

Referee 3: 3. The author reiterated in both the manuscript and response that the models' error can potentially be mitigated through the use of a pre-processing technique, such as SMOGN. Therefore, it would be beneficial to investigate and compare the efficacy of SMOGN against the models' performance on the raw datasets.

Response: Thank you for this suggestion. As the reviewer mentioned we were trying to provide an additional approach to further improve the modeling results. We also mentioned that: "This algorithm was not implemented, because the user must assign more importance to the predictive performance obtained for some poorly represented ranges, in comparison 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."

However, we think that a reasonable implementation of SMOGN and the availability of the code can be beneficial for the manuscript and for the readers. We found a balanced way to implement SMOGN, considering:

1) that this process can be very time-consuming;
2) the initial manuscript structure and methodological approach.

We have generated 100 synthetic training datasets for each of the 12 raw datasets that were initially considered to define the 12 optimized models. Hence, for each of the 12 datasets, the best model (random forest) was optimized and trained with 100 different training datasets for each of the 12 stations. The methodology and discussion sections have been updated and a new section has been included to describe the modeling results (See Section 4.4). The SMOGN code has been added to the code and data repository (Almeida and Coelho, 2023).

Referee 3: 4. Some sentences in the discussion repeat the introduction, for example, in line 651 "but can also be associated with the uncertainty caused by the fact that river WT is not only affected by local environmental conditions, but also by upstream conditions" and line 64 "The predictor variables can represent a significant source of uncertainty, as river WT is not only affected by local

environmental conditions, but also by upstream conditions (Moore et al., 2005)." To avoid such repetition, it is recommended to consolidate related ideas and present them cohesively throughout the manuscript.
Response: Thank you, the manuscript text has been reviewed.

Minor comments:
Referee 3: 1. Figure 8a: Change the legend "validation" to "testing".
Response: Actioned.

Referee 3: 2. Figure 9: Add a legend explaining the different colored dots, and consider plotting a regression line in green that represents the regression during the testing period in Figure 9a and 9c instead of separating Figure 9b and 9d from Figure 9a and 9c.
Response: Actioned.

Referee 3: 3. Table 2 and 5: Change "Stdev" to "Standard deviation".
Response: Corrected. Thank you.

Referee 3: 4. Line 158: "The watershed discharge data used to force the models and the water temperature considered for the model's validation are also available from Portuguese Water Resources Information System (SNIRH)." Change the word "validation" to "test".
Response: Corrected. Thank you.

Referee 3: 5. Line 203: "Following this initial analysis, the models (vide Sect. 3.1 to 3.6) were applied to each of the 83 input datasets, divided between a training (70% of the entire dataset) and a test dataset (the remaining 30%)." It may not be appropriate to use the terms "training" and "test" for air2stream model. In hydrology, calibration and validation are more accepted terms.
Response: Thank you. We agree with the reviewer and the following sentence has been included:

Line 214 "It should be noted that, in the case of the Air2stream model, 70% of the initial dataset corresponds to the calibration dataset and the remaining 30% to the validation dataset."

Referee 3: 6. Line 289: "The Air2stream model solves a lumped heat-exchange budget between an unknown river section volume, its tributaries, and the

atmosphere (Toffolon and Piccolroaz, 2015).” It is suggested to add groundwater term to the sentence describing the air2stream model.
Response: Thank you for the suggestion, the sentence has been modified.

Line 281. “The Air2stream model solves a lumped heat-exchange budget between an unknown river section volume, its tributaries, groundwater, and the atmosphere (Toffolon and Piccolroaz, 2015).”

Referee 3: 7. Line 305: “In this study five versions of this model were considered to model WT. The 3, 4, 5, 7 and 8 parameter versions.” A dot is missing in front of the sentence.
Response: Corrected. Thank you.

Referee 3: 8. Eq (4): Specify the input variables in the equation to improve clarity.
Response: Actioned.

Referee 3: 9. Eq (8): Explain what ol mean?
Response: Thank you. ol has been replaced with $\overline{o}$, which is the observed values mean defined in the original document.

References

Almeida, M.C. and Coelho, P.S.: mcvta/WaterPythonTemp: Release 0.2.0, Zenodo [code]. https://doi.org/10.5281/zenodo.7870379, 2023

Hauenstein, S., Wood, S.N. and Dormann C.F..: Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation, Communications in Statistics - Simulation and Computation, 47:5, 1382-1396, DOI: 10.1080/03610918.2017.1315728, 2018.

Virro H, Kmoch A, Vainu M, Uuemaa, E.: Random forest-based modeling of stream nutrients at national level in a data-scarce region. Science of the Total Environment 840: 156613. Volume 840, https://doi.org/10.1016/j.scitotenv.2022.156613, 2022.