Hydrology and Earth System Sciences,
Manuscript #HESS-2024-256
January 25th, 2025

Subject: Response to Comments on Review Paper "**Machine Learning in River/Stream Water Temperature Modeling: a review and metrics for evaluation**"

Dear Dr. Christa Kelleher, Mr. Jeremy Diaz and referees #1 and #2,

We thank you for your time and patience in reviewing our manuscript. We have revised the manuscript in line with referee feedback and think the manuscript is much improved as a result. This document serves as our written author response to referee comments.

This document separates responses by referee: referee #1 (pages 2 - 14), #2 (pages 15 - 20), and #3 (pages 21 – 30). Where referee comments had several parts, we separated the comments into "1a, 1b, etc.," to both indicate that the part belonged to one comment, and allow for a more organized response. Otherwise, all comments were kept in their original format.

For revisions, new/edited text is in BLUE, while removed text is shown as being crossed out. The statement "*revised lines XXX-XXX*" indicates the in-line placement in the track-changes manuscript "**2-Track-Changes-Manuscript-HESS-2024-256**", where the described changes will be found. The manuscript "**3-Clean-Manuscript-HESS-2024-256**" is the "final" version.

Once again, thank you kindly for your time and consideration.

Sincerely,

Claudia R. Corona
Postdoctoral Fellow
Colorado School of Mines

Terri S. Hogue
Dean, Earth and Society Programs
Colorado School of Mines

Referee #1 Comments

The manuscript on "*ML in Stream/River Water Temperature Modeling, a review and metrics for evaluation"* focuses on providing a comprehensive review of ML 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:

AUTHOR RESPONSE: We appreciate the referee's feedback and think the manuscript is much improved as a result. For reference, we separated some referee comments into a, b, etc., to provide a more organized response. Thank you for your time and insight. Proposed new/edited text is in BLUE. Revised lines in the track-changes manuscript are indicated by the statement: *(revised lines XXX-XXX).*

1a. 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.

AUTHOR RESPONSE: We agree that the SWT studies vary spatially/temporally and that ML models risk overfitting. We appreciate the referee's comments in pointing out areas of improvement and we suggest adding the following: 1) a new subsection 2.4.1 "SWT Predictions using ML" on overfitting/underfitting that suggests the need for temporal- and spatially-focused testing as suggested by the referee, and 2) a diagram showing initial steps to mitigate overfitting. The new text is below:

*new Section 2.4.1, Identifying Model Complexity *(revised lines 464-483)*

> The strong success of ML-use in SWT modeling warrants a brief and broad overview on identifying model complexity to minimize overfitting and underfitting" of models. When a model is too complex, i.e., has too many features or parameters relative to the number of observations, or is forced to overextend its capabilities, i.e., make predictions with insufficient training data, the model runs the risk of overfitting (Srivastava et al., 2014). An overfitted model fits the training data "too well", capturing noise and details that provide high accuracy on a training dataset, only to perform poorly once the model encounters "unseen" data in testing/validation (Xu and Liang, 2021). Scenarios where overfitting may be temporarily acceptable are: 1) model development is at preliminary stages, the interest is in a "proof of life" concept, 2) when the objective is to identify heavily-relied on features by the model, i.e., feature importance, or 3) in highly-controlled modeling environments where the expected data will be consistently similar to the training dataset. The latter is more likely in industrial applications and unlikely in the changing nature of hydrology.
>
> In contrast, underfitting occurs when a model is too simple to capture any patterns in the data, which can also lead to unsatisfactory performance in training, testing and validation. Underfitting can occur with inadequate model features, poor model complexity or when regularization techniques, (e.g., L1 or L2 regularization), are over-used, making the model too rigid and unable to respond to changes in the data. Given the propensity of ML models to effectively learn the training data, underfitting is less of an issue in ML whereas overfitting can be widespread. In Figure 1, we present an example

workflow that researchers can use to transition away from overfitting and towards generalizability. In the five-step outline (Fig. 1), we suggest the need for "Temporal, Unseen, Ungaged Region Tests" (TUURTs), which is a call for temporal and spatially-focused testing that can be used to strengthen model robustness.

Revised lines 484-486:

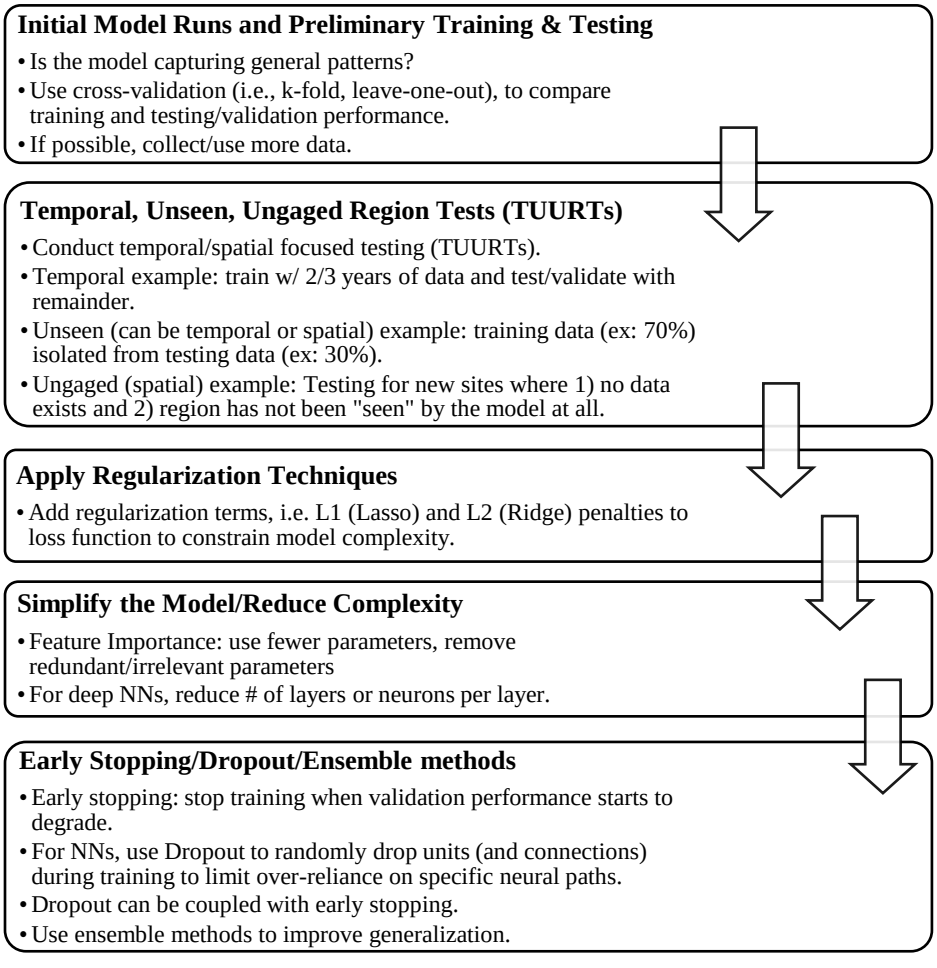


Figure 1. Diagram outlining steps that can be taken in modeling process to mitigate overfitting.

1b. Therefore, it is acceptable to get lower performance on ungauged basins, however, the metrics should not be vastly 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 SWT, I would rather to have a model that passes all these three tests (temporal, ungaged, unseen regions) with relatively close metrics, rather than having a model that gives high performance in temporal tests and low performance in the other two tests.

AUTHOR RESPONSE: We agree. The referee mentions a key point that having a SWT model pass all three tests for temporal, ungaged, and unseen regions may be more qualitatively sound, but as of this review, we had not yet seen any ML-SWT papers that test for all three cases. For example, Topp et al. (2023) held out a region to be considered "unseen" but did not test for ungaged basins. Hani et al. (2023) used an inverse weighted distance interpolation method to estimate values for ungaged sites but did not test for "unseen" data. Souaissi et al. (2023) used a leave-one-out cross-validation technique to mimic the estimation of quantiles at ungaged sites by temporarily removing the gaged site information, which is

arguably not testing for new, ungaged sites but rather "unseen" (i.e., tested only within the development dataset, not for new sites). Siegel et al. (2023), a non-ML paper tested for "ungaged" and "unseen" data, but did not perform a temporal test. A newly published example, Philippus et al. (2024), which considered spatial testing on ungaged basins, has been added. We further agree with the theory posited by the referee that a model capturing true relations should perform acceptably, however, we have yet to see a study that has adequately captured all true relations.

We have added a few sentences (blue is new) to the Discussion subsection 4.3, "ML Use for Knowledge Discovery" where we further urge for the use of TUURTs (Temporal, Unseen, Ungaged Region Tests)' *(revised lines 914-925):*

> While it is understandable that not every ML-SWT paper aims to explain physical processes, the SWT community should agree on a baseline of tests that all ML-SWT models undergo to assess model robustness and transferability. Specifically, we urge use of TUURTs (Temporal, Unseen, Ungaged Region Tests) for future ML-SWT models as a helpful step towards better modeling practices, increased model transparency and robustness (Fig.1). As stated in figure 1, for TUURTs, testing for "unseen" cases means testing only within the developmental dataset, whereas testing for "ungaged" cases means testing for new sites that have no data and have not been previously seen by the model at all. Due to the difficulty of conducting spatial tests compared to temporal tests, few ML-SWT studies have applied one or two of the tests, and rarely all three (Topp et al., 2023; Hani et al., 2023, Souassi et al., 2023). For example, Siegel et al. (2023), a non-ML SWT paper, tested for ungaged regions and unseen data but did not perform a temporal test. To our knowledge, Philippus et al. (2024), appears to be the only published SWT-ML study that applied TUURTs with some success. We further encourage modelers to shift towards more generalizable models, which are in theory, more capable of performing well across diverse scenarios and datasets, and stand to become increasingly important with the unpredictability of climate extremes.

2. Evaluation of Data Requirements: The manuscript does not extensively discuss the challenges that ML ST 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, ML models are dependent on data. If we compare the availability of streamflow observation data availability versus the SWT 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.

AUTHOR RESPONSE: We agree with the referee that issues remain with data requirement limitations. We propose adding a new 'Discussion' subsection, titled 'ML Data Requirements vs. Availability' stating the following:

*new section 4.2 ML Data Requirements vs. Data Availability *(revised lines 881-906).*

> While, in recent years, access to hydrologic data has improved (Miller et al., 2016; CUAHSI, 2024), data remains scarce for many hydrologic applications including SWT research, particularly because continual project management and funding to place and maintain stream temperature sensors, can be expensive and/or time-consuming to undertake. As a result, in the 21st century, the scarcity of data remains a large impediment for the application of machine learning in SWT modeling. What is more, the question of data quantity (how much data do you have?) versus quality (how much diverse data is needed?) continues to hinder ML use in hydrologic applications. Xu and Liang (2021) make the excellent point that one year of streamflow data (can swap for stream temperature) at 15-minute intervals equals about ~35,000 points, which may seem extensive, but is unlikely to be enough to properly train a ML model due to autocorrelation and limited exposure to diverse types of data that

are naturally encountered with a longer time-series (Xu and Liang, 2021). For example, machine learning models may only predict flood volumes they have previously seen (Kratzert et al., 2019). While data requirements for ML remain high, there are some strategies that researchers have used to alleviate this impact.

One strategy that hydrologists in other fields have used to tackle this problem is data augmentation, which can be applied spatially or temporally to create new training examples that the ML model can learn from. Spatial augmentation can be done by means of interpolation methods, i.e., kriging or distance weighting to create new data points or by generating synthetic data based on expected physical patterns to fill gaps in data coverage (Baydaroğlu and Demir, 2024). Temporal data augmentation can be done by shifting, scaling or adding noise to existing time series to create new training examples (Skoulikaris et al., 2022). Alternatively, and not a new idea, is to use the statistical technique known as seasonal decomposition, which breaks down a time series into its main components, i.e., the trend, seasonal patterns and residual components (Apaydin et al., 2021; He et al., 2022). These can then be recombined to generate new data and train the model for improved accuracy (Apaydin et al., 2021). In addition to data augmentations, data requirements can be alleviated by considering the help of unsupervised transfer learning, i.e., use pre-trained models on similar tasks to reduce amount of data needed for training, or semi-supervised learning, such as few shot learning, i.e., combine a small percent of labeled data with larger percent of unlabeled data to improve model performance (Yang et al., 2023). By implementing these strategies, researchers in other hydrologic fields have shown that models can be improved with less data, strategies that are likely transferable to SWT research.

3. 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.

AUTHOR RESPONSE: We agree and appreciate the referee's feedback. We propose adding a new 'Discussion' subsection, titled '4.4 Future Directions of SWT Modeling', with the following:

*new Section 4.4 Future Directions of SWT Modeling, *(revised lines 944-991):*

The utility of ML in hydrologic modeling has advanced significantly, with interest seemingly growing exponentially (Nearing et al., 2021). With the novelty of ML, it is easy to over-value model performance and ignore the physics of the system , but with several decades of ML-experience, we advocate it is necessary to purposefully use ML to address physically-meaningful questions and not just create ML for the sake of creating. Given this, Varadharajan et al. (2022) laid out an excellent discussion on opportunities for advancement of ML in water quality modeling, see section 3 of publication of Varadharajan et al. (2022). (Varadharajan et al., 2022)Here we highlight some of the questions from Varadharajan et al. (2022) that can be considered in the context of what objectives the SWT community should be using in the ML era, namely: 1) How do we use physical knowledge (re: heat exchange equations, radiation influence) to improve models and process understanding? Rahmani et al. (2023) coupled NNs with the physical knowledge from SNTEMP, a one-dimensional stream temperature model that calculates the transfer of energy to or from a stream segment by either heat flux equations or advection, but found that even with SNTEMP, their flexible NNs exhibited substantial variance in prediction and needed to be constrained by further multi-dimensional

assessments (Rahmani et al., 2023). In short, if our use of physics in machine learning makes our models worse, we should understand why.

A second question that needs addressing is 2) How do we deal with predictive uncertainty in ML used for SWT modeling? According to Moriasi et al. (2007), uncertainty analysis is the process of quantifying the level of confidence in any given model output based on five guidelines: 1) the quality and amount of observations (data), 2) the lack of observations due to poor or limited field monitoring, 3) the lack of knowledge of physical processes or operational procedures (instrumentation), 4) the approximation of our mathematical equations, and 5) the robustness of model sensitivity analysis and calibration. For example, in rainfall-runoff modeling, researchers have proposed benchmarking to examine uncertainty predictions of ML rainfall-runoff modeling (Klotz et al., 2022). For stream temperature modeling, researchers have attempted to address the role of uncertainty in deep learning model (RGCN, LSTM) predictions using the Monte Carlo Dropout (Zwart, Oliver, et al., 2023) and a unimodal mixture density network approach (Zwart, Diaz, et al., 2023).

Other questions that SWT-ML studies should consider is 3) How do we make ML models generalize better, specifically with regards to ungaged basins? And 4) How can ML models be improved to predict extremes? As ML models advance to use satellite data, include more sensor networks and/or couple with climate models, there is a logical next step toward creating generalizable models that can account for extremes. In our review, only two papers by the same group (Rahmani et al., 2020, 2023) conducted a CONUS-scale approach towards SWT-ML modeling, omitting hydrologically important regions in the southwest (CA) and southeast (FL). Recently, a satellite remote sensing paper used RF to model monthly stream temperature across the CONUS and tested for temporal (walk-forward validation), unseen and 'true' ungaged regions (Philippus et al., 2024). We have also learned that ML models such as LSTMs, generally only make predictions within the bounds of their training data (Kratzert et al., 2019), which is a limitation for predicting extremes. Thus, we strongly urge the community to work towards ML models that generalize better and/or are more robust towards predictions of extremes.

Finally, 5) How can we build ML models such that they are seen as trustworthy and interpretable by the hydrologic community? To answer this question, we must address a technical barrier (black-box issues, data limitations, model uncertainty) and a social barrier (i.e., educated skepticism of ML due to novelty, little understanding of computer science basics and/or coding experience). If we are to incorporate ML into decision-making processes, it makes sense that ML must be transparent and understandable to more than just computer or data scientists (Varadharajan et al., 2022). For example, Topp et al. (2023) recently used explainable AI to elucidate how ML architectures affected the SWT model's spatial and temporal dependencies, and how that in turn affected the model's accuracy. Addressing this technical barrier can also be done by improving access to data, which has seen remarkable progress thanks to web repositories such as NSF-funded CUAHSI's Hydro share (CUAHSI, 2024) and GitHub (GitHub, 2024). In the United States, data access to state and locally-based data remains limited, and should be addressed. In terms of the social barrier, education about ML and ML-use is key. Societal interest in ML has thankfully also lead to a plethora of educational resources and ML walk-through videos and tutorials in Tensorflow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), and Google Colab (Bisong, 2019). With the speed at which ML-use is evolving, short communication pieces (Lapuschkin et al., 2019) and opinion pieces (Kratzert et al., 2024) with clear examples about an ML-issue and practical solutions will also help make ML challenges more transparent and therefore accessible to the hydrologic community-at-large.

Added citations used for new subsection, 4.4 Future Directions of SWT Modeling:

1) Apaydin, H., Taghi Sattari, M., Falsafian, K., and Prasad, R.: Artificial intelligence modelling integrated with Singular Spectral analysis and Seasonal-Trend decomposition using Loess approaches for streamflow predictions, Journal of Hydrology, 600, 126506, https://doi.org/10.1016/j.jhydrol.2021.126506, 2021.

2) Baydaroğlu, Ö. and Demir, I.: Temporal and spatial satellite data augmentation for deep learning-based rainfall nowcasting, Journal of Hydroinformatics, 26, 589–607, https://doi.org/10.2166/hydro.2024.235, 2024.

3) CUAHSI. 2024. Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Water Data Portal: https://www.cuahsi.org/community/water-data-portals, last access: 13 November 2024.

4) Kratzert, F., Gauch, M., Klotz, D. and Nearing, G., 2024. HESS Opinions: Never train an LSTM on a single basin. Hydrology and Earth System Sciences Discussions, 2024, pp.1-19.

5) Kwak, J., St-Hilaire, A., and Chebana, F.: A comparative study for water temperature modelling in a small basin, the Fourchue River, Quebec, Canada, Hydrological Sciences Journal, 1–12, https://doi.org/10.1080/02626667.2016.1174334, 2016.

6) Philippus, D., Sytsma, A., Rust, A., and Hogue, T. S.: A machine learning model for estimating the temperature of small rivers using satellite-based spatial data, Remote Sensing of Environment, 311, 114271, https://doi.org/10.1016/j.rse.2024.114271, 2024.

7) Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

8) Skoulikaris, C., Venetsanou, P., Lazoglou, G., Anagnostopoulou, C., and Voudouris, K.: Spatio-Temporal Interpolation and Bias Correction Ordering Analysis for Hydrological Simulations: An Assessment on a Mountainous River Basin, Water, 14, 660, https://doi.org/10.3390/w14040660, 2022.

9) Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15, 30, 2014.

10) Yang, M., Yang, Q., Shao, J., Wang, G., and Zhang, W.: A new few-shot learning model for runoff prediction: Demonstration in two data scarce regions, Environmental Modelling & Software, 162, 105659, https://doi.org/10.1016/j.envsoft.2023.105659, 2023.

11) GitHub. 2024. About Git and Github: https://docs.github.com/en/get-started/start-your-journey/about-github-and-git, last access: 14 November 2024.

12) Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), p.1096.

13) Zwart, J.A., Oliver, S.K., Watkins, W.D., Sadler, J.M., Appling, A.P., Corson-Dosch, H.R., Jia, X., Kumar, V. and Read, J.S., 2023. Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. JAWRA Journal of the American Water Resources Association, 59(2), pp.317-337.

14) Zwart, J.A., Diaz, J., Hamshaw, S., Oliver, S., Ross, J.C., Sleckman, M., Appling, A.P., Corson-Dosch, H., Jia, X., Read, J. and Sadler, J., 2023. Evaluating deep learning architecture and data assimilation for improving water temperature forecasts at unmonitored locations. *Frontiers in Water*, *5*, p.1184992.

15) Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S. and Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrology and Earth System Sciences, 26(6), pp.1673-1693.

16) M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. TensorFlow. Website: https://www.tensorflow.org/

17) A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Website: https://arxiv.org/abs/1912.01703

18) Bisong, E. (2019). Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. Website: https://doi.org/10.1007/978-1-4842-4470-8_7

4. 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.

AUTHOR RESPONSE: We appreciate the referee's feedback. In the Supplementary Materials, Table S1 contains some of the suggested data by the referee, such as: period considered, region examined, temporal resolution of SWT, spatial scale of study, and hydrometeorological parameters used for modeling. We provided the information as Supplementary Material because Tables S1 and S2 are seven pages alone, which may risk making the review lengthier than it already is. We have added text to the manuscript regarding model inputs and moved the LASSO paragraph (original lines 247-253) to this section because we think it can more smoothly follow the paragraph on feature importance.

*new section 2.4.2 Model Inputs for ML-SWT (*revised lines 488-516*):

> Using air temperature (AT) to better understand SWT has been considered since the 1960s, when Ward (1963) and Edinger et al. (1968) discussed the influence of air temperature on SWT. Since then, studies have used varying input variables (see Table S1), however, the model inputs of AT and SWT continue to be the most used in ML-modeling studies. In particular, studies have used AT from time periods outside of the known SWT record to improve model performance (Sahoo et al., 2009; Piotrowski et al., 2015; Graf et al., 2019). In addition to AT and SWT, flow discharge has been used to attempt to constrain SWT (Foreman et al., 2001; Tao et al., 2008; St-Hilaire et al., 2011; Grbić et al., 2013; Piotrowski et al., 2015; Graf et al., 2019; Qiu et al., 2020). Traditionally-used model inputs include precipitation (Cole et al., 2014; Jeong et al., 2016; Rozos, 2023), wind direction/speed (Hong and Bhamidimarri, 2012; Cole et al., 2014; Jeong et al., 2016; Kwak et al., 2016; Temizyurek and Dadaser-Celik, 2018; Abdi et al., 2021; Jiang et al., 2022), barometric pressure (Cole et al., 2014), landform attributes (Risley et al., 2003; DeWeber and Wagner, 2014; Topp et al., 2023; Souaissi et al., 2023), and many more (see Table S1).
>
> In the last few years, including the day-of-year as an input, DOY (Qiu et al., 2020; Heddam et al., 2022; Drainas et al., 2023; Rahmani et al., 2023) and humidity (Cole et al., 2014; Hong and Bhamidimarri, 2012; Kwak et al., 2016; Temizyurek and Dadaser-Celik, 2018; Abdi et al., 2021), have also shown to better capture the seasonal patterns of SWT (Qiu et al., 2020; Philippus et al., 2024). With improved access to remote sensing data, there has also been a notable increase of satellite products such as estimates of sky cover (Cole et al., 2014), solar radiation (Kwak et al., 2016; Topp et al., 2023; Majerska et al., 2024), sunshine per day (Drainas et al., 2023) and potential ET (Rozos, 2023; Topp et al., 2023). However, more research is needed to better understand the influence of newer model inputs on SWT (Zhu and Piotrowski, 2020).
>
> Recently, SWT studies focused on the CONUS-scale have chosen to use as many model inputs as available, with Wade et al. (2023), a point-scale CONUS ML study using over 20 variables, while Rahmani et al. (2023) created a LSTM model and considered over 30 variables to simulate SWT. Despite the use of diverse data, the models in these studies performed only satisfactorily and were deemed not generalizable, leaving much room for improvement in CONUS-scale modeling of SWT. With the compilation of larger and larger datasets, feature importance in ML, that is the process of using techniques to assign a score to model input features based on how good the features are at predicting a target variable, can be an efficient way to improve data comprehension, model performance, and model interpretability, the latter of which can dually serve as a transparency marker of which features are driving predictions. Methods for measuring feature importance include using correlation criteria (Pearson's r, Spearman's rho), permutation feature importance (shuffling feature values, measuring decrease in model performance), linear regression feature importance (larger absolute values indicate greater importance), or if using CART/RF/gradient boosting, entropy impurity measurements can be insightful (Venkateswarlu and Anmala, 2023).

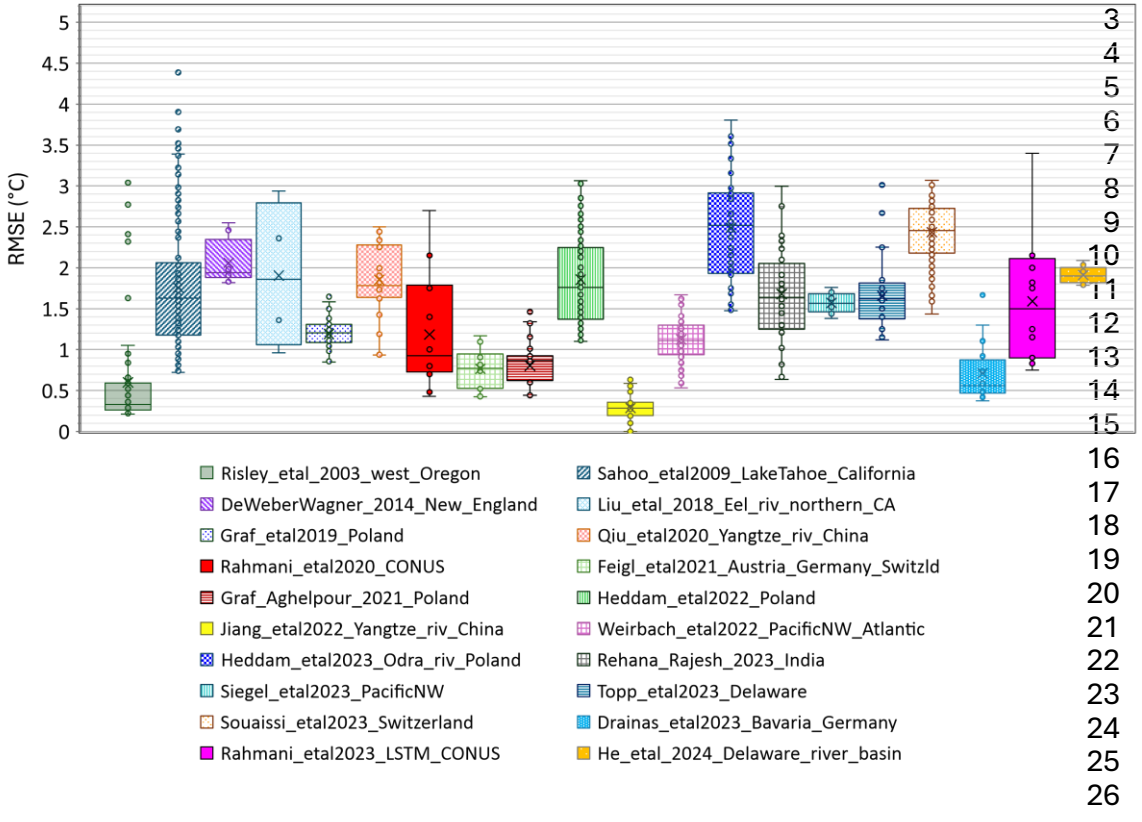
Moved part of section 2.3.1, original (lines 246-253) to section 2.4.2 Model Inputs for ML-SWT *(moved to lines 517-523):*

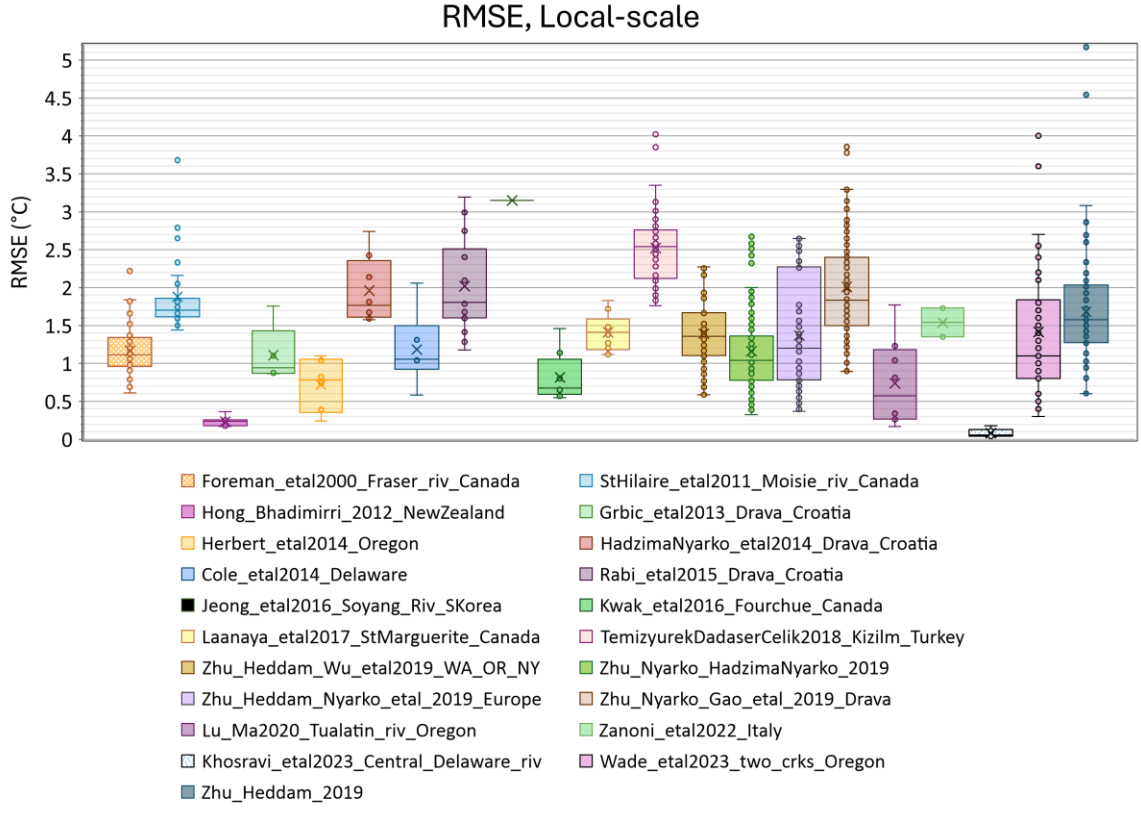
> For example, one technique that can be used to improve ML model parameter selection is the *Least Absolute Shrinkage and Selection Operator (LASSO)*, a regression technique used for feature selection (Tibshirani, 1996). Research utilizing ML models for SWT frequency analysis at ungaged basins used the LASSO method to select explanatory variables for two ML models (Souaissi et al., 2023). The LASSO method consists of a shrinkage process where the method penalizes coefficients of regression variables by minimizing them to zero (Tibshirani, 1996). The number of coefficients set to zero depends on the adjustment parameter, which controls the severity of the penalty. Thus, the method can perform both feature selection and parameter estimation, an advantage when examining large datasets (Xu & Liang, 2021).

5. 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.

AUTHOR RESPONSE: We appreciate the opportunity to clarify. Initially, we compiled a performance metric comparison by spatial scale for the most-cited metric, RMSE (42 papers cited) and plotted RMSE by study for regional/CONUS scale and local scale (located in HYDROSHARE repository but not the manuscript, see plots on next page for comparison). The comparison found minimal difference in RMSE between the regional/CONUS studies and the local scale studies, which we summarize in Table 1.

Given the speed at which ML is advancing and being applied for hydrologic applications, we do not think it wise to opinionate on which ML model is better or worse. Instead, our aim is to inform the reader on current studies and metrics, which we do by providing data as supplementary info, such as Tables S1 which states the time scale, spatial scale, region and time period considered of each study while Table S2 lists the data analysis techniques and/or ML algorithms used, as well as the training/validation/testing percentages/time periods as reported by the study. We think that summarizing publications (see Tables S1 and S2) and compiling performance metrics allows the reader to identify what has already been done in the ML-SWT field so that they can then make their own informed decisions depending on their research questions, model selection, project time frame, etc.


RMSE, Region/CONUS-scale
RMSE (°C)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Risley_etal_2003_west_Oregon
DeWeberWagner_2014_New_England
Graf_etal2019_Poland
Rahmani_etal2020_CONUS
Graf_Aghelpour_2021_Poland
Jiang_etal2022_Yangtze_riv_China
Heddam_etal2023_Odra_riv_Poland
Siegel_etal2023_PacificNW
Souaissi_etal2023_Switzerland
Rahmani_etal2023_LSTM_CONUS
Sahoo_etal2009_LakeTahoe_California
Liu_etal_2018_Eel_riv_northern_CA
Qiu_etal2020_Yangtze_riv_China
Feigl_etal2021_Austria_Germany_Switzld
Heddam_etal2022_Poland
Weirbach_etal2022_PacificNW_Atlantic
Rehana_Rajesh_2023_India
Topp_etal2023_Delaware
Drainas_etal2023_Bavaria_Germany
He_etal_2024_Delaware_river_basin


RMSE, Local-scale
RMSE (°C)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Foreman_etal2000_Fraser_riv_Canada
Hong_Bhadimirri_2012_NewZealand
Herbert_etal2014_Oregon
Cole_etal2014_Delaware
Jeong_etal2016_Soyang_Riv_SKorea
Laanaya_etal2017_StMarguerite_Canada
Zhu_Heddam_Wu_etal2019_WA_OR_NY
Zhu_Heddam_Nyarko_etal_2019_Europe
Lu_Ma2020_Tualatin_riv_Oregon
Khosravi_etal2023_Central_Delaware_riv
Zhu_Heddam_2019
StHilaire_etal2011_Moisie_riv_Canada
Grbic_etal2013_Drava_Croatia
HadzimaNyarko_etal2014_Drava_Croatia
Rabi_etal2015_Drava_Croatia
Kwak_etal2016_Fourchue_Canada
TemizyurekDadaserCelik2018_Kizilm_Turkey
Zhu_Nyarko_HadzimaNyarko_2019
Zhu_Nyarko_Gao_etal_2019_Drava
Zanoni_etal2022_Italy
Wade_etal2023_two_crks_Oregon

6. 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.

AUTHOR RESPONSE: We agree with the referee that the purpose of the review should be more clearly stated. We drafted this paper to serve as a middle ground between traditional modelers and more well-versed ML users. The intended audience are hydrologic modelers who have heard of AI/ML and want a summary of what has been done in SWT modeling using ML. Our dual objective is also for this to be a reference for assessment of ML performance. At the same time, we want ML researchers to be aware of where their models stand compared to other modelers while communicating that an "A+ grade" is actually more common (and therefore the new average) relative to what they are used to in hydrologic modeling. We have added a few sentences in the introduction, under section 1.2 'Study Objectives' of the manuscript to state who the intended audience is:

1.2. Study Objective (new in blue, *revised lines 64-68***)**

> *The current work includes an extensive literature review of studies that used ML algorithms/models for river/SWT modeling, hindcasting and forecasting.* The intent of this review is two-fold: 1) to introduce ML for hydrologists who have modeling experience and are interested in pursuing ML-use for their SWT studies, and 2) to provide a broad overview of machine learning applications in SWT. For ML experts, we think that this review could also prove useful as reference for how ML has been applied in the field of SWT modeling and where improvement is needed. Overall, this article aims to serve as a bridge between hydrologists and machine learning experts. *Our review includes papers cited by Zhu and Piotrowski (2020), who previously conducted a study of ANNs used in SWT modeling, however, we provide a comprehensive examination of peer-reviewed journals that use any type of artificial intelligence/ML algorithm to model or evaluate river/SWT […]*

7a. While the paper provides an extensive review of ML applications in 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 & 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 to lines 136 – 143 & lines 146 – 159 & lines 263 - 292.

AUTHOR RESPONSE: Thank you for the opportunity to clarify. We provided supplementary tables to summarize study information, for example, Tables S1 includes summarized information stating the time scale, spatial scale, region and time period considered of each study while Table S2 lists the data analysis techniques and/or ML algorithms used, as well as the training/validation/testing percentages/time periods as reported by the study. We think the "pros/cons" and "strengths/weakness" vary depending on the research goal and question, and the robustness of ML models allows them to cater to most problems, which is why rather than opinionating, we provide concrete specifications on the models used and allow the reader to decide based on their objectives.

7b. 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.

AUTHOR RESPONSE: We agree with the reviewer. We have revised the text to describe the advantages and disadvantages of K-nn, section 2.3.1.1 (crossed out is deleted, *revised lines 166-181*):

K-nearest neighbors (K-nn) is a ~~type of~~ versatile supervised ML algorithm (Fix & Hodges, 1952; Cover & Hart, 1967) used to solve nonparametric classification and regression problems. ~~It is one of the oldest algorithms (Fix & Hodges, 1952; Cover & Hart, 1967) considered within classical ML~~. The K-nn algorithm uses proximity between data points to make classifications or evaluations about the grouping of any given data point (Acito, 2023). K-nn gained popularity in the 2000s due to its simplicity in implementation and understanding, making it readily accessible to hydrologic researchers and practitioners. ~~While less used today, -~~For example, St.-Hilaire et al. (~~2012~~2011) used various K-nn model configurations to model SWT for the Moisie River in northern Quebec, Canada, finding that~~. T~~ the best K-nn model required prior-day SWT data and day-of-year (DOY), an indicator of seasonality ~~(St. Hilaire et al., 2011)~~. Advantages of K-nn include its non-assumptions of the underlying distribution of the data, allowing it to handle nonlinear complexities without requiring a solid model structure as is the case for some physical models (St. Hilaire et al., 2011). Disadvantages of K-nn are that it is computationally intensive, may require extensive cross-validation, performance can be affected by irrelevant/redundant features, and due to its high memory and computational needs, is impractical for large-scale applications, i.e., scalability issues, (Acito, 2023). For example, Heddam et al. (2022) For example, Heddam et al. (2022) ~~For five stream stations in Poland, Heddam et al. (2022)~~ compared K-nn with other ML algorithms, finding that K-nn was outperformed by other MLs such as least squares support vector machine and neural networks. ~~performed poorly compared to other ML algorithms.~~ The use of K-nn may still be reasonable for simple, local cases but we advise considering other MLs for more complex or larger-use cases.

7c. Lines 146 – 153 explains PCA & ck-means clustering on data reduction application, however, it is not clear under what conditions we can use them.

AUTHOR RESPONSE: We agree. We propose adding text to clarify:

Added lines 192-193:

Krishnaraj and Deka (2020) used *K-means* to organize spatial grouping for water quality monitoring stations for dry and wet regions along the Gangas River basin in India to identify whether pollution patterns could be discerned.

Added lines 198-203:

Using *PCA*, Krishnaraj and Deka (2020) found that certain water quality parameters (dissolved oxygen, sulfate, electrical conductivity) were more dominant in the dry season compared to the wet season (total dissolved solids, sodium, potassium, sodium, chlorine, chemical oxygen demand), data which could be used to cater the monitoring program to the important parameters. In their study, SWT was not a dominant parameter, likely in part because the SWT of large downstream rivers like the Gangas River are generally less variable due to their larger volume and stronger thermal buffer. ~~Used k-means and PCA in the Ganga River Basin of India to find spatiotemporal patterns of water quality parameters, including SWT.~~

7d. 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.

AUTHOR RESPONSE: We agree and added text on feature importance to a section on model inputs as suggested (please see comment #4 for full text).

The text specific to feature importance is below (*revised lines 506-516*):

Recently, SWT studies focused on the CONUS-scale have chosen to use as many model inputs

as available, with Wade et al. (2023), a point-scale CONUS ML study using over 20 variables, while Rahmani et al. (2023) created a LSTM model and considered over 30 variables to simulate SWT. Despite the use of diverse data, the models in these studies performed only satisfactorily and were deemed not generalizable, leaving much room for improvement in CONUS-scale modeling of SWT. With the compilation of larger and larger datasets, feature importance in ML, that is the process of using techniques to assign a score to model input features based on how good the features are at predicting a target variable, can be an efficient way to improve data comprehension, model performance, and model interpretability, the latter of which can dually serve as a transparency marker of which features are driving predictions. Methods for measuring feature importance include using correlation criteria (Pearson's *r*, Spearman's *rho*), permutation feature importance (shuffling feature values, measuring decrease in model performance), linear regression feature importance (larger absolute values indicate greater importance), or if using CART/RF/gradient boosting, entropy impurity measurements can be insightful (Venkateswarlu and Anmala, 2023).

7e. Lines 263 – 292 are organized in three paragraphs while providing general knowledge about ANNs with relatively less direct relations to water temperature application.

AUTHOR RESPONSE: We appreciate the reviewer's feedback and are open to making changes to improve the manuscript for the reader. Referee #3 made a similar comment about this section, and we now wonder if it would be better to provide the description of ANN variants and alternatives (lines 263-320) as part of an appendix. We think it would still be helpful to keep the information, but we also agree that it may be too extensive for the main text. In this way, the manuscript can be made more concise while also keeping the details as a section of the manuscript for anyone who is interested in reading further.

Following this line of thinking, we can point the reader to the appendix (*revised lines 326-327*):

"For more detail on traditional ANNs, with descriptions of ANN variants and backpropagation alternatives, we refer the reader to Appendix A."

We have added Appendix A after the conclusion (*revised lines 1014 – 1075*).

Minor corrections:

1. Line 13: There is a typo that changes the meaning of the sentence. It should be "… with in situ …" or "… with in-situ …".

AUTHOR RESPONSE: Thank you for pointing this out, we have fixed the typo to read "with in-situ" (*revised line 13*).

2. 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.

AUTHOR RESPONSE: Thank you! We have revised the text to read "long short-term memory" (*revised line 160*) and reviewed the text accordingly.

3. 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.

AUTHOR RESPONSE: Yes, we agree with both points. We have clarified that a feed-forward NN was used and revised the sentence to make it more accurate (*revised lines 263-266*):

> ~~In the case of~~ A SWT modeling study comparing the output of three model versions of DT, GPR, and feed-forward neural networks for ~~daily SWT modeling~~ multiple sites ~~and prediction~~, found that DTs ~~can~~ could perform similarly to GPR and feed-forward neural networks when detailed statistics of air temperature, day-of-year, and discharge were included ~~NNs~~ (Zhu, Nyarko, Hadzima-Nyarko, Heddam, et al., 2019).

4. Line 541: "at" is missed. It is .. All journals examined used at least …"

AUTHOR RESPONSE: Thank you! We have added the word "at" (*revised line 681*).

Referee #2 Comments

This is a meaningful manuscript that provides a thorough review of ML approaches for SWT modeling and their evaluation metrics. I believe that the current scientific community has indeed developed a broad understanding of the integration of ML into stream temperature modeling. Hence, while the manuscript presents a comprehensive overview, incorporating more in-depth insights could enhance its appeal to readers and significantly increase its contribution to the field. The review covers a wealth of content, including recent articles and other reviews, but the sections are somewhat loosely structured, with key points relatively briefly mentioned.

AUTHOR RESPONSE: We thank the referee for their time and feedback, we believe the manuscript is stronger as a result. We address specific referee comments below. For reference, we separated some referee comments into a, b, etc., to provide a more organized response. Proposed new/edited text is in BLUE. Revised lines in the track-changes manuscript are indicated by: *(revised lines XXX-XXX).*

1. For instance, in the first section (Overview: SWT Model Types), the author provides a solid overview of statistical, physical, and ML models. However, a more detailed analysis of the comparative strengths and weaknesses of physical and ML models would strengthen the discussion. The models are presented in a nearly linear developmental order in this review, but it would be beneficial to mention some points, for example, [if] physical models perform well, why ML models are adopted[?].

AUTHOR RESPONSE: The referee makes a good point with regards to the question of "if physical models perform well, why are ML models being adopted?". We have expanded the section 2.3 "Artificial Intelligence Models in SWT Modeling" to discuss this (*revised lines 136-156):*

> In the last decade, computing advances in AI have started to offer several advantages for using machine learning (ML) in hydrology that are comparable to physically based models (Cole et al., 2014; Zhu et al., 2019; Rehana and Rajesh, 2023). In contrast to traditional physically based models, the code underlying ML models are generally open-source and publicly available allowing for near real-time accessible advances and user feedback, whereas the source code for some physically based models may be inaccessible to the public due to being privately managed (MIKE suite of models) or the model software may be publicly available but take years to publish updates (USGS MODFLOW, Simunek's HYDRUS). One advantage that has made ML increasingly appealing includes its ability to learn directly from the data (i.e., data driven), which can be useful when the underlying physics are not fully understood or are considered too complex to model accurately.
>
> Additionally, ML models are more efficient in making predictions compared to the time-intensive solvers of physically based models. ML models can also handle the challenge of scalability, that is managing large datasets and seamlessly deploying across various computer platforms and applications (Rehana and Rajesh, 2023). Air2stream, a hybrid statistical-physically based SWT model (Toffolon and Piccolroaz, 2015; Piccolroaz et al., 2016), initially outperformed earlier ML models such as Gaussian Process Regression (Zhu et al., 2019). However, in the last few years, Air2stream has had its performance matched and even exceeded by recent neural networks models (Feigl et al., 2021; Rehana and Rajesh, 2023).
>
> Finally, with computer processing power improving and the emergent field of quantum computing, there is a strong belief that using ML and by extension AI, in science applications will drive innovation to the point where natural patterns and insights not currently apparent in physical modeling will be uncovered (Varadharajan et al., 2022). Thus, while physically based models are considered tried-and-true, thereby invaluable for their interpretability and grounding in established physics, ML models have the potential for growth – where they can be used to first complement and eventually lead as powerful tools for prediction, optimization, and understanding in increasingly complex and data-rich environments.

New citation:

Toffolon, M. and Piccolroaz, S., 2015. A hybrid model for river water temperature as a function of air temperature and discharge. *Environmental Research Letters*, *10*(11), p.114011.

2. How to gain the trust of traditional model users in ML methods? (This question is inherently challenging, as model users often have preferences based on their own familiarity with certain models and may exhibit biases against alternative approaches. However, it may be worthy to acknowledge this in the review.) This discussion could extend to the choice between different ML models as well, as conclusions favoring one model over another often depend on the specific context of the study. Many conclusions are applicable only under particular circumstances, so a generalization such as "a certain model is better suited to a particular type of problem" is more appropriate.

AUTHOR RESPONSE: We agree and appreciate the referee's feedback. We address this comment in our response to referee #1 for comment #3 (copied below) where we discuss how researchers can work to present their ML models as trustworthy. For this, we propose adding a new 'Discussion' subsection titled '4.4 Future Directions of SWT Modeling' *(revised lines 944-991)*:

The utility of ML in hydrologic modeling has advanced significantly, with interest seemingly growing exponentially (Nearing et al., 2021). With the novelty of ML, it is easy to over-value model performance and ignore the physics of the system, but with several decades of ML-experience, we advocate it is necessary to purposefully use ML to address physically-meaningful questions and not create ML for the sake of creating. Given this, Varadharajan et al. (2022) laid out an excellent discussion on opportunities for advancement of ML in water quality modeling, see section 3 of publication Varadharajan et al., (2022). Here we highlight some of the questions from Varadharajan et al. (2022) that can be considered in the context of what objectives the SWT community should be using in the ML era, namely: 1) How do we use physical knowledge (re: heat exchange equations, radiation influence) to improve models and process understanding? Rahmani et al. (2023) coupled NNs with the physical knowledge from SNTEMP, a one-dimensional stream temperature model that calculates the transfer of energy to or from a stream segment by either heat flux equations or advection, but found that even with SNTEMP, their flexible NNs exhibited substantial variance in prediction and needed to be constrained by further multi-dimensional assessments (Rahmani et al., 2023). In short, if our use of physics in machine learning makes our models worse, we should understand why.

A second question that needs addressing is 2) How do we deal with predictive uncertainty in ML used for SWT modeling? According to Moriasi et al. (2007), uncertainty analysis is the process of quantifying the level of confidence in any given model output based on five guidelines: 1) the quality and amount of observations (data), 2) the lack of observations due to poor or limited field monitoring, 3) the lack of knowledge of physical processes or operational procedures (instrumentation), 4) the approximation of our mathematical equations, and 5) the robustness of model sensitivity analysis and calibration. For example, in rainfall-runoff modeling, researchers have proposed benchmarking to examine uncertainty predictions of ML rainfall-runoff modeling (Klotz et al., 2022). For stream temperature modeling, researchers have attempted to address the role of uncertainty in deep learning model (RGCN, LSTM) prediction using the Monte Carlo Dropout (Zwart, Oliver, et al., 2023) and a unimodal mixture density network approach (Zwart, Diaz, et al., 2023).

Other questions that SWT-ML studies should consider is 3) How do we make ML models generalize better, specifically with regards to ungaged basins? And 4) How can ML models be improved to predict extremes? As ML models advance to use satellite data, include more sensor networks and/or couple with climate models, there is a logical next step toward creating generalizable models that can account for extremes. In our review, only two papers by the same group (Rahmani et al., 2020, 2023) conducted a CONUS-scale approach towards SWT-ML modeling, omitting hydrologically important regions in the southwest (CA) and southeast (FL). Recently, a satellite remote sensing paper used RF to model monthly stream temperature across the CONUS and tested for temporal (walk-forward validation), unseen and 'true' ungaged regions (Philippus et al., 2024). We have also learned that ML models such as LSTMs, generally only make predictions within the bounds of their training data (Kratzert et al., 2019), which is a limitation for predicting extremes. Thus, we

strongly urge the community to work towards ML models that generalize better and/or are more robust towards predictions of extremes.

Finally, 5) How can we build ML models such that they are seen as trustworthy and interpretable by the hydrologic community? To answer this question, we must address a technical barrier (black-box issues, data limitations, model uncertainty) and a social barrier (i.e., educated skepticism of ML due to novelty, little understanding of computer science basics and/or coding experience). If we are to incorporate ML into decision-making processes, it makes sense that ML must be transparent and understandable to more than just computer or data scientists (Varadharajan et al., 2022). For example, Topp et al. (2023) recently used explainable AI to elucidate how ML architectures affected the SWT model's spatial and temporal dependencies, and how that in turn affected the model's accuracy. Addressing this technical barrier can also be done by improving access to data, which has seen remarkable progress thanks to web repositories such as NSF-funded CUAHSI's Hydro share (CUAHSI, 2024) and GitHub (GitHub, 2024). In the United States, data access to state and locally-based data remains limited, and should be addressed. In terms of the social barrier, education about ML and ML-use is key. Societal interest in ML has thankfully also lead to a plethora of educational resources and ML walk-through videos and tutorials in Tensorflow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), and Google Colab (Bisong, 2019). With the speed at which ML-use is evolving, short communication pieces (Lapuschkin et al., 2019) and opinion pieces (Kratzert et al., 2024) with clear examples about an ML-issue and practical solutions will also help make ML challenges more transparent and therefore accessible to the hydrologic community-at-large.

3a. Furthermore, the author may not clearly (separately) present the generalization capabilities of ML models in temporal and spatial contexts, which is crucial for data split. The model ability of generalization over time is particularly meaningful for climate change studies, where overfitting (common for ML studies) may lead to highly unreliable projections. Spatial generalization is useful for applying models to new regions or watersheds (ungauged stream/river/watershed).

AUTHOR RESPONSE: We agree. Referee #1 made a similar comment (ref #1, comment #1A) about overfitting and having ML undergo more testing and we propose to address both comments by adding: 1) a new subsection 2.4.1, titled "Identifying model complexity", which discusses overfitting/underfitting, with 2) a diagram with initial steps to mitigate overfitting. The new text is below:

*new Section 2.4.1, Identifying Model Complexity *(revised lines 464-483)*

The strong success of ML-use in SWT modeling warrants a brief and broad overview on identifying model complexity to minimize overfitting and underfitting" of models. When a model is too complex, i.e., has too many features or parameters relative to the number of observations, or is forced to overextend its capabilities, i.e., make predictions with insufficient training data, the model runs the risk of overfitting (Srivastava et al., 2014). An overfitted model fits the training data "too well", capturing noise and details that provide high accuracy on a training dataset, only to perform poorly once the model encounters "unseen" data in testing/validation (Xu and Liang, 2021). Scenarios where overfitting may be temporarily acceptable are: 1) model development is at preliminary stages, the interest is in a "proof of life" concept, 2) when the objective is to identify heavily-relied on features by the model, i.e., feature importance, or 3) in highly-controlled modeling environments where the expected data will be consistently similar to the training dataset. The latter is more likely in industrial applications and unlikely in the changing nature of hydrology.

In contrast, underfitting occurs when a model is too simple to capture any patterns in the data, which can also lead to unsatisfactory performance in training, testing and validation. Underfitting can occur with inadequate model features, poor model complexity or when regularization techniques, (e.g., L1 or L2 regularization), are over-used, making the model too rigid and unable to respond to changes in the data. Given the propensity of ML models to effectively learn the training data, underfitting is less an issue in ML whereas overfitting can be widespread. In Figure 1, we present an

example workflow that researchers can use to transition away from overfitting and towards generalizability. In the five-step outline (Fig. 1), we suggest the need for "Temporal, Unseen, Ungaged Region Tests" (TUURTs), which is a call for temporal and spatially-focused testing that can be used to strengthen model robustness.

Revised lines 484-486:

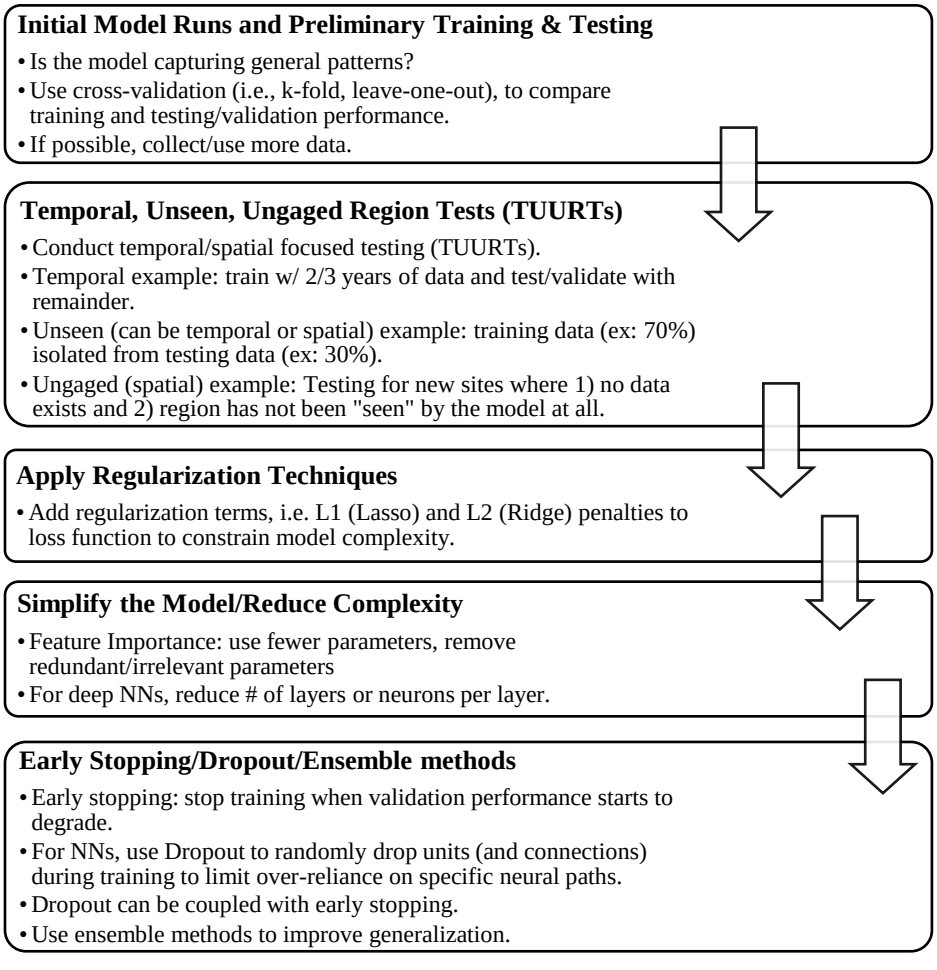


Figure 1. Diagram outlining steps that can be taken in modeling process to mitigate overfitting.

We propose to address the comment about generalization and a similar one made by ref #1 (comment # 3) by adding a new Discussion subsection, '*4.4 Future Directions of SWT Modeling'*. Below is our selected response where we state that models should work towards generalizability (*revised lines 966-975*). For full text, please see comment #2:

> Other questions that SWT-ML studies should consider are 3) How do we make ML models generalize better, specifically with regards to ungaged basins? And 4) How can ML models be improved to predict extremes? As ML models advance to use satellite data, include more sensor networks and/or couple with climate models, there is a logical next step toward creating generalizable models that can account for extremes. In our review, only two papers by the same group (Rahmani et al., 2020, 2023) conducted a CONUS-scale approach towards SWT-ML modeling, omitting hydrologically important regions in the southwest (CA) and southeast (FL). Recently, a satellite remote sensing paper used RF to model monthly stream temperature across the CONUS and tested for temporal (walk-forward validation), unseen and 'true' ungaged regions (Philippus et al., 2024). We have also learned that ML models such as LSTMs, generally only make predictions within the bounds of their training data (Kratzert et al., 2019), which is a limitation for predicting extremes. Thus, we strongly urge the community to work towards ML models that generalize better and/or are more

robust towards predictions of extremes.

3b. Additionally, the review does not systematically address the critical issue of model input selection, which is essential in ML modeling. Model inputs for SWT modeling may include hydrometeorological and physical parameters (or other attributes used in different studies), they play a role in model performance and should be discussed in this part.

AUTHOR RESPONSE: Thank you for pointing out this area in need of clarity. Referee #1, comment #4 had a similar question about model input, and we propose adding the paragraph below in response to both. Additionally, we want to note that we included in Supplementary Materials, Table S1, which contains some of the suggested data by the referee, such as: period considered, region examined, temporal resolution of SWT, spatial scale of study, and hydrometeorological parameters used for modeling.

*new subsection 2.4.2, Model Inputs for ML-SWT (*revised lines 488-516*):

Using air temperature (AT) to better understand SWT has been considered since the 1960s, when Ward (1963) and Edinger et al. (1968) discussed the influence of air temperature on SWT. Since then, studies have used varying input variables (see Table S1), however, the model inputs of AT and SWT continue to be the most used in ML-modeling studies. In particular, studies have used AT from time periods outside of the known SWT record to improve model performance (Sahoo et al., 2009; Piotrowski et al., 2015; Graf et al., 2019). In addition to AT and SWT, flow discharge has been used to attempt to constrain SWT (Foreman et al., 2001; Tao et al., 2008; St-Hilaire et al., 2011; Grbić et al., 2013; Piotrowski et al., 2015; Graf et al., 2019; Qiu et al., 2020). Traditionally-used model inputs include precipitation (Cole et al., 2014; Jeong et al., 2016; Rozos, 2023), wind direction/speed (Hong and Bhamidimarri, 2012; Cole et al., 2014; Jeong et al., 2016; Kwak et al., 2016; Temizyurek and Dadaser-Celik, 2018; Abdi et al., 2021; Jiang et al., 2022), barometric pressure (Cole et al., 2014), landform attributes (Risley et al., 2003; DeWeber and Wagner, 2014; Topp et al., 2023; Souaissi et al., 2023), and many more (see Table S1).

In the last few years, including the day-of-year as an input, DOY (Qiu et al., 2020; Heddam et al., 2022; Drainas et al., 2023; Rahmani et al., 2023) and humidity (Cole et al., 2014; Hong and Bhamidimarri, 2012; Kwak et al., 2016; Temizyurek and Dadaser-Celik, 2018; Abdi et al., 2021), have also shown to better capture the seasonal patterns of SWT (Qiu et al., 2020; Philippus et al., 2024). With improved access to remote sensing data, there has also been a notable increase of satellite products such as estimates of sky cover (Cole et al., 2014), solar radiation (Kwak et al., 2016; Topp et al., 2023; Majerska et al., 2024), sunshine per day (Drainas et al., 2023) and potential ET (Rozos, 2023; Topp et al., 2023). However, more research is needed to better understand the influence of newer model inputs on SWT (Zhu and Piotrowski, 2020).

Recently, SWT studies focused on the CONUS-scale have chosen to use as many model inputs as available, with Wade et al. (2023), a point-scale CONUS ML study using over 20 variables, while Rahmani et al. (2023) created a LSTM model and considered over 30 variables to simulate SWT. Despite the use of diverse data, the models in these studies performed only satisfactorily and were deemed not generalizable, leaving much room for improvement in CONUS-scale modeling of SWT. With the compilation of larger and larger datasets, feature importance in ML, that is the process of using techniques to assign a score to model input features based on how good the features are at predicting a target variable, can be an efficient way to improve data comprehension, model performance, and model interpretability, the latter of which can dually serve as a transparency marker of which features are driving predictions. Methods for measuring feature importance include using correlation criteria (Pearson's r, Spearman's rho), permutation feature importance (shuffling feature values, measuring decrease in model performance), linear regression feature importance (larger absolute values indicate greater importance), or if using CART/RF/gradient boosting, entropy impurity measurements can be insightful (Venkateswarlu and Anmala, 2023).

Moved part of section 2.3.1, original (lines 246-253) to section 2.4.2 Model Inputs for ML-SWT (moved to lines 517-523):

> For example, one technique that can be used to improve ML model parameter selection is the Least Absolute Shrinkage and Selection Operator (LASSO), a regression technique used for feature selection (Tibshirani, 1996). Research utilizing ML models for SWT frequency analysis at ungaged basins used the LASSO method to select explanatory variables for two ML models (Souaissi et al., 2023). The LASSO method consists of a shrinkage process where the method penalizes coefficients of regression variables by minimizing them to zero (Tibshirani, 1996). The number of coefficients set to zero depends on the adjustment parameter, which controls the severity of the penalty. Thus, the method can perform both feature selection and parameter estimation, an advantage when examining large datasets (Xu & Liang, 2021).

4. In the second section, the authors do an excellent job summarizing model evaluation metrics. However, considering that ML models are often optimized to achieve superior performance on these metrics, there is (always) a risk of overfitting. Thus, beyond focusing on metrics, the review should also highlight the importance of more rigorous evaluation to further assess generalization ability. For instance, if a SWT model is built to run climate change scenarios, additional testing and more rigorous designs are essential to evaluate the model's ability to generalize over time. For robust long-term predictions, the model is supposed to maintain robust predictive performance in completely unseen periods, rather than being limited to a specific temporal range.

AUTHOR RESPONSE: We agree. This comment has similar themes to our response to #3a regarding overfitting and highlighting the need for generalization, please see comment #3a for a full response.

For the comment regarding having ML undergo more rigorous testing, we propose adding the following discussion for more rigorous testing for MLs. We added a few sentences (blue is new) to the Discussion subsection 4.3 “ML as Knowledge Discovery” where we urge for TUURTs (Temporal, Unseen, Ungaged Region Tests)’ (*revised lines 914-925*):

> While it is understandable that not every ML-SWT paper aims to explain physical processes, the SWT community should agree on a baseline of tests that all ML-SWT models undergo to assess model robustness and transferability. Specifically, we urge use of TUURTs (Temporal, Unseen, Ungaged Region Tests) for future ML-SWT models as a helpful step towards better modeling practices, increased model transparency and robustness (Fig.1). As stated in figure 1, for TUURTs, testing for “unseen” cases means testing only within the developmental dataset, whereas testing for “ungaged” cases means testing for new sites that have no data and have not been previously seen by the model at all. Due to the difficulty of conducting spatial tests compared to temporal tests, few ML-SWT studies have applied one or two of the tests, and rarely all three (Topp et al., 2023; Hani et al., 2023, Souassi et al., 2023). For example, Siegel et al. (2023), a non-ML SWT paper, tested for ungaged regions and unseen data but did not perform a temporal test. To our knowledge, Philippus et al. (2024), appears to be the only published SWT-ML study that applied TUURTs with some success. We further encourage modelers to shift towards more generalizable models, which are in theory, more capable of performing well across diverse scenarios and datasets and stand to become increasingly important with the unpredictability of climate extremes.

Overall, this review is informative and well-researched, and with more refined organization and deeper exploration of these key issues, it could make a substantial contribution to the field of SWT research.

AUTHOR RESPONSE: Thank you! This would certainly not be possible without the insightful feedback from referees.

Referee #3 Comments

I believe that this manuscript is a very useful and extensive methods literature review regarding stream temperature modeling. I would recommend approval with minor revisions to provide additional details from the reviewed literature and correct minor writing aspects; I had no problem with the general structure/flow or quality.

AUTHOR RESPONSE: We thank the referee for their time and feedback, we believe the manuscript is better as a result. We address specific referee comments below. Proposed new/edited text is in BLUE. Revised lines in the track-changes manuscript are indicated by the statement: *(revised lines XXX-XXX).*

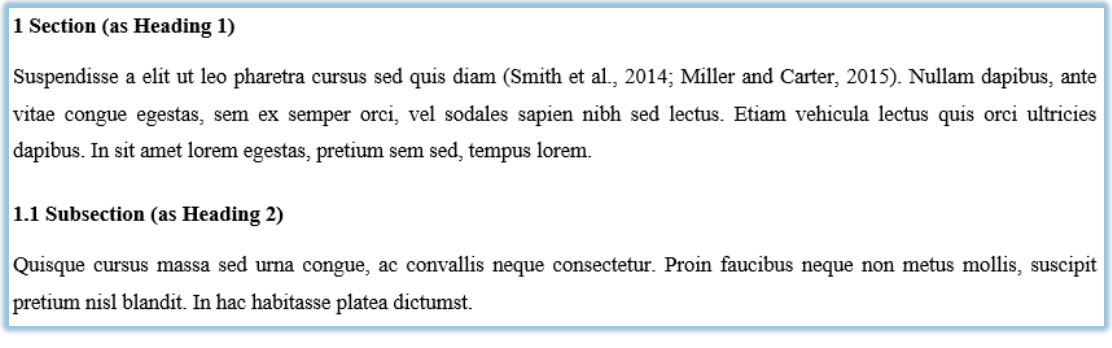
1. Section 2.3.3 ("Newer/recent ML algorithms") introduces RNNs, CNNs, and GNNs sufficiently, but it should probably give some description and reference to attention-based transformers. I am not aware of their application to SWT, but they are responsible for broader interest in ML (e.g., ChatGPT, which was cited earlier) and have had mixed success in hydrologic modeling. This class of models seems easily placed as a future direction.

AUTHOR RESPONSE: We agree. A literature search on Google Scholar at the end of 2024 found no publications specifically using attention-based transformers for SWT modeling, but we can add text about their potential to section 2.3.3 (*revised lines 456-462*):

> Attention-based transformers are a more novel type of deep learning that has led to advancements in natural language processing, in the form of ChatGPT, Microsoft's CoPilot, Google's Gemini and others. Due to their exponential success in the last few years, attention-based transformer models have been used in geological science fields such as oceanography for sea surface temperature prediction (Shi et al., 2024), hydrology for streamflow and runoff prediction (Ghobadi and Kang, 2022; Wei, 2023) and remote sensing for streambed land use change classification (Bansal and Tripathi, 2024). As a relatively new AI tool, attention-based transformers have yet to be used for SWT (to our knowledge), but their applications in other geological science fields suggest it is only a matter of time before their use is observed in SWT modeling.

2. There are some examples of unusual subsection and paragraph formatting. For example, section 1.1 is one paragraph which is approximately 1 page long. It seems that this is excessively large for one paragraph and that a named subsection should perhaps be more than just one (regularly sized) paragraph. Line 201 has another approximately 1-page-long paragraph, this area might be better organized with another level of subsections rather than fitting the more extensive references of decision trees into 1 paragraph.

AUTHOR RESPONSE: Thank you for pointing this out. We have consulted other published HESS articles and it appears that the first paragraph of a section is not indented but the subsequent paragraphs are. We have revised the manuscript to follow the Copernicus manuscript template (screenshot below).



1 Section (as Heading 1)

Suspendisse a elit ut leo pharetra cursus sed quis diam (Smith et al., 2014; Miller and Carter, 2015). Nullam dapibus, ante vitae congue egestas, sem ex semper orci, vel sodales sapien nibh sed lectus. Etiam vehicula lectus quis orci ultricies dapibus. In sit amet lorem egestas, pretium sem sed, tempus lorem.

1.1 Subsection (as Heading 2)

Quisque cursus massa sed urna congue, ac convallis neque consectetur. Proin faucibus neque non metus mollis, suscipit pretium nisl blandit. In hac habitasse platea dictumst.

At the referee's suggestion, we can add subsections to section 2.3.1 to distinguish algorithms as follows:

2.3.1.1 K-nearest neighbors (starts line 165),
2.3.1.2 Cluster analysis and variants (line 182),
2.3.1.3 Support vector machine and regression (line 204),
2.3.1.4 Gaussian Process Regression (line 234),
2.3.1.5 Decision trees and Classification and Regression Trees (line 255),
2.3.1.6 Random Forests and XGBoost (line 272)

We propose making section 2.3.1.6 (original lines 226-253) more concise given the newly separate Model Inputs for ML-SWT section (new *section 2.4.2*). Below is the revised text for *section 2.3.1.6* (blue is new, *revised lines 284-313*):

> RF and XGBoost, have been used to predict~~for~~ daily SWT prediction ~~in 10~~for Austrian catchments, ~~Results~~ with results showing ~~ed~~ minor differences in model performance, with a median RMSE difference of 0.08 °C between tested ML models (Feigl et al., 2021). Using RF and XGBoost along with four other ML models, Jiang et al. (2022) ~~tested the performance of six ML models in~~ estimated~~ing~~ daily SWT below dams in China, finding~~. They found that~~ day of year, ~~was most influential for the prediction of SWT, followed by~~ stream flow flux and AT to be most influential in the prediction of SWT (Jiang et al., 2022). Weierbach et al. (2022) used XGBoost and SVR to predict SWT at monthly time scales for the Pacific Northwest region of the U.S., ~~finding~~ showing that an ensemble XGBoost outperformed all modeling configurations for spatiotemporal predictions in unmonitored basins, ~~In contrast to Jiang et al. (2022), Weierbach et al. (2022) found~~with AT identified as the primary driver of monthly SWT. ~~for all 78 sites in the Pacific Northwest region of the U.S. (which included areas affected by dams), followed by month of year and solar radiation.~~ Zanoni et al. (2022) used RF and a deep learning model to develop regional models of SWT and other water quality parameters, ~~finding that~~with RF performance ~~was~~ comparitively less effective at detecting non-linear relationships than ~~to~~ the deep learning model, though both models identified ~~They found~~ AT ~~to be~~as most influential~~, with day of the year, and year of observation as possible replacements where AT was not available~~ (Zanoni et al., 2022).
>
> Souassi et al. (2023) tested the performance of ~~two ML models~~, RF and XGBoost, with non-parametric models for the regional estimation of maximum SWT at ungaged locations in Switzerland, finding no significant differences between the ML performance and the non-parametric model performances, which was attributed to the lack of a large dataset ~~as required by the ML models~~. Hani et al. (2023) used four supervised ML models – MARS, GAM, SVM, and RF to model potential thermal refuge area (PTRA) at an hourly timestep for two tributary confluences of the Sainte-Marguerite River in Canada. RF had the highest accuracy at both locations in terms of hourly PTRA estimates and modeling SWT (Hani et al., 2023). Wade et al. (2023) conducted a CONUS-scale study using 410 USGS sites with four years of daily SWT and discharge to examine maximum SWT. They used RF to estimate max SWT and thermal sensitivity (Wade et al., 2023), finding that AT was the most influential control followed by other properties (watershed characteristics, hydrology, anthropogenic impact).

3. There is an extensive background of traditional ANNs (2.3.2) which is debatably too extensive given the description of ANN variants and backpropagation alternatives (e.g., lines 284-320), which are relatively niche and rare. The content already exists and is not wrong, but if length were a concern, I would reduce this area.

AUTHOR RESPONSE: We appreciate the reviewer's feedback and propose making changes to improve the manuscript for readability. Referee #1 made a similar comment about this section, and we propose

moving the description of ANN variants and alternatives (lines 263-320) to Appendix A. We think it would still be helpful to keep the ANN information, while also agreeing that it may be too extensive for the main text. In this way, the manuscript can be made more concise while also keeping the details as a section of the manuscript for anyone who is interested in reading further. Following this line of thinking, we removed lines 322-379 from the main text and moved to Appendix A. We provide the following to note the appendix (*revised lines 326-327*):

> "For more detail on traditional ANNs, with descriptions of ANN variants and backpropagation alternatives, we refer the reader to appendix A."

We have added Appendix A after the conclusion (*revised lines 1015 – 1075*).

4. This work does not address predictive uncertainty, or the lack thereof associated with the ML literature review. I think that would be a worthwhile addition because I suspect most efforts lack that (e.g., referring to https://doi.org/10.5194/hess-26-1673-2022). A counterexample to the lack of uncertainty quantification, which may also be relevant to section 2.5, could be work led by Jacob Zwart focusing on SWT for reservoir operations (thermal releases). Examples being https://doi.org/10.1111/1752-1688.13093 or https://doi.org/10.3389/frwa.2023.1184992

AUTHOR RESPONSE: We appreciate the referee's insight in bringing these publications to our attention. Based on their relevancy, we have added Klotz et al. 2022, Zwart et al. 2023a and 2023b and included their RMSE values in our review. First, we added text on predictive uncertainty in the new 'Discussion' subsection titled, '4.4 *Future Directions of SWT Modeling'*, which also addresses ref #1, comment #3, (revised lines 944-991):

> The utility of ML in hydrologic modeling has advanced significantly, with interest seemingly growing exponentially (Nearing et al., 2021). With the novelty of ML, it is easy to over-value model performance and ignore the physics of the system , but with several decades of ML-experience, we advocate it is necessary to purposefully use ML to address physically-meaningful questions and not just create ML for the sake of creating. Given this, Varadharajan et al. (2022) laid out an excellent discussion on opportunities for advancement of ML in water quality modeling, see section 3 of publication of Varadharajan et al. (2022). (Varadharajan et al., 2022)Here we highlight some of the questions from Varadharajan et al. (2022) that can be considered in the context of what objectives the SWT community should be using in the ML era, namely: 1) How do we use physical knowledge (re: heat exchange equations, radiation influence) to improve models and process understanding? Rahmani et al. (2023) coupled NNs with the physical knowledge from SNTEMP, a one-dimensional stream temperature model that calculates the transfer of energy to or from a stream segment by either heat flux equations or advection, but found that even with SNTEMP, their flexible NNs exhibited substantial variance in prediction and needed to be constrained by further multi-dimensional assessments (Rahmani et al., 2023). In short, if our use of physics in machine learning makes our models worse, we should understand why.
>
> A second question that needs addressing is 2) How do we deal with predictive uncertainty in ML used for SWT modeling? According to Moriasi et al. (2007), uncertainty analysis is the process of quantifying the level of confidence in any given model output based on five guidelines: 1) the quality and amount of observations (data), 2) the lack of observations due to poor or limited field monitoring, 3) the lack of knowledge of physical processes or operational procedures (instrumentation), 4) the approximation of our mathematical equations, and 5) the robustness of model sensitivity analysis and calibration. For example, in rainfall-runoff modeling, researchers have proposed benchmarking to examine uncertainty predictions of ML rainfall-runoff modeling (Klotz et al., 2022). For stream temperature modeling, researchers have attempted to address the role of uncertainty in deep learning model (RGCN, LSTM) predictions using the Monte Carlo Dropout (Zwart, Oliver, et al., 2023) and a unimodal mixture density network approach (Zwart, Diaz, et al., 2023).

Other questions that SWT-ML studies should consider is 3) How do we make ML models generalize better, specifically with regards to ungaged basins? And 4) How can ML models be improved to predict extremes? As ML models advance to use satellite data, include more sensor networks and/or couple with climate models, there is a logical next step toward creating generalizable models that can account for extremes. In our review, only two papers by the same group (Rahmani et al., 2020, 2023) conducted a CONUS-scale approach towards SWT-ML modeling, omitting hydrologically important regions in the southwest (CA) and southeast (FL). Recently, a satellite remote sensing paper used RF to model monthly stream temperature across the CONUS and tested for temporal (walk-forward validation), unseen and 'true' ungaged regions (Philippus et al., 2024). We have also learned that ML models such as LSTMs, generally only make predictions within the bounds of their training data (Kratzert et al., 2019), which is a limitation for predicting extremes. Thus, we strongly urge the community to work towards ML models that generalize better and/or are more robust towards predictions of extremes.

Finally, 5) How can we build ML models such that they are seen as trustworthy and interpretable by the hydrologic community? To answer this question, we must address a technical barrier (black-box issues, data limitations, model uncertainty) and a social barrier (i.e., educated skepticism of ML due to novelty, little understanding of computer science basics and/or coding experience). If we are to incorporate ML into decision-making processes, it makes sense that ML must be transparent and understandable to more than just computer or data scientists (Varadharajan et al., 2022). For example, Topp et al. (2023) recently used explainable AI to elucidate how ML architectures affected the SWT model's spatial and temporal dependencies, and how that in turn affected the model's accuracy. Addressing this technical barrier can also be done by improving access to data, which has seen remarkable progress thanks to web repositories such as NSF-funded CUAHSI's Hydro share (CUAHSI, 2024) and GitHub (GitHub, 2024). In the United States, data access to state and locally-based data remains limited, and should be addressed. In terms of the social barrier, education about ML and ML-use is key. Societal interest in ML has thankfully also lead to a plethora of educational resources and ML walk-through videos and tutorials in Tensorflow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), and Google Colab (Bisong, 2019). With the speed at which ML-use is evolving, short communication pieces (Lapuschkin et al., 2019) and opinion pieces (Kratzert et al., 2024) with clear examples about an ML-issue and practical solutions will also help make ML challenges more transparent and therefore accessible to the hydrologic community-at-large.

Added citations used for new subsection, 4.4 Future Directions of SWT Modeling:

1) Apaydin, H., Taghi Sattari, M., Falsafian, K., and Prasad, R.: Artificial intelligence modelling integrated with Singular Spectral analysis and Seasonal-Trend decomposition using Loess approaches for streamflow predictions, Journal of Hydrology, 600, 126506, https://doi.org/10.1016/j.jhydrol.2021.126506, 2021.

2) Baydaroğlu, Ö. and Demir, I.: Temporal and spatial satellite data augmentation for deep learning-based rainfall nowcasting, Journal of Hydroinformatics, 26, 589–607, https://doi.org/10.2166/hydro.2024.235, 2024.

3) CUAHSI. 2024. Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Water Data Portal: https://www.cuahsi.org/community/water-data-portals, last access: 13 November 2024.

4) Kratzert, F., Gauch, M., Klotz, D. and Nearing, G., 2024. HESS Opinions: Never train an LSTM on a single basin. Hydrology and Earth System Sciences Discussions, 2024, pp.1-19.

5) Kwak, J., St-Hilaire, A., and Chebana, F.: A comparative study for water temperature modelling in a small basin, the Fourchue River, Quebec, Canada, Hydrological Sciences Journal, 1–12, https://doi.org/10.1080/02626667.2016.1174334, 2016.

6) Philippus, D., Sytsma, A., Rust, A., and Hogue, T. S.: A machine learning model for estimating the temperature of small rivers using satellite-based spatial data, Remote Sensing of Environment, 311, 114271, https://doi.org/10.1016/j.rse.2024.114271, 2024.

7) Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

8) Skoulikaris, C., Venetsanou, P., Lazoglou, G., Anagnostopoulou, C., and Voudouris, K.: Spatio-Temporal Interpolation and Bias Correction Ordering Analysis for Hydrological Simulations: An Assessment on a Mountainous River Basin, Water, 14, 660, https://doi.org/10.3390/w14040660, 2022.

9) Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15, 30, 2014.

10) Yang, M., Yang, Q., Shao, J., Wang, G., and Zhang, W.: A new few-shot learning model for runoff prediction: Demonstration in two data scarce regions, Environmental Modelling & Software, 162, 105659, https://doi.org/10.1016/j.envsoft.2023.105659, 2023.

11) GitHub. 2024. About Git and Github: https://docs.github.com/en/get-started/start-your-journey/about-github-and-git, last access: 14 November 2024.

12) Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), p.1096.

13) Zwart, J.A., Oliver, S.K., Watkins, W.D., Sadler, J.M., Appling, A.P., Corson-Dosch, H.R., Jia, X., Kumar, V. and Read, J.S., 2023. Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. JAWRA Journal of the American Water Resources Association, 59(2), pp.317-337.

14) Zwart, J.A., Diaz, J., Hamshaw, S., Oliver, S., Ross, J.C., Sleckman, M., Appling, A.P., Corson-Dosch, H., Jia, X., Read, J. and Sadler, J., 2023. Evaluating deep learning architecture and data assimilation for improving water temperature forecasts at unmonitored locations. *Frontiers in Water*, *5*, p.1184992.

15) Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S. and Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrology and Earth System Sciences, 26(6), pp.1673-1693.

16) M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015. TensorFlow. Website: https://www.tensorflow.org/

17) A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Website: https://arxiv.org/abs/1912.01703

18) Bisong, E. (2019). Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA. Website: https://doi.org/10.1007/978-1-4842-4470-8_7

We have also added text to section 2.5 Decision Support with the provided citations (*revised lines 663-668*):

> Further focusing on the Delaware River Basin, Zwart, Oliver, et al. (2023) used data assimilation and an LSTM to generate 1-day and 7-day forecasts of daily maximum SWT for the purpose of aiding reservoir managers in decisions about when to release water to cool streams. Following up on this study was Zwart, Diaz, et al. (2023), who used a LSTM and a RGCN, to generate 7-day forecasts of daily maximum SWT for monitored and unmonitored locations in the Delaware River Basin. The study found that the RGCN with data assimilation performed best for ungaged locations and for higher SWT, which can serve as valuable information for reservoir operators to consider while drafting release schedules.

5. In section 3 (e.g., 3.1, 3.3, 3.4), I would recommend adding some discussion regarding the equivalence or lack of between lower-case r and r-squared, upper-case R-squared, and NSE. I am very comfortable

stating that for the purpose of this continuously valued model evaluation, upper case R-squared and NSE are equivalent, but I am less comfortable making the assertion that lower case r and r-squared are (in all the papers reporting this value). This is likely further complicated by the reviewed literature using the lower-case r-squared and R-squared interchangeably, but given the 0-1 range, the high value skew, and the special case/conditional equivalences, I believe these values should all be reported together to characterize goodness of fit – especially that upper case R-squared and NSE should not be separated.

AUTHOR RESPONSE: We agree. We propose the following to address the referee's comments:

- Revise section 3.1 text to distinguish between lower-case *r*, r-squared r^2, and upper-case R^2 (*revised lines 697-710*):

> Pearson's r, also known as the correlation coefficient, is useful for determining the strength and direction (i.e., positive, negative) of a simple linear relationship (Helsel and Hirsch, 2002). Values of r, range from -1 to +1, where r < 0 indicates a negative correlation and r > 0 indicates a positive correlation (Legates and McCabe, 1999). The square of r is denoted as ~~R²~~ r^2, ~~or~~ known as the square of the correlation coefficient, with values of r^2 ranging from 0 to 1. The r^2 metric is commonly used in simple linear regression to assess the goodness of fit by ~~of determination, which represents~~ measuring the fraction of the variance in one variable (i.e., observations) that can be explained by the other variable (i.e., predictors). The metric r^2 tends to be confused with R^2, the latter which is a statistical measure that represents the proportion of variance explained by the independent variable(s) in a multiple linear regression model (Helsel and Hirsch, 2002). Part of the confusion may be related to the fact that R^2 shares the same range of ~~from~~ 0 to 1, with R^2 = 1 ~~suggesting~~ indicating that the model can explain all the variance, and vice versa. We note that while both r^2 and R^2 share similarities in that they measure the proportion of variance, R^2 is more commonly used for multiple linear regression context, while r^2 is best suited for simple linear regressions. To reduce confusion, we strongly suggest that *r*, r^2 and R^2 always be reported together (even if as a supplement to a manuscript) to characterize goodness-of-fit. ~~The *r* and R^2 metrics are typically used for normally distributed data that follows a bivariate normal distribution (Helsel and Hirsch, 2002).~~

- Add text stating that upper R^2 and NSE should always be provided together in section 3.4:

1st paragraph, added after 1st sentence (*revised lines 785-789*):

> Having reviewed the literature and in agreement with previous published recommendations (Moriasi et al., 2007), we recommend that a combination of standard regression (i.e., *r*, r^2, R^2), dimensionless (i.e., NSE), and error index statistics (i.e., RMSE, MAE, PBIAS) be used for model evaluation and reported together in future publications.

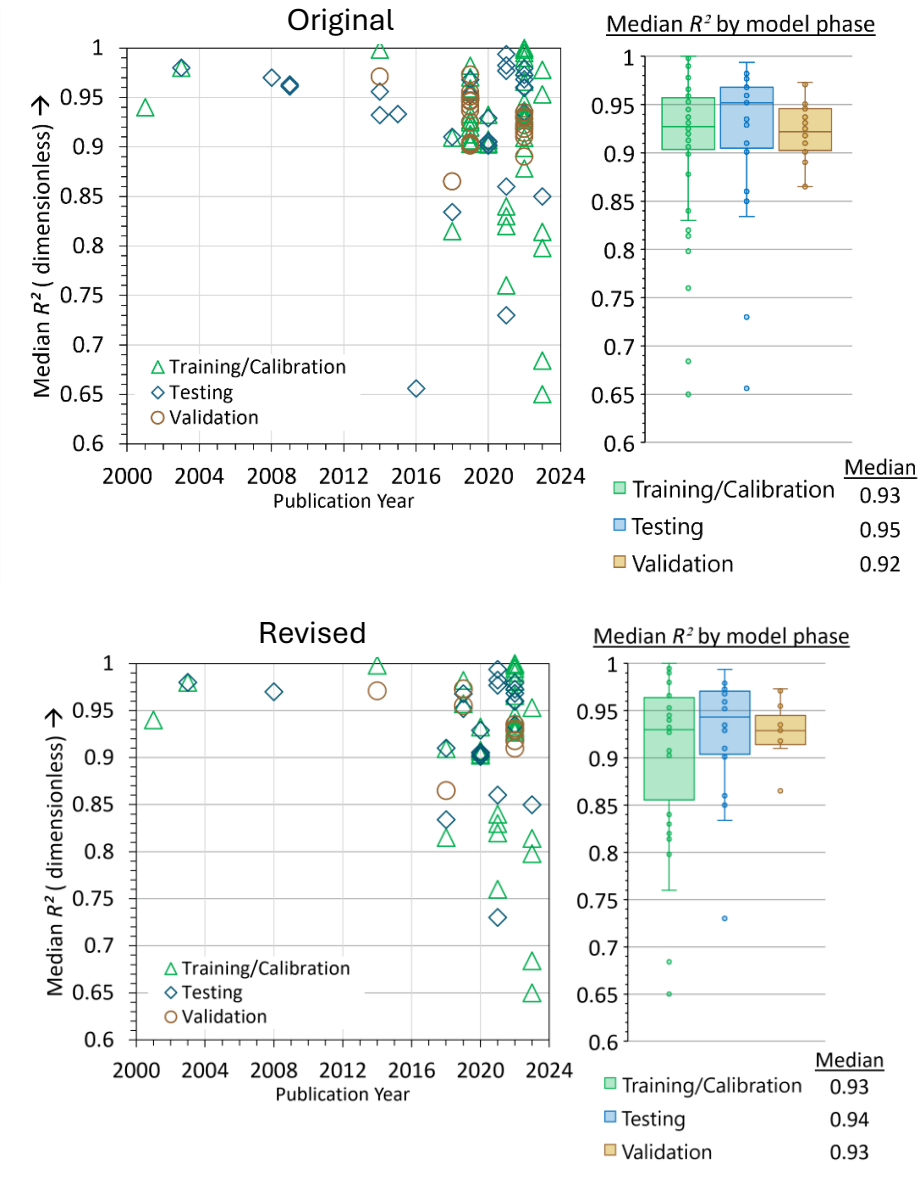
2nd paragraph, remove the statement about r and R2 (*revised lines 792-793*):

> ~~We note that for the 11 studies that used Pearson's r (see Table S1), and given that r and R2 are directly related, we converted r to R2 for ease of comparison on fig. 1.~~

3rd paragraph, added last sentence (*revised lines 803-805*):

> Overall, these complimentary metrics should always be reported together as they provide a broader evaluation of model performance, i.e., NSE measures a model's predictive skill and error variance, while R^2 assesses how well the model explains the variability of the data.

- In section 3.4, remove all r^2 values from Figure 1, only R^2 citations (17) remain. The median R^2 for training stayed the same (0.93), while the testing R^2 went from 0.95 to 0.94, and the validation R^2 went from 0.92 to 0.93. Overall, changes were insignificant. Below is a screenshot of the "Original (top)" and "Revised (bottom)" Figure 1 for reference.



6. In line 761, it feels controversial and a step too far to say ML models should be held to a higher standard. It feels less problematic to apply these higher, seemingly attainable standards to all SWT models. For example, a physics-based model is not "very good" by virtue of being a physics-based model, instead it is the same "satisfactory" label because its physics are not sufficient or accurate enough to do what the ML models can.

AUTHOR RESPONSE: We appreciate the referee's point of view. Perhaps instead of saying "separate, higher standard", we can say "additional standards" (see *revised line 1002*), but we think that additional standards are warranted nonetheless, not only in terms of performance metrics but also to improve model transparency, eradicate black-box confusion and encourage user confidence. We disagree that a physics-based model should be in the same "satisfactory" performance metric category because the intention of performance metrics is to identify what fits the data best (which data-driven ML excel at), whereas the general intention of physics-based models is to adhere to whatever governing equations have been employed. Our review indicates that we have been somewhat blinded by the excellence of ML performance metrics relative to physics-based and statistically-based models, and more awareness is needed moving forward.

7. If possible, in addition to considering spatial extents and temporal resolution of the papers, it would be interesting to know the aggregation level of data - if that is reported and what all the possibilities are. For example, individual gages with input data collected at the same gage location in situ, remotely sensed data subset to the drainage area for the reach that a gage is on. Are any works modeling dense transects along a river or modeling raster grid cells up and across a river (i.e., the 2D surface area), etc.

AUTHOR RESPONSE: Thank you for the opportunity to clarify. We provided supplementary table S1 to summarize study information regarding time period, temporal resolution, spatial resolution and hydrometeorological parameters considered by the cited studies. Responding to your comment, in our review, we saw that the aggregation level of data is more often than not, left unreported and unclear by studies (and reporting is not mandatory as a lot of data is pre-processed before utilization in modeling, adding to transparency questions). We do think discerning all the possibilities of data aggregation could make for an interesting follow-up study for the larger hydrologic community, which could focus solely on data manipulation, processing and augmentation for ML.

Additional literature to consider. Not necessary

8. The paragraph at line 385 related to process guidance prompted me to recommend https://doi.org/10.1029/2023WR035327 as very relevant. The reference is concerned with comparing different hybrid ML methods for SWT modeling to represent groundwater processes which aren't as represented here (e.g., relative to reservoir influence/reservoir adjacent modeling).

AUTHOR RESPONSE: Thank you for the suggestion, we agree that the challenge of including groundwater influence in SWT modeling warrants more research. We want to clarify that we did not include this reference as it appears to be a conference paper and not subjected to journal standards of peer review. That being said, the authors of the suggested manuscript went on to publish similar work in Water Resources Research, which we cite in this review (Topp et al., 2023).

9. In section 4.2, https://doi.org/10.1029/2020WR028091 may be a very relevant addition in-line with the author's narrative.

AUTHOR RESPONSE: Thank you for the suggestion, we enjoyed reading it and think it insightful. We added it to new 'Discussion' subsection, titled '4.4 Future Directions of SWT Modeling', in the first sentence (please see our response to ref #1, comment #4 for the full text), *revised lines 944-945*:

> "The utility of ML in hydrologic modeling has come a long way, with interest seemingly growing exponentially (Nearing et al., 2021)."

Minor writing comments:

1.The sentence beginning on line 51 perhaps uses too bold language when stating "AI … create reasonable choices". Many users of AI and scientists have concerns regarding the reasonableness of AI. Maybe it would be more accurate to further connect with the latter part of that sentence and say that "AI … learn optimal patterns to meet stated objectives" (which may or may not be broadly reasonable)

AUTHOR RESPONSE: That is a good point. Reasonableness is fluid. We agree with the referee and have updated the sentence as follows (*revised lines 51-53*):

> "Artificial intelligence (AI) describes technologies that can incorporate and assess inputs from an environment, ~~create reasonable choices,~~ learn optimal patterns and implement actions to meet stated

objectives or performance metrics (Xu & Liang, 2021; Varadharajan et al., 2022)."

2. Starting at line 131, "We define newer ML as those introduced in hydrologic modeling in the few years," perhaps this should say "in recent years"?

AUTHOR RESPONSE: We agree, thank you for the suggestion, we have updated the text to say, "in recent years" (*revised line 159*).

3. At line 380, although it can be inferred, "WNN" is never explicitly defined.

AUTHOR RESPONSE: Thank you for catching that, we have defined the acronym (*revised line 445*).

4. At line 541, "all journals examined used least one", perhaps this should say, "at least one"

AUTHOR RESPONSE: Thank you! We have added the word "at" (*revised line 681*).

5. By typo/mistake, it appears that two subsections in section 3 are titled "Model Performance Metrics: Error Indices"

AUTHOR RESPONSE: Yes, thank you for catching this error. Subsection 3.3 should have said "Dimensionless" because the subsection summarizes dimensionless metrics. We have updated the subsection header accordingly (*title, revised line 739*).

6. At line 610, there is a typo claiming an upper bound of -1

AUTHOR RESPONSE: Yes, that was a typo. Thank you for catching that, we have updated the text to just say "0 to 1" (*revised line 759*).

7. I have the benefit of reviewing 3rd, so I read the other reviewer's comments after making my own. I agree that a characterization of the validation and test sets used would be very beneficial (e.g., spatial, temporal, spatiotemporal exclusion, etc.), but I believe the concerns of overfitting are potentially overstated by the other reviewers given that this manuscript reports train, validation, and test set metrics (and the very strong agreement between the three).

AUTHOR RESPONSE: Thank you for your time and energy in reviewing this manuscript. With regards to the concerns of overfitting, we include below our response to referee #1, comment #1A. We think that the referee comment with regard to "characterization of the validation and test sets" is related to referee #1, comment #1B, which we also include below:

New subsection 2.4.1 Identifying Model Complexity (*revised lines 464-483*):

> The strong success of ML-use in SWT modeling warrants a brief and broad overview on identifying model complexity to minimize overfitting and underfitting" of models. When a model is too complex, i.e., has too many features or parameters relative to the number of observations, or is forced to overextend its capabilities, i.e., make predictions with insufficient training data, the model runs the risk of overfitting (Srivastava et al., 2014). An overfitted model fits the training data "too well", capturing noise and details that provide high accuracy on a training dataset, only to perform poorly once the model encounters "unseen" data in testing/validation (Xu and Liang, 2021). Scenarios where

overfitting may be temporarily acceptable are: 1) model development is at preliminary stages, the interest is in a "proof of life" concept, 2) when the objective is to identify heavily-relied on features by the model, i.e., feature importance, or 3) in highly-controlled modeling environments where the expected data will be consistently similar to the training dataset. The latter is more likely in industrial applications and unlikely in the changing nature of hydrology.

In contrast, underfitting occurs when a model is too simple to capture any patterns in the data, which can also lead to unsatisfactory performance in training, testing and validation. Underfitting can occur with inadequate model features, poor model complexity or when regularization techniques, (e.g., L1 or L2 regularization), are over-used, making the model too rigid and unable to respond to changes in the data. Given the propensity of ML models to effectively learn the training data, underfitting is less of an issue in ML whereas overfitting can be widespread. In Figure 1, we present an example workflow that researchers can use to transition away from overfitting and towards generalizability. In the five-step outline (Fig. 1), we suggest the need for "Temporal, Unseen, Ungaged Region Tests" (TUURTs), which is a call for temporal and spatially-focused testing that can be used to strengthen model robustness.

Response to ref #1, comment #1B: We have added a few sentences (blue is new) to the Discussion subsection 4.3, "ML Use for Knowledge Discovery" where we further urge for the use of TUURTs (Temporal, Unseen, Ungaged Region Tests)' *(revised lines 914-925):*

While it is understandable that not every ML-SWT paper aims to explain physical processes, the SWT community should agree on a baseline of tests that all ML-SWT models undergo to assess model robustness and transferability. Specifically, we urge use of TUURTs (Temporal, Unseen, Ungaged Region Tests) for future ML-SWT models as a helpful step towards better modeling practices, increased model transparency and robustness (Fig.1). As stated in figure 1, for TUURTs, testing for "unseen" cases means testing only within the developmental dataset, whereas testing for "ungaged" cases means testing for new sites that have no data and have not been previously seen by the model at all. Due to the difficulty of conducting spatial tests compared to temporal tests, few ML-SWT studies have applied one or two of the tests, and rarely all three (Topp et al., 2023; Hani et al., 2023, Souassi et al., 2023). For example, Siegel et al. (2023), a non-ML SWT paper, tested for ungaged regions and unseen data but did not perform a temporal test. To our knowledge, Philippus et al. (2024), appears to be the only published SWT-ML study that applied TUURTs with some success. We further encourage modelers to shift towards more generalizable models, which are in theory, more capable of performing well across diverse scenarios and datasets, and stand to become increasingly important with the unpredictability of climate extremes.

Disclaimer from Reviewer: I propose some additional literature (n = 4-5), and I am a coauthor on 1 of them. I do not view including that literature as mandatory, and only proposed additional sources based on their relevance to the content of this manuscript. I selected "No" to anonymity to avoid any appearance of subversive influence.

AUTHOR RESPONSE: Thank you! This would certainly not be possible without the insightful feedback from referees.