Review of "Which strategy to improve the performances of an LSTM-based model for extreme stream temperature values?"

General comments

This paper compares the performance of LSTM models to predict daily stream water temperature over the whole year, but notably also during the days with the highest 10% of observed water temperatures. Different sets of models are tested based on: (i) local vs regional models, (ii) different sets of input variables, and (iii) different loss functions. Data from several stations in the Garonne River in France are used. The main finding is that regional multi-station training including static attributes improves performance, whereas customized loss functions do not improve performance.

Overall, the manuscript is well-written, and the figures are clear. The manuscript could be interesting to the readership of HESS. While the knowledge that regional multi-station training including static attributes improves LSTM performance is not very novel, its comparison against the change in performance for different loss functions is valuable. Nevertheless, some important considerations are still needed. Please find below some specific comments and suggestions.

Specific comments

1. The computation of catchment average potential evapotranspiration (PE) according to Oudin et al. (2005) seems unnecessary. Catchment average Ta and information on day of the year could be used instead of PE, unless PE is strictly necessary for obtaining Qsim.

2. It could be useful to have an additional table showing the values from the different input variables for all stations, even if it is in the appendix.

3. It could be useful to include the long-term mean and standard deviation of daily water temperature from each station as a static variable, but I understand this might not be feasible if it means that all models need to be re-trained.

4. L211: It would be useful to do a more detail assessment for choosing the hyperparameters of the LSTM models, considering the findings of Feigl et al. (2021). Doing hyperparameter optimization as in Kraft et al. (2025) would be a good option.

   Kraft, B., Schirmer, M., Aeberhard, W. H., Zappa, M., Seneviratne, S. I., and Gudmundsson, L.: CH-RUN: a deep-learning-based spatially contiguous runoff reconstruction for Switzerland, Hydrol. Earth Syst. Sci., 29, 1061–1082, https://doi.org/10.5194/hess-29-1061-2025, 2025.

5. An important point when training deep learning models is their inherent randomness. It would be useful to assess for each model setup the variability in the performance when retraining the model with different random seeds. In this way, the differences in performance from the different strategies tested in the paper can be put into context with the uncertainty in performance from varying random seeds.

6. L240–244: I would suggest constructing the validation set using 15% of the observations at all stations. Using more data from the non-test stations could bias the models to better fit to these stations instead of to the test stations. Please also clarify if this 15% corresponds to at least one continuous year of observations.

7. Suggestion to move section 3.2.3 to 3.2.1 to be more consistent with the order proposed in the last paragraph of section 1 and in section 4.

8. Important: I don't see the need to have the denominator in Eq. 3, and it seems to be counterproductive. When having high Tw and u, the denominator increases faster the numerator, thus reducing the loss for higher values of Tw (see Table below with example data). If this is the case, then the loss function does not serve its intended purpose to give higher weights to errors when Tw values are high. I think only the numerator of Eq. 3 should be used as loss function.

u =	3				
lambda =	2				
T_w_obs	T_w_sim	T_w_obs_bar	Eq. 3 numerator	Eq. 3 denominator	Loss
20	19.5	13	342371.2656	33674809	0.010166985
15	14.5	13	106520.6406	1387684	0.076761453
10	9.5	13	20341.89063	1432809	0.01419721

9. L282: Explain what u does in Eq. 3, i.e. having higher powers on higher Tw values would lead to larger errors, thus emphasizing the weight on high Tw, if I understood it correctly.

10. L287–289: This is important. It would be useful to add another sentence or example to clarify that having higher powers on higher Tw values would lead to larger errors, thus emphasizing the weight on high Tw.

11. L324: Report the number of cases out of the 21 for which the performance improved. This is more informative than saying it is not statistically significant.

Minor comments and technical corrections

1. The study from Padrón et al. (2025) could be useful for section 1.2 and the second paragraph of section 5.

   Padrón, R. S., Zappa, M., Bernhard, L., and Bogner, K.: Extended-range forecasting of stream water temperature with deep-learning models, Hydrol. Earth Syst. Sci., 29, 1685–1702, https://doi.org/10.5194/hess-29-1685-2025, 2025.

2. Table 1: Please clarify what are the min, median and max values reported. Are these average values across all 21 stations? Otherwise, it is not consistent with the value of 21C reported in L381.

3. L238–239: Mention here that the 15% of the available records span at least one full year.

4. Figs. 3 and 4: Clarify if the bottom row corresponds to the best loss function averaged over all sets of input variables. If this is not the case, then why is the MAE of the "Reference" loss function (1.29) lower than that of the "Best" loss function (1.48) for the model with only Ta as input in Fig. 4.

5. L381: Should "which" be replaced by "with".

6. Fig. 5: suggestion to reduce the size of the black dots to improve visualization.

7. Fig. 5 caption: "(a and c)" should be exchanged with "(b and d)" and vice versa.