Hydrology and Earth System Sciences,
Manuscript #HESS-2024-256
March 23rd, 2025

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

Dear Dr. Christa Kelleher,

We thank you for your time and patience in handling the review of our manuscript. We also appreciate the referee's feedback and provide our text regarding the "prediction in ungaged basins" below. For revisions, new/edited text is in BLUE, removed text is ~~crossed out,~~ and original text is left in black. The statement "*revised lines XXX-XXX*" indicates the in-line placement of the described changes in "**2-Track-Changes-Manuscript-HESS-2024-256_v3**". The document "**3-Clean-Manuscript-HESS-2024-256_v3**" is the "final" version.

Referee Comments from Report #1

I would like to thank the authors for addressing the reviewers' comments. I believe they have significantly improved the manuscript. I would like to bring to the authors' attention a minor issue: the topic of 'prediction in ungaged basins' has been explored in SWT modeling for at least a decade. I encourage the authors to review the following two papers, as well as the references cited within them, which may help provide additional context and background: https://doi.org/10.5194/hess-19-3727-2015 and https://doi.org/10.1002/hyp.14400

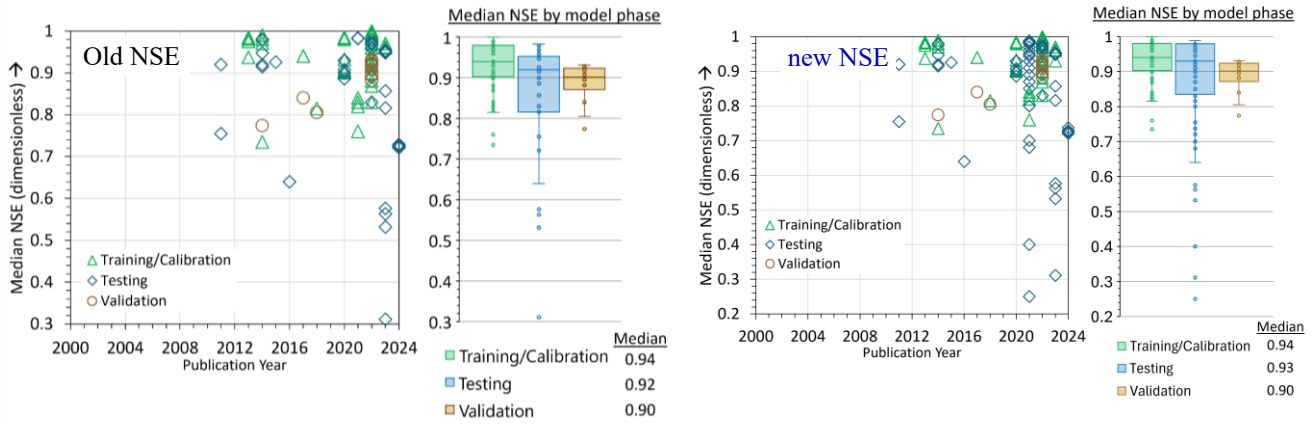
AUTHOR RESPONSE: We agree that the topic of 'prediction in ungaged basins' has been previously explored and we appreciate the opportunity to clarify. The added text is below (*revised lines 894-914*) and includes the references suggested by the reviewer (in **bold**):

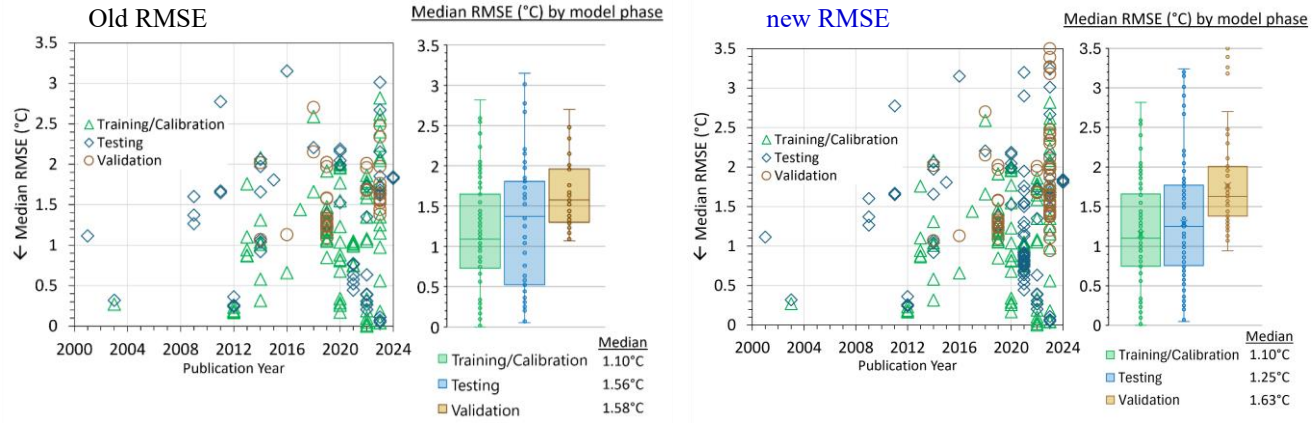
> The challenge of prediction in ungaged basins in SWT modeling has been explored for at least a decade by process-based (Dugdale et al., 2017) and statistically based (**Gallice et al., 2015**, Isaak et al., 2017; Wanders et al., 2019; Siegel et al., 2023) models. Unfortunately, process-based models continue to be limited by data requirements and memory or processing/programming impediments (Dugdale et al., 2017; Ouellet et al., 2020), while statistically based models struggle to account for changing physical conditions (Benyahya et al., 2007; Arismendi et al., 2014; Lee et al., 2020). Physics-derived statistically based models have been applied in ungaged regions (Gallice et al., 2015) but models tend to be region-specific and not generalizable. We posit that a future direction of ML models is to expand on their ability to learn, identify and mimic the complexity needed to improve SWT predictions for ungaged basins. To date, researchers have used ML to model SWT for partially ungaged (i.e., discharge used as input) regions across the CONUS (**Rahmani et al., 2020**, 2021), though limitations persist in ~~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~~ complex and critical regions in the ~~south~~west (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). Given community-wide modeling interest expanding from SWT prediction to forecasting (Zhu and Piotrowski, 2020; Jiang et al., 2022; Zwart, Diaz, et al., 2023), ML-use could prove essential in capturing unknown, complex SWT patterns in space and time (Philippus, Corona, et al., 2024) and with shifting baselines. ~~We have also learned that~~ With regards to ML models such as LSTMs predicting extremes, a limitation that must be addressed ~~with ML models such as LSTMs,~~ is that they generally only make predictions within the bounds of their training data (Kratzert et al., 2019) though researchers are looking to improve on this by using ML-hybridizations (Rozos et al., 2023). ~~, which is a limitation for predicting extremes. Thus, we strongly urge~~ Overall, there is promising work in the community towards creating ML models for SWT that generalize better and/or are more robust ~~towards~~ for predictions of extremes.

Additionally, we describe the Rahmani et al. (2021) study (*revised lines 539-542)*:

> A follow-up study by **Rahmani et al. (2021)** used six years of SWT data and relevant meteorological parameters for 455 sites across the CONUS (minus California and Florida) to test LSTM models for data-scarce, dammed, and semi-ungaged basins (discharge used as input). The follow-up study showed improved performance, but the LSTM models remained limited in capturing the influence of latent contributions such as base-flow and subsurface storage.

We updated our calculations of performance metrics (screenshots below) to include the suggested Rahmani et al. (2021), for **NSE** (top, fig.3) and **RMSE** (bottom, fig.4). We note no significant changes (*revised lines 730-735*):





We updated Table *1 (revised line 755)* with added Rahmani et al. (2021) publication, screenshot below:

Table 1. Average, median, maximum, and minimum RMSE (°C) for studies grouped by local/watershed and regional/CONUS spatial scales.

	Local/Watershed (< 100 km² area)	Regional/CONUS (> 100 km² area)
Number of data points	900	13692~~2~~
Average	1.52~~4~~	1.55
Median	1.38	1.42~~3~~
Maximum	5.170	4.387
Minimum	0.038	0.0002

We updated suggested ratings on Table 2 *(revised line 770)* with Rahmani et al. (2021), screenshot below:

Table 2. Suggested ratings for performance metrics (median) using metrics published by ML studies examining SWT.

	R^2			NSE		
Rating	Training	Testing	Validation	Training	Testing	Validation
Very Good (>)	0.99	0.99	0.96	0.99	0.~~97~~98	0.93
Good (range)	0.89 - 0.99	0.92 - 0.99	0.94 - 0.96	0.92 - 0.99	0.~~88~~ 84 - 0.~~97~~98	0.88 - 0.934
Satisfactory (range)	0.79 - 0.92	0.86 - 0.92	0.91 - 0.94	0.~~84~~ 85 - 0.92	0.70 - 0.8~~4~~8	0.83 - 0.88
Unsatisfactory (<)	0.79	0.86	0.91	0.~~84~~85	0.~~70~~61	0.83
	RMSE (°C)			**MAE (°C)**		
Rating	Training	Testing	Validation	Training	Testing	Validation
Very Good (>)	0.25	0.2~~6~~2	1.15	0.33	0.42	0.86
Good (range)	1.34 - 0.25	1.5~~1~~46 - 0.262	1.80 - 1.15	1.18 - 0.33	1.12 - 0.42	1.32 - 0.86
Satisfactory (range)	2.43 - 1.34	2.770 - 1.5~~1~~46	2.45 - 1.80	1.70 - 1.01	1.97 - 1.19	1.79 - 1.32
Unsatisfactory (<)	2.43	2.770	2.45	1.70	1.97	1.79

***AUTHOR EDITS, NOTE FOR EDITOR:**

Upon proof-reading, we realized that the in-text description for fig. 1 was towards the end of the manuscript (original line ~837) and far from the location of fig. 1 (original line ~405). To aid the reader, we moved the text for fig. 1 (original lines 837-842) closer to where the figure is mentioned (*revised lines 393-404*):

> In the five-step outline (Fig. 1), we suggest the need for "Temporal, Unseen, Ungaged Region Tests" (TUURTs) in SWT ML modeling. The idea behind TUURTs has been applied for decades in SWT process-based (Dugdale et al., 2017) and statistically based models (Benyahya et al., 2007; **Gallice et al., 2015**) to improve SWT ~~which is a call for temporal and spatially focused testing that can be used to strengthen~~ model robustness. In 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. Some statistically based models, such as DynWat (Wanders et al., 2019) and the Pacific Northwest (PNW) SWT model (Siegel et al., 2023) have tested for ungaged regions and unseen data. In the last few years, ML-SWT studies have begun applying TUURTs (**Rahmani et al.,** 2020, **2021**, 2023; Topp et al., 2023; Hani et al., 2023, Souassi et al., 2023; Philippus et al., 2024) but more ML-SWT studies need to apply these tests to improve user confidence in extrapolation capability. ~~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 researchers 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.

For transparency, we would also like to point out minor revisions for sentence clarity or to update text due to the addition of Rahmani et al. (2021):

Revised lines 18-20: edited sentence structure and updated publications from 56 to 57.

> Our review found that in the recent five years (2020–2024), ~~a similar number (28) of~~ more studies ~~publications~~ using ML for SWT were published, ~~as were published in the~~ than had been in the previous 20 years, (2000–2019), totaling 57.

Revised line 139: "...software may be publicly available but could take years to publish updates"

Revised lines 152-153: removed filler words and added "in various fields of hydrology" to clarify where ML models have the potential for growth.

> "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 in various fields of hydrology,"

Revised line 383: "...at preliminary stages, ~~the interest is in~~ such as a "proof of life" concept,..."

Revised line 385: "~~like~~ similar to the training dataset".

Revised lines 411-412: "Since then, ~~studies have used varying~~ various input variables have been tested..."

Revised line 413: "For example, studies have used..."

Revised line 416: "~~Traditionally used~~ Other model inputs..."

Revised line 425: "...satellite product inputs such as estimates of sky cover..."

Revised lines 535-536: "study ~~could~~ did not explicitly state what physical laws ~~(if any)~~ were followed..."

Added text in parentheses: *line 610 "*(Pearson's *r*, R^2)", *line 611* "(NSE, KGE)", and *line 613* "(RMSE, MAE)"

*Revised lines 705-706:"*RMSE (~~44~~ 45 citations), NSE (~~24~~ 25), ~~and~~ MAE..."

Revised line 723: "...with a median NSE of 0.93 across ~~24~~ 25 studies (fig.3)"

*Revised lines 742-743: "*The median RMSE ~~values~~ was 1.~~40~~35 °C across ~~44~~ 45 studies (fig. 4)".

*Revised line 766: "*for the local (~ 1.~~51~~52 °C)..."

Revised line 841: "(see section 2.4.1, fig.1)."

Revised line 891: plural - "studies should consider ~~is~~ are:"

Once again, we appreciate the opportunity to proof-read and revise the manuscript and think the manuscript is better as a result. Should you have any questions, please email me at claudia.corona@mines.edu.

Thank you kindly for your time and consideration. Much appreciated!

Sincerely,

Claudia R. Corona
Postdoctoral Fellow, CO Mines

Terri S. Hogue
Dean, Earth and Society Programs, CO Mines