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.

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 in November 2024 found no publications specifically using attention-based transformers for SWT, but we are happy to add some text about their potential to section 2.3.3:

> 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 DL tool, attention-based transformers have yet to be used for SWT, but their aforementioned applications in other geological science fields suggest it is only a matter of time before we see their use 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: We appreciate the opportunity to clarify. For section 1.1 (line 35), the 2nd paragraph begins on line 46, with the words “*Aided by the continued…*”. The same occurs after Line 201, where the RF and XGBoost paragraph begins on line 238. The manuscript follows the Copernicus manuscript template (screenshot below) which appears to not provide for paragraph indentation.

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 138),
2.3.1.2 Cluster analysis and variants (line 145),
2.3.1.3 Support vector machine and regression (line 160),
2.3.1.4 Gaussian Process Regression (line 189),
2.3.1.5 Decision trees and Classification and Regression Trees (line 202),
2.3.1.6 Random Forests and XGBoost (line 215)

We also think that we can make section 2.3.1.6 (lines 226-253) more concise now because model inputs are now a separate section (section 2.4.X). Below is our suggested reduction, with the last LASSO paragraph also being moved to model inputs and selection:

~~Feigl et al. (2021) tested the performance of six ML models, including~~ RF and XGBoost have been used to predict SWT for Austrian catchments with minor differences in model performance, ~~for daily SWT prediction in 10 Austrian catchments. Results showed minor difference 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 estimating~~ estimated daily SWT below dams in China, finding~~. They found that~~ day of year, stream flow flux and AT to be ~~was~~ most influential for the prediction of SWT~~, followed by stream flow flux and AT~~ (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 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~~ AT 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 RF performance was comparatively less effective at detecting non-linear relationships, though both models identified AT as most influential ~~than to the deep learning model. They found AT to be 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~~ compared 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 RF ~~410 USGS sites~~ with four years of daily SWT and discharge to examine maximum SWT. ~~They used RF to estimate~~found that ~~max SWT and thermal sensitivity (Wade et al., 2023), finding~~ Study findings identified~~that~~ AT ~~was the~~as 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 are open to making changes to improve the manuscript for readability. Referee #1 made a similar comment about this section, and we now propose providing 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 ANN 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 add the following to point the reader to the appendix:

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

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 to our manuscript and included their RMSE values in our review. First, we added text on predictive uncertainty in the new 'Discussion' subsection, titled 'Future Directions of SWT Modeling' (this section also addresses ref #1, comment #3):

> The utility of ML in hydrologic modeling has come a long way, with interest seemingly growing exponentially (Nearing et al., 2021). With the novelty of ML, it is easy to get lost in the value of how well a model performs and ignore the science, but with several decades of ML-experience, we think it necessary to urge the scientific community to purposefully use ML 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 (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 the objectives of the SWT community should be 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 must know 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 more of the decision-making process, it makes sense that ML must be transparent and understandable to more than just computer 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., 2015), PyTorch (Abadi et al., 2015), and Google Colab (Bison, 2019). With how fast 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 could also help make ML challenges more transparent and therefore accessible to the hydrologic community-at-large.

We have added a few lines to section 2.5 Decision Support with the provided citations:

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 clearly distinguish between lower-case r, r-squared r^2, and upper-case R^2 :

The square of r is denoted as 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 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 0 to 1, with $R^2 = 1$ indicating that the model can explain all the variance, and vice versa. We note here 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 prevent 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:

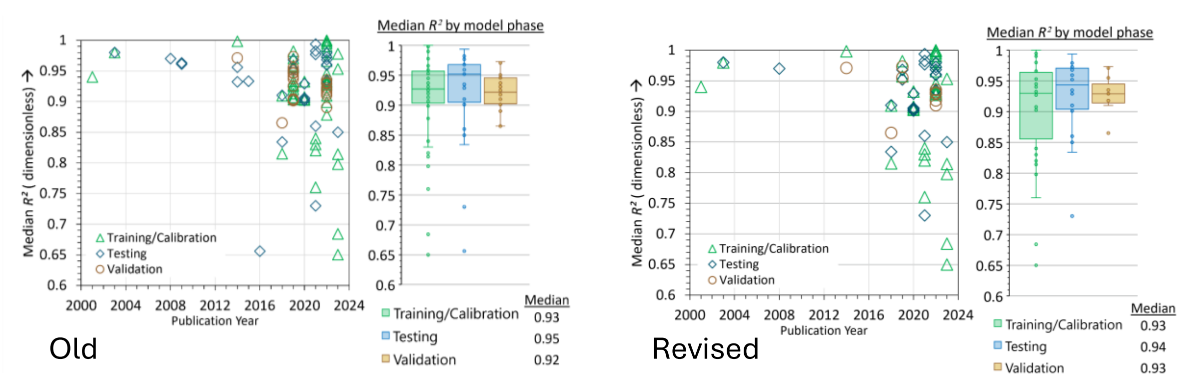
1st paragraph, added after 1st sentence:

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.

3rd paragraph, added last sentence:

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 "Old (left)" and "Revised (right)" 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 and are open to discussion. Perhaps instead of saying "higher standard", we can say "additional standards", 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. This review shows that we have been blinded by the excellence of ML performance metrics relative to physics-based and statistically-based models, and we need to be aware of this short sight 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 a proposed new 'Discussion' subsection, titled 'Future Directions of SWT Modeling', in the first sentence (please see our response to ref #1, comment #4 for the full text):

> "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:

> "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".

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.

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".

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 that mistake. Subsection 3.3 should have said "Model Performance Metrics: Dimensionless" because the subsection summarizes NSE, KGE, etc. We have updated the subsection header accordingly.

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".

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 comment #1B, which we also include below:

Section 2.4.X Overfitting and Underfitting:

When a model is too complex, i.e., has too many features or too many 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 overfitting 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 those where: 1) model development is at its preliminary stages, where 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 certain 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 terrible 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 machine learning models to effectively learn the training data, underfitting is less of an issue in ML whereas overfitting can be widespread. In the following diagram, we present an example workflow to transition away from overfitting and towards generalizability. We further encourage modelers to actively transition towards making more generalizable models, which are in theory, more capable of performing well across diverse scenarios and datasets, which will become increasingly important with the persistence of climate extremes.

Response to ref #1, comment #1B: We have added a few sentences (blue is new) to the Discussion subsection titled "ML as Knowledge Discovery" where we urge for TUURTs (Temporal, Unseen, Ungaged Region Tests)':

Our review finds that ML studies examining SWT have been conducted from a computational perspective, one with a focus on comparing techniques and performance metrics as opposed to explaining the nature of SWT dynamics or influencing processes. While it is understandable that not every ML-SWT paper aims to explain physical processes, we think the SWT community should come together and agree on a baseline of tests that all ML-SWT models should undergo for model robustness and transferability. Along these lines, we urge consideration of TUURTs (temporal, unseen, ungaged region tests) for future ML-SWT models as a helpful step towards not only better modeling practices but also increased model transparency and robustness. For this, we clarify that testing for "unseen" cases means testing only within the developmental dataset, whereas testing for "ungaged" cases means testing for new sites that have not been previously seen by the model at all. Recent ML-SWT studies have only applied one or two of the tests, but not all three (Topp et al., 2023; Hani et al., 2023, Souassi et al., 2023). Siegel et al. (2023), a non-ML SWT paper, tested for ungaged and unseen data but did not perform a temporal test. A relatively new study, Philippus et al. (2024), appears to be the only published SWT-ML study that purposefully applied TUURTs with some success.

Disclaimer: 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.