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

by

Mohamed Saadi, Louis Guichard, Gabrielle Cognot, Laurent Labbouz, Hélène Roux

Submitted to *Hydrology and Earth System Sciences*

Manuscript ID: egusphere-2025-3393

Answer to comments of the Anonymous Referees, round 2

23 April 2026

1 Summary of main changes

We first would like to thank the two Anonymous Referees for their time and their constructive feedback and comments, which helped us clarify our methodological choices and improve the discussion of the limitations of our methodological setup. In response to the comments of both Referees, we made the following main changes:

- We clarified the role of catchment-scale potential evapotranspiration as a proxy for catchment-scale air temperature (see our answer to Comment 1 of Referee #1).
- We justified the choice of (1) MAE as an evaluation metric and (2) the 10% range of the test period to evaluate model performances over extreme stream temperature values as a response to Comments 1 and 2 of Referee #2. Note that we limited the impact of the choice of MAE on the interpretation of our results by including an additional test evaluation using RMSE instead of MAE in Appendix C. As for the choice of the top 10% range, this (seemingly arbitrary) choice is limited by the size of the data available for test.
- We improved the description of our choice of the lookback value by explicitly mentioning that we optimized it on the validation set for each model configuration.

In the following, we provide a detailed response to Referee #1's comments in Section 2, and a detailed response to Referee #2's comments in Section 3.

2 Detailed response to Anonymous Referee #1

General comment: *"I have read the revised manuscript and the authors' response to my first review. Overall, the response is constructive and the revisions address the main issues raised previously. The manuscript is in good shape and I recommend minor revisions."*

Authors' response: We thank the Referee for their constructive and positive feedback!

Comment 1: *"Catchment-scale air temperature vs PE. Please avoid implying that PE is fully equivalent to catchment-scale air temperature. Temper the wording (e.g., describe PE as a proxy) and, if possible, add a brief supporting diagnostic (e.g., Ta–PE correlation) or a short explanatory sentence."*

Authors' response: We did not intend to imply that potential evapotranspiration (PE) is *fully* equivalent to catchment-scale air temperature, because (as pointed out by the Referee) PE is only a proxy for catchment-scale air temperature. The only place in the manuscript where this could be misinterpreted is in Lines 262-266 of the revised manuscript, which read:

"Finally, comparing $T_{amn} + T_{amx} + P + PE$ and $T_{amn} + T_{amx} + Q_{sim}$ will show whether the LSTM models are able of maintaining similar (or obtaining better) performances by exploiting the catchment-scale forcing (P and PE, with PE almost linearly dependent on catchment-average air temperature according to Oudin et al., 2005) instead of the more relevant station-scale streamflow (Q_{sim})."

Here, the linear dependence between PE and catchment-scale air temperature is simply given by the temperature-based formula for PE of Oudin et al. (2005). As asked by the Referee, we computed the correlations between daily catchment-scale air temperature and daily potential evapotranspiration for our set of 37 stations between the years 1988 and 2022. We found that these correlations vary between 0.92 and 0.93, highlighting high strong linear correlations between the two variables. Following the Referee's comment, and to avoid confusion, we modified the relevant part of the manuscript as follows:

"Finally, comparing $T_{amn} + T_{amx} + P + PE$ and $T_{amn} + T_{amx} + Q_{sim}$ will show whether the LSTM models are able of maintaining similar (or obtaining better) performances by exploiting the catchment-scale forcing (P and PE, with PE being a good proxy for catchment-average air temperature according to the temperature-based formula in Oudin et al., 2005) instead of the more relevant station-scale streamflow (Q_{sim})."

where we underlined that PE is a good proxy for catchment-scale air temperature, as suggested by the Referee.

3 Detailed response to Anonymous Referee #2

General comment: *"This article demonstrates a data-driven approach to simulate extreme stream flow temperatures. The authors conducted an extensive ablation of input features and loss terms in training LSTMs for extreme streamflow temperature. They develop one catchment LSTM across multiple stations in the Garonne catchment and compare this with local LSTMs, trained on stations separately. In their ablation the authors focus especially on 1) using multi-scale data that is catchment scale forcing and auxiliary information, and 2) modifications to the loss term that aim at penalizing extreme values.*

While the approach to optimising simulation of extreme temperature through targeted loss functions would be a valuable experiment to show and learn, one major thing I am sceptical about is the evaluation and choices regarding their ablation. In order to make this a more robust part of their study, I would kindly ask the authors to add the full picture to their methods description on the evaluation and consider the comments below to revise or discuss their comparative evaluation.

Because statistical testing was used for quantifying significant performance differences, the choice on the scoring function gets blurred in the evaluation framework (section 3.3). My reading is that all models were evaluated on the test set with the MAE computed on the upper 10% - which is a limitation, and this choice needs to be much more visible in the methods. E.g., its not given that this is a robust choice for example for figure 5 where MAEs are computed over the whole test period."

Authors' response: We thank the Referee for their constructive feedback! As a response, we clarified our methodological choices regarding the MAE as an evaluation metric (see our response to Comment 1) and the choice of the top 10% range (see our response to Comment 2). We agree that these choices can have an impact on the interpretation of our results, and we acknowledged these limitations in the Discussion section (see our response to Comment 2).

Comment 1: *"MAE and MSE as loss functions optimize on different properties of the predictive distribution: The MAE optimizes toward median and the MSE toward the mean, and which one is appropriate is, in principle, non-arbitrary (e.g. 10.5194/gmd-15-5481-2022). What for this work matters is that mean and median are only equivalent if the residual distributions are symmetric - which I assume is not the case for streamflow temperature, hence the focus on capturing extreme values. Using only the MAE as an evaluation score on models that are trained with MSE is not a fair comparison if the target distributions are asymmetric. This may particulary be relevant for evaluation on whole test periods, i.e. for figure 5."*

Authors' response: We completely agree with the Referee's comment, and we thank them for suggesting the nice paper by Hodson (2022). We would like to emphasize two points as an answer to this comment.

First, we chose MAE as an evaluation metric simply because it is easy to interpret. Since this metric will put more emphasis on model errors around the median/most frequently observed values of stream temperature (as underlined by the Referee), and in order to evaluate model errors on extreme values, we recomputed this metric over the top 10% of the observations during the test period. We believe that this choice allows for an easy interpretation of model performances for both the overall range of stream temperature values *and* the extreme, top 10% range. The Referee cites Figure 5 in their comment to argue that MAE may only be relevant for a whole-period test evaluation, but if we look at Figure 5 of the manuscript, we can actually see that MAE computed on the top 10% range (i.e., MAE_{10}) emphasizes model errors on extreme, high temperature values. For example, Figure 5a shows the performances on the whole test period for models trained without static attributes. The MAE scores over the whole period are 0.74°C for the model with $T_{amn} + T_{amx} + P + PE$ as input set and 0.86°C for the model with $T_{amn} + T_{amx}$ as input set, and the simulations satisfactorily reproduce the overall dynamics of stream temperature over the whole period. But if we look at Figure 5c, where we zoom in on the stream temperature values higher than 15°C, we can see that these same models perform worse, and this is well reflected by significant increases in MAE from 0.74°C/0.86°C over the whole period to 1.34°C/1.18°C over the top 10% range. This highlights that although being suited for reflecting model performances on median values, the computation of MAE on the top 10% range reflects – to a satisfactory extent – the model performances over the extreme range of stream temperature values.

Second, as a response to the first round of review, we also included an evaluation using the root-mean-squared errors (RMSE) to control whether the choice of MAE as an evaluation metric favours the MAE-based loss functions. We found that this was not the case, at least not systematically. This is shown in Figure C1 of Appendix C of the revised manuscript (Figure R1 of this response), where MAE-based loss functions (with $\mu = 1$) also score systematically among the best in terms of RMSE (despite the fact that MSE without standardization is the best loss function for regional models according to RMSE, MAE with/without standardization rank second).

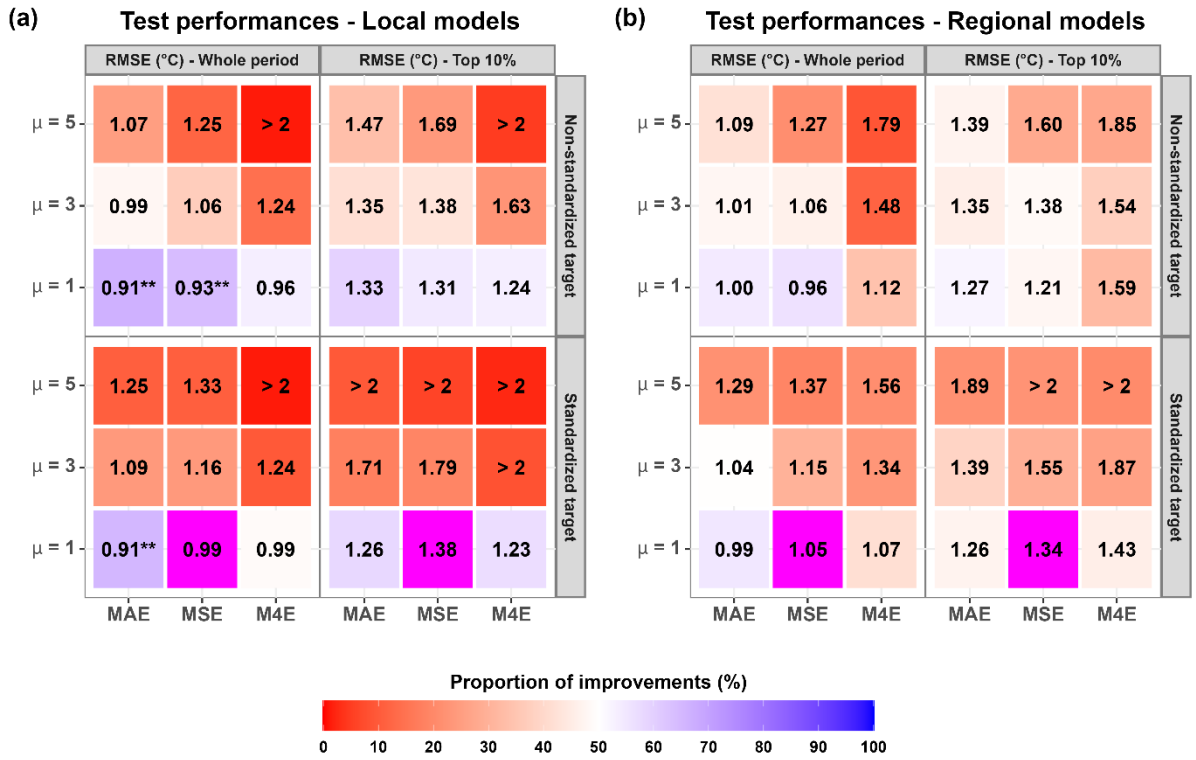

Figure R1: Median test performances (RMSE, in °C) of the local models (a) and the regional models (b) according to the loss function used for training. This figure is similar to Fig. 2 but with RMSE as a criterion to evaluate test performances. Colours indicate the proportion of cases for which the use of the loss function yielded better results than the reference loss function (MSE with standardization, shown in magenta). Asterisks indicate that the custom loss function is significantly better than the reference loss function according to the binomial test: * for a significance threshold of 5%, ** for a threshold of 1%, and * for a threshold of 0.1%.**

To summarize, (1) we used MAE because it's intuitive and easier to interpret, and (2) we also used RMSE to control whether our choice of MAE biases the evaluation in favor of MAE-based loss functions. To better justify our choices, we modified the first paragraph of Section 3.3 of the manuscript (***"Evaluation framework"***) as follows:

"To quantify model performances, we computed the mean absolute error MAE (in °C) between the predictions of each model and the observed values during the test period. We chose MAE as an evaluation metric simply because it allows for easier interpretation of model performances. We computed the MAE score first on the whole test period, then on a restricted period during which observed T_w exceeded the 90th percentile, which corresponds to the 10% highest T_w values observed during the test period. Targeting higher ranges of extreme T_w values (for example, top 5% or 1%) is limited by the relatively short test samples available at our 21 test stations, which have sizes that vary from 390 to 1074 days. To control whether choosing MAE as an evaluation metric favours MAE-based loss functions, we also evaluated the model performances using RMSE as an evaluation metric instead of MAE (see Appendix C for the corresponding results)."

In Appendix C, we show the results with RMSE as an evaluation metric instead of MAE. We modified the first paragraph of Appendix C that summarizes the results as follows:

"To control whether the choice of MAE as an evaluation metric biases the results in favour of MAE-based loss functions, and in order to compare our results with previous studies (e.g., Rahmani et al., 2021b), we also evaluated the test performances using RMSE as a criterion. Figure C1 shows the evolution of the test performances with respect to the choice of the loss function, and comparing it with Fig. 2 suggests that

using MAE as a criterion for test performances does not bias our conclusions in favour of MAE-based loss functions (see the results with local models, Fig. C1a). Figures C2 and C3 show the performances of the locally trained (Fig. C2) and the regionally trained (Fig. C3) LSTM models, which are replicates of Figs. 3 and 4 but with RMSE instead of MAE as an evaluation metric. These figures confirm the importance of regional training with catchment attributes in improving the LSTM performances especially for the range of extreme (top 10%) T_w values."

Comment 2: *"An evaluation with MAE for the median may be inconsistent with the training objective – hence the evaluation only on the 10% highest values, i.e. an evaluation of the median of that specific subset. But training with high values of mu or lambda may force the model to optimize on higher quantiles than 10% percent. This may be a possible explanation for what the authors state in l. 306 (section 4.1)."*

Authors' response: We thank the Referee for this great observation. This might be an explanation why increasing μ or λ was not rewarded enough with our evaluation metrics. In order to do this and fully respond to the Referee's comment, we should shrink the range on which we compute the MAE score to target higher and higher quantiles. Unfortunately, we are limited by the number of datapoints in our test sets: they vary from 390 (slightly above 1 year) values up to a maximum of 1074 values (about 3 years' worth of observations), meaning that we computed our MAE_{10} over a number of values that vary from 39 to 107 datapoints. If we aim at higher quantiles, say 5% and 1%, corresponding MAE_5 and MAE_1 would be computed on only 20-54 datapoints and 4-11 datapoints, respectively. This is too low to provide a robust evaluation of model performances over the extreme stream temperature ranges.

To sum up, the choice of the 10% range is a compromise between (1) targeting a range of extreme stream temperature values (thus shrinking the pool of datapoints used to compute MAE), and (2) having a minimum number of datapoints that is sufficient to provide robust evaluation of model performances over the extreme range of stream temperature values (thus choosing a threshold of frequency of non-exceedance that is not too high). We justified this somewhat arbitrary choice in Section 3.3 of the revised manuscript as follows:

"We computed the MAE score first on the whole test period, then on a restricted period during which observed T_w exceeded the 90th percentile, which corresponds to the 10% highest T_w values observed during the test period. Targeting higher ranges of extreme T_w values (for example, top 5% or 1%) is limited by the relatively short test samples available at our 21 test stations, which have sizes that vary from 390 to 1074 days."

In the Discussion section, we mentioned the Referee's observation as a possible, alternative explanation of unsatisfactory results over the top 10% range with higher values of μ and λ. Accordingly, we added the following lines to the revised manuscript:

"We tested several loss functions to see whether increasing the weight of high T_w values could result in better performances over the top 10% range, following previous works in process-based environmental modelling (see e. g. Jadon et al., 2024; Thirel et al., 2024). Our results indicate that this is actually detrimental not only to the reproduction of extreme values, but also to the reproduction of the overall thermal response. This is perhaps due to the fact that some of our tested loss functions put too much emphasis on errors over large T_w values, thus limiting the information content that the LSTM models were able to extract from the whole range of observations. This suggests that in order to satisfactorily perform during extreme thermal conditions, LSTM-based models should first learn the overall thermal behaviour observed under "usual" conditions. An alternative explanation may be that higher values of μ and λ drive model optimization towards more extreme values than the top 10% range, but a robust evaluation of this argument requires longer test periods in order to quantify model

performances over a sufficient number of values exceeding higher frequencies of non-exceedance."

Comment 3: *"It should be made more obvious if the same lookbacks were chosen for all the models, or if the comparative evaluation that we see in Figures 2 and 3 allows varying lookbacks, as they were seemingly part of the ablation too (explained in section 3.3)."*

Authors' response: We already specified in Section 3.3 that the lookback value was chosen as the one that minimized the MSE over the validation set. In the revised manuscript version, we improved the clarity of our evaluation framework by explicitly mentioning that the lookback value is optimized on the validation set, and that the models that were kept for test do not necessarily share the same lookback value:

"We repeated each training configuration (local vs. regional training, input variables, loss functions) three times corresponding to each of the three pre-chosen lookback values (30, 90, and 365), meaning that we trained a total of 3 × 4 × 18 × 21 = 4536 local models and 3 × 7 × 18 = 378 regional models. For each configuration, we kept only the lookback value that provided the lowest MSE on the validation set (in other words, we optimized the lookback value on the validation set). Thus, we evaluated a total of 4 × 18 = 72 local models and 7 × 18 = 126 regional models on each of the 21 test stations, which do not necessarily share the same lookback value."

In addition, we added a clarification to the captions of Figures 2, 3, and 4 to mention that the models kept for test do not share the same lookback. For example, the caption of Figure 2 now reads:

"Figure 2: Median test performances (MAE, in °C) of the local models (a) and the regional models (b) according to the loss function used for training. These median values were computed from a set of 21 × 4 = 84 points for the local models and 21 × 7 = 147 points for the regional models. Note that the models we kept for test do not necessarily share the same lookback. Colours indicate the proportion of cases for which the use of the loss function yielded better results than the reference loss function (MSE with standardization, shown in magenta). Asterisks indicate that the custom loss function is significantly better than the reference loss function according to the binomial test: * for a significance threshold of 5%, ** for a threshold of 1%, and *** for a threshold of 0.1%."

Comment 4: *"Fig. 3: So the middle row left and right columns use the same loss function, while the bottom row different ones? Clarify."*

Authors' response: Yes, exactly! The middle row of Figure 3 (and also Figure 4) shows the results for the same loss function, i.e., MSE computed on standardized target. The panels of the bottom row show results for a priori different loss functions depending on the range. This is clarified in the caption of Figure 3, which now reads:

"Figure 3: Distributions of the test performances (MAE, in °C) of the local models over the whole test period (left column) and the test subperiod corresponding to the highest 10% observed values (right column). The top row shows the distributions over all the loss functions (18 × 21 = 378 points per distribution). The middle row shows the performances for the reference loss function, MSE with standardization, for both the whole period and the top 10% (21 points per distribution). The bottom row shows the performances for the best loss function over all sets of input variables (MAE without standardization for the whole period and MAE with standardization for the top 10%, 21 points per distribution). Numerical values under the boxes represent the median value for each distribution. Letters under the numerical values rank the distributions, and are defined such as distributions that share at least one letter are not significantly different

according to the binomial test. Note that the models we kept for test do not necessarily share the same lookback."

Similarly, we also updated the caption of Figure 4 as follows:

"Figure 4: Distributions of the test performances (MAE, in °C) of the regionally trained models over the whole test period (left column) and the test subperiod corresponding to the highest 10% observed values (right column). Top row shows the distributions over all the loss functions (18 × 21 = 378 points per distribution). The middle row shows the performances for the reference loss function, MSE with standardization, for both the whole period and the top 10% range (21 points per distribution). The bottom row shows the performances with the best loss function over all sets of input variables (MAE with standardization for the whole period and MSE without standardization for the top 10%, 21 points per distribution). Numerical values under the boxes represent the median value for each distribution. Letters under the numerical values rank the distributions, and are defined such as distributions that share at least one letter are not significantly different according to the binomial test. Note that the models we kept for test do not necessarily share the same lookback."

Comment 5: *"l. 190: This builds quite a tension arc towards section 3.3., but in this section you don't state a best lookback length either. Could you be more clear about your choice, and at best already state here which lookback length you chose?"*

Authors' response: We agree with the Referee's comment. The problem is that the section in question, Section 3.1, is focused on providing an overview of our choices for the LSTM model structure. A detailed explanation of the choice of the lookback value requires presenting the choices of the splitting strategy (training, validation, test), which are part of Section 3.2.2. In order to satisfy the Referee's request while keeping Section 3.1 as concise as possible, we modified the lines in question by providing more details on the general strategy we adopted to choose the best lookback value:

"The LSTM performances were sensitive to the lookback, which was expected due to the strong differences in T_w dynamics across our catchment set. For this reason, we chose to test three lookbacks (30, 90, 365) to accommodate for T_w dynamics at the monthly, seasonal, and annual scales. Our strategy to choose the best lookback value among the three tested values is based on splitting our data sample into three subsets for training, validation, and test of the LSTM models (see Sect. 3.2.2 for more details), and then keeping only the lookback value that minimized model errors on the validation set. More details are provided in Sect. 3.3."

Comment 6: *"Section 3.3.: Here the choice of the lookback is stated as part of an original ablation, so do lookback lengths vary across the results we see? (see also c) in major comment above)"*

Authors' response: Yes, lookback lengths vary across the results in Figures 2, 3, and 4! We improved the description of our methodology by adding clarifications to Sections 3.1 and 3.3 and to the captions of Figures 2 to 4. Please see our detailed response to Comment 3.

Comment 7: *"Section 3.2.1: nice description of the loss term."*

Authors' response: We thank the Referee for their nice feedback.

Comment 8: *"Tab. 2: Listing input variables and models of those variables feels a bit redundant. I'd say a table of just input variable combinations is enough. Regarding CatAttrs –*

are reach attributes are explicitly named in Tab.1 or do they fully intersect with catchment attributes?"

Authors' response: We agree with the Referee that the listing in Table 1 is redundant, but we decided to keep it as is because we think it helps explain the meaning behind the acronyms in column 2 (such as "T_{amn} + T_{amx} + Q_{sim}"), and this is necessary since these acronyms are later used in Figures 3 to 5, and Figures C2-C3 of Appendix C. To better clarify this, we changed the title of the second column of Table 2 from ***"Input variables"*** to ***"Acronyms for the sets of input variables"*** and the title of the third column from ***"Models"*** to ***"Corresponding models"***.

As for the difference between catchment vs. reach attributes, these are all explained in Table 1 under the category of ***static*** attributes, which include 1 reach attribute (station elevation) and 9 catchment attributes (aridity index, mean catchment elevation, catchment area, percentage of urban areas, forest cover, and agricultural areas, and mean content of sand, silt, and clay). We added this clarification to the title of Table 2 as follows:

"Table 2: Summary of tested local and regional models. "CatAttrs" refers to static, catchment and reach attributes (see Table 1). Reach attributes include only station elevation. Catchment attributes consist of aridity index, mean catchment elevation, catchment area, land cover properties (percentage of urban, forest, and agricultural areas), and soil properties (sand, silt, and clay content)."

Comment 9: *"l. 470: The punctuation in this sentence is weird."*

Authors' response: We agree with the Referee. We modified the sentence in question to enhance its clarity. It now reads:

"Future research includes a ranking of local/reach attributes and regional/catchment attributes by quantifying their importance in controlling T_w dynamics and spatial variability."

4 Cited References

Hodson, T.O., 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci. Model Dev. 15, 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Jadon, A., Patil, A., Jadon, S., 2024. A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting, in: Sharma, N., Goje, A.C., Chakrabarti, A., Bruckstein, A.M. (Eds.), Data Management, Analytics and Innovation. Springer Nature, Singapore, pp. 117–147. https://doi.org/10.1007/978-981-97-3245-6_9

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. J. Hydrol. 303, 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., Shen, C., 2021. Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. Environ. Res. Lett. 16, 024025. https://doi.org/10.1088/1748-9326/abd501

Thirel, G., Santos, L., Delaigue, O., Perrin, C., 2024. On the use of streamflow transformations for hydrological model calibration. Hydrol. Earth Syst. Sci. 28, 4837–4860. https://doi.org/10.5194/hess-28-4837-2024