The manuscript on "*Machine Learning in Stream/River Water Temperature Modeling, a review and metrics for evaluation"* focuses on providing a comprehensive review of Machine Learning studies, including traditional and recent methods in ML and AI, on stream temperature modeling and prediction. Overall, the manuscript is well-written and covers most of the relevant papers, but there are a few strategic points I would like to share with the authors:

- Figures 1 & 2 & 3 & table 2: The manuscript provides a table for multiple metrics such as R2, NSE, RMSE, and MAE, and suggested a rate of numbers to rate the ML methods' performances. This table is based on the metrics that have been achieved by the studies in the previous years which are reflected in figures 1 & 2 & 3. However, those studies vary in terms of case studies, number of basins included in the study, running regional or local models. We know that ML models are prone to overfitting, especially for stream temperature that follows a relatively sinusoidal curve through a year, which means it is more predictable for complex models such as LSTM. However, it means the models are prone to easily overfit. Therefore, I suggest the authors encourage the stream temperature researchers to go towards making more generalizable models and less overfitted. For example, instead of suggesting performance metrics, the authors can provide a few steps to make sure the models are not overfitted or underfitted. For instance, always considering a spatial test on ungauged sites (basins). We know that spatial tests are more difficult tasks rather than temporal tests. Therefore, it is acceptable to get lower performance on ungauged basins, however, the metrics should not be very different from temporal tests. A more challenging experiment is to test the trained model on regions that have not been seen by the model. In theory, if a model has been able to capture true relations between the driving factors on stream temperature, it should achieve a relatively decent performance on basins

with different hydrologic, geologic, and climatic characteristics from the trained basins. As a researcher on stream water temperature, I would rather to have a model that passes all these three tests (temporal, ungauged, unseen regions) with relatively close metrics, rather than having a model that gives very high performance in temporal tests and low performance in the other two tests.

- Evaluation of Data Requirements: The manuscript does not extensively discuss the challenges that ML stream temperature modelers are facing with. Different ML models have varying data requirements, but the review does not thoroughly discuss the data needs for each type of model. For example, machine learning models are dependent on data. If we compare the availability of streamflow observation data availability versus the stream water temperature observation data, we realize there is a massive gap here, which impacts the studies and reduces the SWT model performances. I suggest, while the authors encouraging the researchers and water institutes to collect more data, they add their comments on this issue and discuss how researchers can reduce the impact of this problem in their models.
- Future Directions Could Be Expanded: Although the paper concludes with a general discussion of future challenges, it does not offer specific, actionable directions for future research. Highlighting key areas where ML can advance, such as the use of satellite data, sensor networks, or the fusion of climate models with ML, would provide more meaningful insights. In this concept, we can learn from hydrologic community and capitalize on their experience and what they learned. The ML hydrologic community is moving toward making global models, incorporating mechanistic models into their ML framework and learning the governing factors, flow prediction with predicted inputs (predicted

meteorological inputs) and last but not least, providing a seamless simulation in streams in CONUS/global scale. Therefore, I would ask the authors to add their comments on where the future direction of SWT community should be and how SWT community can achieve the future objectives and what the barriers are.

- The manuscript walked through many ML and AI models. An important factor of the ML and AI models are the inputs. I assume you faced a variety of inputs that have been used in the models. That would be informative to the readers, if the authors add their observations that what kind of inputs that have been missed to be used, either because it is not available yet or it is even missed. For instance, whether there is any geophysical attribute, climatic attributes, or any forcings that is worth to be extracted and used in ML models.
- Lack of Clear Structure in the Evaluation: Although the paper aims to summarize the performance evaluation metrics for ML models in SWT prediction, the organization of these sections feels somewhat scattered. A more systematic approach could improve clarity, such as separating the analysis based on time scales (e.g., hourly, daily, monthly) or spatial scales (local, regional, continental). This would make it easier for readers to find the relevant insights based on their application. For instance, a stream temperature model in monthly scale is different from a daily or hourly scale models on many aspects. As an example, the complexity of a daily model is different from a monthly temperature models. A monthly model may not need all inputs of a daily model to capture the monthly changes. The authors can add their overall opinion of what types of models are better fitted to which time scale. In ML models, it is important to know the scope of the model, whether it is a local model that needs to be calibrated site by site, or it is a model that is designed to work

for multiple sites (a regional model). I believe that would be informative to consider the modeling approach when methods are compared.

- I believe the authors need to decide first who are the readers of the papers. Whether the paper serves to new-commers to ML and AI methodologies in stream temperature community or it serves to researchers that are already familiar with basics of ML and AI methods. While the paper provides an extensive review of machine learning (ML) applications in stream water temperature (SWT) modeling, it focuses heavily on listing the types of ML models used rather than deeply analyzing their applications, strengths, weaknesses, and performance differences. A more critical analysis of the pros and cons of each model type could provide greater value to researchers choosing the appropriate model for their specific needs. To provide a few examples, I refer you to lines 136 – 143 & lines 146 – 159 & lines 263 - 292. The first half of the paragraph that is written in lines 136 – 143 explains the fundamentals of the method, which may not be necessary to be long, and the rest is an example of the method usage. However, this paragraph could have been enriched by statements like the advantages and disadvantages of this method compared to other existing ML methods or even to a linear regression method, or a 1D mechanistic method (although they are not ML methods, but the comparison is beneficial to the readers). The authors also can add their statement of under what conditions they think the method is beneficial. Lines 146 – 153 explains PCA and ck-means clustering on data reduction application, however, it is not clear here under what conditions we can use them. Additionally, that would be nice for readers if the authors add feature importance to their comparison as it has been used more frequently in streamflow and soil moisture prediction studies. Lines 263 – 292 are organized in three paragraphs while providing general

knowledge about ANNs with relatively less direct relations to water temperature application.

- Line 13: There is a typo that changes the meaning of the sentence. It should be "… with in situ …" or "… with in-situ …".
- Line 132; There is a typo here too. It should be "long short-term memory". Although I am trying to catch them, there is a chance that I miss some of them. I recommend the authors to carefully re-read the manuscript or ask help from a fresh pair of eyes to find these types of typos.
- Lines 208 – 210: to make the sentence more accurate, it needs to be stated whether these are local models or one model for multiple sites. Additionally, I believe by "NNs" here, the authors mean feedforward neural network, which are totally different from recurrent neural networks.
- Line 541 : "at" is missed. It is .. All journals examined used at least …"