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.

Major comments:

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.

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 ^{o}C), 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.

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.

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.

Minor comments:

1. Figure 8a: Change the legend "validation" to "testing".

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.

3. Table 2 and 5: Change "Stdev" to "Standard deviation".

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".

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.

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.

7. Line 305: “In this 305 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.

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

9. Eq (8): Explain what o_l mean?