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 the Editor's and the Anonymous Referees' comments

29 November 2025

1 Summary of main changes

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

- We reduced the length of Section 1.2 (Introduction) and moved the methodological details regarding the reconstruction of streamflow at stream temperature stations to a new Appendix (Appendix A of the revised manuscript; see our response to Comments 1 to 4 of Referee #1). In addition, we re-organized the subsections of Section 3.2 in response to Comment 7 of Referee #2, and updated Figure 5 for clarity;
- We justified the exclusion of predictors that explicitly encode time features (see our response to Comment 6 of Referee #1) and clarified our choice of using catchment-scale potential evapotranspiration instead of catchment-scale air temperature (see our response to Comment 5 of Referee #1 and Comment 1 of Referee #2);
- We made additional, computationally expensive experiments with a lower dropout rate and a higher learning rate to test new hyperparameter values (number of layers, number of hidden cells per layer, batch size). These tests show that model performances are weakly sensitive to hyperparameter values and that our choices of hyperparameters, which are in line with previous studies, remain optimal (see our response to Comment 9 of Referee #1 and Comment 4 of Referee #2);
- We clarified that the baseline loss function used in literature are part of the loss functions tested in our study, and we enriched the appendices with a replicate of Figure 2 with RMSE as a criterion for test performances to show that our setup does not bias the results in favour of MAE-based loss functions (see our response to Comments 11 and 12 of Referee #1). Finally, we explained in detail the role of parameters μ and λ of the loss function we proposed (see our response to Comments 8, 9 and 10 of Referee #2) and underlined, in the manuscript, their importance in emphasizing the weight of extreme stream temperature values in the training phase.

In the following, we provide a short answer to the Editor's comment in Section 2, a detailed answer to Referee #1's comments in Section 3, and a detailed answer to Referee #2's comments in Section 4.

2 Response to the Editor's comment

Editor's comment: *"[…] That said, I am very pleased with the review round and the provided answers. I am sure the manuscript will improve fast. One personal note in this regard: In your answer to comment 1 by reviewer two you state that the reason for not including time-based features because you want that the LSTM remains similar to process based models. I think this is a fine motivation, but I fail to see how this follows from your stated research goals (Line 114ff in the original manuscript). I am not necessarily saying that any new experiments are required, but rather I propose to rethink the orientation of the manuscript accordingly."*

Authors' response: We thank the Editor for his encouraging decision and nice feedback!

Regarding the motivation behind not using time-based features, it's not that we want that the LSTM be "similar" to process-based models but be comparable, in the sense that process-based models answer the question "Given some atmospheric forcing and a catchment/reach with some properties, what would be the stream temperature?" instead of "Given the day of the year (among other features), what would be the stream temperature?" Keeping the choice of the input features to physically relevant, atmospheric and landscape features in our implementation is hence not contradictory to our stated research goals. In addition, information that could be gained from explicit use of time-based features is already included in some of the input variables of our dataset, namely the catchment-scale potential evapotranspiration and the station-scale air temperature, which feature a very strong seasonality. Finally, our implemented LSTM models, especially the regionally trained ones with static attributes, are already doing a great job by scoring excellent median performances over both the whole test period (median MAE at 0.56-0.57°C and median RMSE at 0.72-0.73°C) and the top 10% range of the test period (median MAE at 0.73-0.74°C and median RMSE at 0.88-0.89°C).

Note that in our answer to Comment 7 of Referee #1, we updated our research questions to better clarify the rationale underpinning our choices of the input features. The corresponding part of the manuscript now reads in the revised version as follows:

"[…] We aimed at answering the following research questions:

- ***To improve the reproduction of extreme, high stream temperature values, what can be gained from increasing the weight of extreme stream temperature values in the loss function used for training?***
- ***How does this strategy compare to a careful selection of the input variables? In particular, what is the contribution of hydrologically relevant variables (namely streamflow)?***
- ***What is the added value of combining regional training with static, catchment and reach attributes in improving the performances of LSTM-based models for high stream temperature values?"***

3 Detailed response to Referee #1's comments

General comment: *"This manuscript addresses the important question of how to improve the performance of LSTM-based models in reproducing extreme stream temperature values. The study focuses on the Garonne river catchment and evaluates three strategies: (i) regional multi-catchment training, (ii) inclusion of static and hydrological variables, and (iii) adaptation of the loss function. The topic is timely and relevant, as accurate modelling of high stream temperatures is critical for ecological and water-management applications.*

The paper is ambitious in scope, draws on a substantial dataset, and tests multiple modelling configurations. It has the potential to contribute meaningfully to the hydrological community by clarifying the role of regionalization and input design for extreme value prediction. However, the manuscript in its current form requires major revision before it can be considered for publication.

Key limitations include the exclusion of essential predictors (notably catchment air temperature and simple temporal features such as day of year or seasonality), an insufficiently clear description of how static variables are incorporated into the LSTM setup, and a narrow framing of the loss-function evaluation that limits the robustness of the conclusions. Together with issues of presentation and readability, these aspects reduce the impact and clarity of the work.

I therefore recommend major revisions. Addressing these issues—by streamlining presentation, clarifying the study's novelty, incorporating or justifying the omission of key predictors, benchmarking against established methods, and refining both methodological detail and evaluation metrics—would substantially strengthen the manuscript and increase its value for the hydrological community."

Authors' response: We thank the Referee for their encouraging and constructive feedback. We did our best to address the key limitations identified by the Referee, specifically:

- We clarified our choices of predictors. In particular, catchment air temperature is linearly dependent on catchment-scale potential evapotranspiration (according to Oudin et al., 2005), which means that, in fact, we did not completely omit this predictor (see our response to Comment 5 in Section 3.2 of this answer). As for temporal features, we intentionally excluded them namely for redundancy and their low relevance compared to input features already included in our study; Our choice is thoroughly explained in our response to Comment 6 (Section 3.2 of this answer).
- We now improved the description of how static attributes are fed to the regionally trained LSTM models (see our response to Comment 10, Section 3.3 of this answer).
- We believe that the framing of the loss function in our study is not that "narrow" as the Referee said, and our application includes a comparison against widely applied (or "established") methods; Please see our responses to Comments 11 and 12 (Section 3.4 of this answer).

Our answers to each of the Referee's comments are provided in the following subsections 3.1 to 3.5.

3.1 Presentation and readability

Comment 1: *"The manuscript is currently too long and dense, which makes it difficult to follow the main arguments."*

Authors' response: We understand that the manuscript could be too long for the Referee's taste, but this is an unavoidable result of our will to be as clear as possible regarding our methodological setup. To satisfy the Referee's request, we merged and reduced two paragraphs of the Introduction section (mainly Section 1.2, see our response to Comment 2), and moved some methodological details on data pre-processing to Appendix A following the Referee's suggestion (see our response to Comment 3).

Comment 2: *"The introduction could be reduced substantially (perhaps to a quarter of its current length), while clearly highlighting the novelty of this work relative to existing literature."*

Authors' response: We would very much like to reduce the length of the Introduction section, but (unfortunately) the Referee did not provide specific details on which parts of the Introduction should be reduced or removed. In the original version, the Introduction section was composed of 7 paragraphs:

- 1 paragraph to highlight the insufficient monitoring of stream temperature despite its socio-economic and ecological importance (Section 1.1);

- 3 paragraphs for a brief overview of process-based and data-driven modelling approaches that are proposed and applied to reconstruct stream temperature records at ungauged locations (Section 1.2);
- 1 paragraph to highlight the research gap left by the existing applications of LSTM for stream temperature modelling (Section 1.3);
- 1 paragraph to summarize the paper's methodology and research questions (Section 1.3); and
- 1 paragraph to summarize the structure of the manuscript (Section 1.3).

To satisfy the Referee's request, we reduced the length of Section 1.2 to 75% of its original length by merging paragraphs 1 and 2 of this section to form one paragraph that briefly reviews the applications of process-based approaches for stream temperature modelling. This new paragraph now reads:

"To overcome this monitoring gap, stream temperature models are typically applied to extend the existing records beyond their temporal coverage or reconstruct missing records at ungauged locations (via model regionalization). These models encode the interactions between stream temperature and other atmospheric and hydrological variables that are more widely available. A first modelling approach consists in explicitly specifying these interactions in the model structure by solving the energy budget at the reach scale. This energy budget accounts for heat advection along the watercourse and heat fluxes at the free surface and at the streambed interface (Caissie, 2006; Dugdale et al., 2017; Leach et al., 2023; Moore et al., 2005). Following this modelling approach, model parameters have a physical meaning and this facilitates the projection of changes in stream temperature in response to climate and landscape changes. Application examples include the characterization of the thermal regimes of large rivers using land surface models (Niemeyer et al., 2018; van Vliet et al., 2013; Wanders et al., 2019), the assessment of the impact of riparian shading at the reach scale (Dugdale et al., 2024), the quantification of heat exchanges at the stream-aquifer interface (Caissie et al., 2014; Kurylyk et al., 2015; Rivière et al., 2020), and the projection of future stream temperature under scenarios of climate drift (Michel et al., 2022). Unfortunately, fully solving the heat budget at the regional, catchment scale is computationally demanding and requires an expensive characterization of stream network morphology and other landscape parameters (such as land-use features). Therefore, process-based approaches resort to adopting several simplifying hypotheses, such as combining the physically based heat balance equation with a statistical approach (Gallice et al., 2015; Toffolon and Piccolroaz, 2015) or using the equilibrium temperature concept to parametrize the heat fluxes at the free surface (Edinger et al., 1968). For instance, variants of this concept have been compared at the Loire river catchment (~10^5 km^2; Bustillo et al., 2014), with advanced model applications that explicitly account for hydrological processes, river network topology (e.g., Strahler order), and riparian vegetation (Beaufort et al., 2016; Seyedhashemi et al., 2023)."

We weren't able to further reduce or remove the remaining paragraphs because (1) the two paragraphs of Section 1.2 are important to provide an overview of stream-temperature modelling approaches, and (2) the two paragraphs of Section 1.3 highlight the research gap and summarize the research questions that are addressed by our study. We remain open to any further explicit suggestions that could help optimize the length of the Introduction section.

Comment 3: *"The description of data collection and preprocessing (e.g. GR6J modelling for discharge) is overly detailed and would be better placed in supplementary material."*

Authors' response: We substantially reduced the description of data pre-processing by moving the details on the reconstruction of streamflow using GR6J to the newly created Appendix A. The part describing the use of streamflow time series in our setup now reads:

"The second set of hydrologically relevant variables is streamflow (Q_{sim}). Since existing streamflow gauging stations did not coincide with the set of T_w stations, we reconstructed streamflow records at each T_w station by feeding the time series of precipitation and potential evapotranspiration to the daily hydrological model GR6J (Pushpalatha et al., 2011), with a parameter transfer approach based on spatial proximity (see Appendix A for details)."

Comment 4: *"Results sections 4.2 and 4.3 repeat exhaustive comparisons of all loss functions, which add little beyond the conclusion already drawn in section 4.1. This makes the results harder to interpret."*

Authors' response: Results in Sections 4.2 and 4.3 are not exactly repeating what has been shown in Section 4.1; Section 4.1 focuses on the contribution of the choice of the loss function, and Sections 4.2 and 4.3 analyze the contribution of regional training and static attributes in improving the performances of LSTM in reproducing extremely high stream temperature values.

We decided to keep the comparison of all loss functions in Sections 4.2 and 4.3 because this comparison allows for verifying that the conclusions on the choice of the input variable set are weakly sensitive to the choice of the loss function. For this reason, we showed the performances considering all loss functions, then considering the reference (or baseline) loss function (MSE on standardized target), and finally considering the best loss function across all sets of input variables. We believe that the interpretation of Figures 3 and 4 is not that difficult given that the most important take-away messages are provided in the accompanying text.

3.2 Input variables and methodological choices

Comment 5: *"A key omission is the absence of catchment-scale air temperature as a predictor. Station air temperature is a proxy that may suffice for small basins but is not adequate for larger catchments where thermal dynamics evolve along the river. This limitation likely explains why models using potential evapotranspiration perform comparatively well, as it implicitly represents catchment-scale air temperature."*

Authors' response: Catchment-scale air temperature is not completely omitted from our setup, because the temperature-based formula that we used to compute the catchment-scale potential evapotranspiration (PE) is (almost) linearly dependent on catchment-scale air temperature (for values higher than -5°C). The daily PE depth (in mm d^{-1}) is computed as follows (Oudin et al., 2005):

$$\mathrm{PE(d)} = \max\left(\frac{\mathrm{R_e}}{\lambda\rho}\frac{\mathrm{T_{a,bv}(d)} + 5}{100}\,;0\right)$$

where $\mathrm{R_e}$ represents the extraterrestrial radiation (MJ m^{-2} d^{-1}), λ represents the latent heat of vaporization (MJ kg^{-1}), ρ represents the water density (kg m^{-3}), and $\mathrm{T_{a,bv}(d)}$ represents the daily catchment-scale average of air temperature. So, this formula clearly shows that the catchment-scale air temperature is not omitted as a predictor (except for situations when this temperature is lower than -5°C, which is less relevant for the paper's main focus of predicting extreme, high stream temperature values).

We used PE instead of catchment-scale air temperature because we wanted to see whether using catchment-scale atmospheric forcing P (precipitation) and PE would lead to comparable performances with using station-scale streamflow, which is physically more relevant as it controls the evolution of thermal dynamics along the river. In addition, comparing the performances between the two options has an important practical application, since P and PE are much more accessible than observed discharge (at least in France). We have explained this choice in Section 3.2.3 of the revised 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 almost linearly dependent on catchment-average air temperature according to Oudin et al., 2005) instead of the more relevant station-scale streamflow (Q_{sim})."

Comment 6: *"No time-based features (e.g. day of year, seasonality) are included, even though prior work (e.g. Feigl et al., 2021, doi.org/10.5194/hess-25-2951-2021) has demonstrated their strong predictive value for stream temperature modelling. These features are straightforward to compute and do not require any additional external datasets. If the authors choose not to include them, it is important to provide a clear justification and to explain why the validity of their results and comparisons is not compromised."*

Authors' response: We agree with the Referee that these features are easily computable and would increase the predictive performance of the LSTM models. However, we intentionally avoided the use of these time-based features because they would overshadow the contribution of more physically relevant variables (air temperature and hydrological variables), knowing that stream temperature has a strong seasonal variability. In addition, since process-based models do not benefit from explicit use of time-based features, we decided not to use them so that LSTM models remain comparable to past and future applications of process-based stream temperature models. Finally, information on seasonality is already contained in the signals of station-scale air temperature and catchment-scale potential evapotranspiration, which means that the information content that would be offered by time-based features is not completely overlooked in our setup. In response to the Referee's comment, we added the following statements at the end of Section 3.2.3 of the revised manuscript to clarify our choice:

"Finally, we avoided using time-based features (month or day of the year; Feigl et al., 2021) so that the performances of the tested LSTM models remain comparable to process-based models that do not benefit from feature engineering. In addition, knowing the strong seasonality of T_w and of some of the input variables (T_a, T_{amn}, T_{amx}, and PE), the use of time-based features would be redundant information-wise and would likely lead to gains in predictive performances high enough to overshadow the contribution of the more physically relevant variables used in our setup."

Comment 7: *"The rationale for testing so many sets of input variables is unclear, as this is not aligned with the stated research questions. Either the scope should be reduced or the research questions reframed."*

Authors' response: We believe that the choices of the sets of input variables are clearly explained in the manuscript, both for the local and for the regional models (please see Section 3.2.2 of the original manuscript version). To satisfy the Referee's request, we reframed the research questions by emphasizing that the selection of input variables is central to the scope of the paper. The research questions now read:

"[…] We aimed at answering the following research questions:

- ***To improve the reproduction of extreme, high stream temperature values, what can be gained from increasing the weight of extreme stream temperature values in the loss function used for training?***
- ***How does this strategy compare to a careful selection of the input variables? In particular, what is the contribution of hydrologically relevant variables (namely streamflow)?***

- ***What is the added value of combining regional training with static, catchment and reach attributes in improving the performances of LSTM-based models for high stream temperature values?"***

Comment 8: *"Why are you predicting daily mean stream temperature values if the stated aim is to model extremes? Since extreme ecological and management impacts are often driven by peak daily temperatures, it would arguably be more appropriate to predict daily maxima rather than means. Please clarify the rationale for focusing on daily mean values, and discuss whether modelling daily maxima might be a more suitable target for assessing extreme conditions."*

Authors' response: We can only agree that daily peak stream temperature values are richer in information and more relevant for assessing extreme conditions than daily mean values. However, to model peak daily temperature values, we need to extract these peaks from sub-daily (e.g., hourly) records of stream temperature, which are obviously more challenging to collect than daily records, and are not always available with a sufficient quality in the Garonne river catchment (to our knowledge). Therefore, we simply focus on daily mean values because we do not have access to records of higher temporal resolution at the scale of the Garonne catchment. Note that these daily averages are still very relevant for water management and ecological applications; For instance, Picard et al. (2022) used daily stream temperature values to compute interannual statistics (average, upper 90% quantile and lower 10% quantile) to implement species distribution models.

This comment invites us to further clarify the aim of our study. Our aim is to look for an LSTM that performs acceptably not only over the whole range of observed daily (average) stream temperature values but also over the range of extreme daily (average) stream temperature values. "Extreme" values are defined here as the top 10% values of the records, or equivalently exceeded 10% of the time at most. These values are encountered mainly during the summer months. We highlighted that this evaluation was absent from applications of LSTM for stream temperature modelling (see Section 1.3 of the original manuscript version). Our results show that if we focus the training of the LSTM model on extreme values (by further penalizing the model errors on the highest temperature values), the LSTM performances are actually worse than when the remaining range is accounted for in the training (see Figure 2 of the manuscript). This suggests that to learn the thermal behaviour during extreme conditions, the LSTM should be first trained on the thermal behaviour frequently observed under "usual" conditions.

For this reason, we further clarified our aim in the Introduction section as follows:

"In our implementation, we focused on strategies to improve LSTM performances for extreme daily stream temperature values (top 10% of the daily observations), while at the same time maintaining satisfactory performances for the remaining range of daily records. Specifically, we compared 18 loss functions, local vs. regional/multi-catchment training, and several combinations of static and dynamic input variables."

In the introductory paragraph of Section 3.2, we defined what we mean by "extreme" stream temperature values in the context of our study:

"In this part, we summarize the strategies that we tested to look for the best approach to improve the LSTM performances at extreme, high T_w values. We define these values as the daily (average) T_w values exceeded less than 10% of the time. Our tested strategies include an adaptation of the loss function to increase the weight of extreme values in the training phase (Sect. 3.2.1), regional multi-catchment training (Sect. 3.2.2), and the inclusion of hydrologically relevant variables and static attributes (Sect. 3.2.3)."

In the Discussion section (Section 5), we highlighted the importance of learning the overall thermal behaviour as a necessary condition to perform well on extreme stream temperature values:

"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."

Finally, among the limitations of our study that we cited in the last paragraph of the Discussion section, we listed the difficulty of modelling daily maxima due to the scarcity of sub-daily records of stream temperature, and discussed their importance in further improving the characterization of extreme conditions:

"Our work can be further improved by addressing some of its limitations. First, our catchment set could be enriched by looking at more catchments with contrasting regional settings, which would shed more light on the regionalization and spatial extrapolation capabilities of LSTM models (see the discussion in Hashemi et al., 2022 and the more rigorous spatial extrapolation tests in Yu et al., 2024). It could also be enriched by collecting records at higher temporal resolutions (e.g., at the hourly timescale), which would enable a better characterization of extreme conditions (van Hamel and Brunner, 2024), and consequently a more relevant assessment of the predictive performances of LSTM models for extreme T_w events."

3.3 LSTM architecture and training details

Comment 9: *"The manuscript states that model performance was insensitive to the number of layers, cells, and batch sizes. This may be an artefact of using a very low learning rate (1e-4) combined with a high dropout rate (0.4). At minimum, additional tests with higher learning rates (e.g. 1e-3) and lower dropout values should be provided."*

Authors' response: We set the dropout rate to 0.1 and the learning rate to 10^{-3} and we made additional, computationally expensive tests to respond to the Referee's request. We trained and tested up to 576 regionally trained models that consist of a combination of the following settings:

- Two values for the number of layers (NL): 1 and 2;
- Two values for the number of cells per layer (HDN_SZ): 128 and 256;
- Two values for the batch size (BTH_SZ): 64 and 256;
- Four loss functions corresponding to μ = 1, λ = 1, 2, with and without standardization of the target variable;
- Six sets of input variables: T_{amn}+T_{amx}, T_{amn}+T_{amx}+CatAttrs, T_{amn}+T_{amx}+P+PE, T_{amn}+T_{amx}+P+PE+CatAttrs, T_{amn}+T_{amx}+Q_{sim}, and T_{amn}+ T_{amx}+Q_{sim}+CatAttrs.
- Three values for lookback: 30, 90, and 365.

Regarding the lookback, we kept only the model with the lookback value that provided the best MSE on the validation set, meaning that only the best 192 models (per lookback) are kept for test. Figure R1 shows that, at best, tuning the number of layers improves the median performances on the top 10% values of the test period, by lowering the median MAE from 0.98°C to 0.95°C. In other words, tuning these hyperparameters (number of layers, number of cells per layers, batch size) has negligible effect on model performances over both the whole test period (R1a to R1c) and the top 10% values (R1d to R1f).

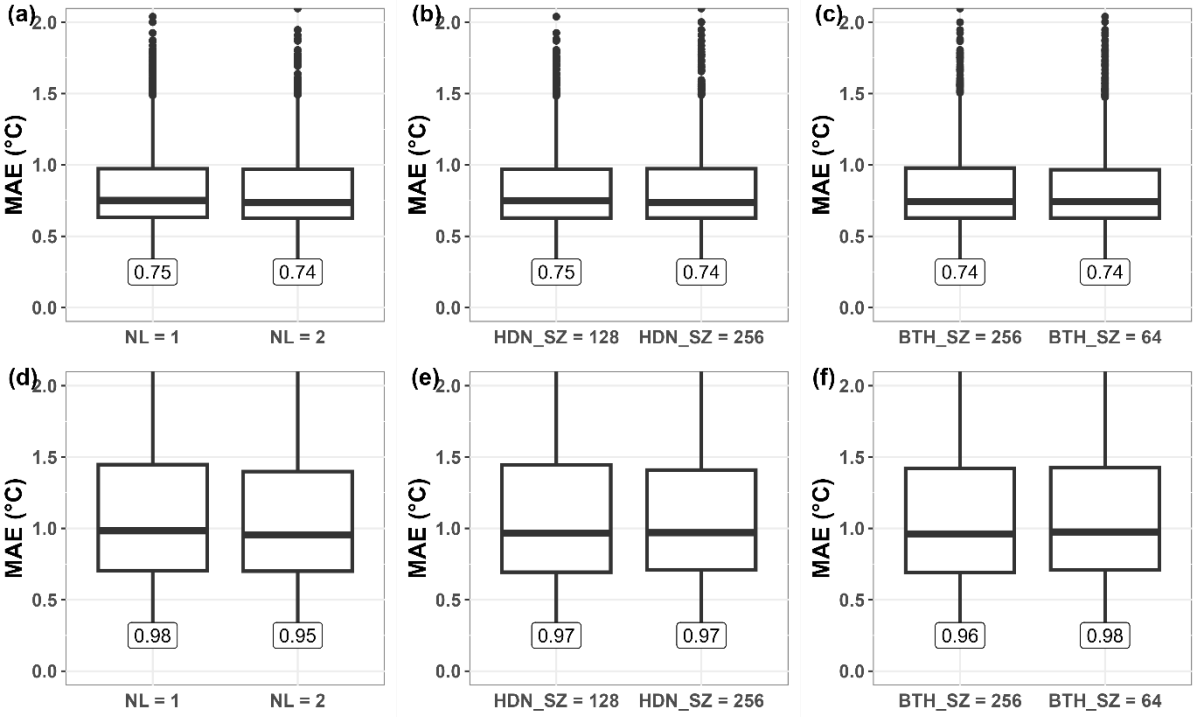

Figure R1: Effect of the number of layers (NL), the number of cells per layer (HDN_SZ), and the batch size (BTH_SZ) on model performances over (a)-(c) the whole test period, and (d)-(f) the top 10% values of the test period. Values under each box indicate the median value. Each distribution contains 2016 points. Note that these model runs are made with a learning rate at 10^{-3} and a dropout rate at 0.1.

Comment 10: *"The description of how static attributes are incorporated into the LSTM is insufficient. Is it via concatenation at each timestep, embeddings, or as additional inputs to the final dense layer? Without this clarity, it is difficult to interpret results."*

Authors' response: We already indicate in the manuscript (last paragraph of Section 3.2.3 of the revised manuscript) that each static attribute was simply repeated at each time step. In other words, we opted for a simple concatenation of static attributes. We rewrote that manuscript part for better clarification by comparing our choice to what has already been done in literature:

"Note that to feed the LSTM model with the static attributes, we opted for a simple integration strategy (see, e.g., Hashemi et al., 2022) in which we repeated the value of each static attribute at each time step to match the length of the dynamic attributes, then we concatenated the columns of the static attributes to those of the dynamic attributes (for each catchment). This strategy compared well against a separate processing of static attributes from dynamic ones using an entity-aware (EA) variant of LSTM networks (Kratzert et al., 2019), and better strategies to encode the static attributes as well as the dynamic variables as inputs to LSTM models have been recently intercompared by Kraft et al. (2025)."

3.4 Loss function evaluation

Comment 11: *"The study introduces several "regional" loss functions but does not benchmark them against published alternatives (e.g. Kratzert et al., 2019, doi.org/10.5194/hess-23-5089-2019) or against a standard MSE baseline. Especially the MSE baseline would be interesting,*

as it is not clear how different catchment water temperature ranges, which do not show as large differences as runoff, affect regional training. This limits the significance of the results."

Authors' response: The loss function tested by Kratzert et al. (2019) has the following expression

$$\mathrm{NSE}^* = \frac{1}{\mathrm{B}} \sum_{\mathrm{b}=1}^{\mathrm{B}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{(\widehat{\mathrm{y_n}} - \mathrm{y_n})^2}{(\mathrm{s(b)} + 0.1)^2}$$

where B is the number of catchments, $\widehat{\mathrm{y_n}} - \mathrm{y_n}$ is the difference between simulated and observed values at time step n, and $\mathrm{s(b)}$ is the standard-deviation of the observations for the catchment b. This loss function is very similar to the MSE loss function

$$\mathrm{MSE} = \sum_{\mathrm{n}} (\widehat{\mathrm{y_n}} - \mathrm{y_n})^2$$

except that NSE* decreases the weight of data points belonging to catchments with high standard-deviation values (or high variances). Looking at Table 2 of Kratzert et al. (2019), this has negligible effect on median performances for regionally trained LSTM models and improves mainly the catchments on which the model has already scored bad performances (which results in improved mean performances). The general expression of the loss functions we tested is

$$\mathcal{L}(\mu, \lambda, \mathrm{g}) = \frac{\sum_{\mathrm{n}} |\mathrm{g}(\mathrm{y_n})^{\mu} - \mathrm{g}(\widehat{\mathrm{y_n}})^{\mu}|^{\lambda}}{\sum_{\mathrm{n}} \left|\mathrm{g}(\mathrm{y_n})^{\mu} - \overline{\mathrm{g}(\mathrm{y_n})^{\mu}}\right|^{\lambda}}$$

the minimization of which is equivalent to the minimization of

$$\mathcal{L}^*(\mu, \lambda, \mathrm{g}) = \sum_{\mathrm{n}} |\mathrm{g}(\mathrm{y_n})^{\mu} - \mathrm{g}(\widehat{\mathrm{y_n}})^{\mu}|^{\lambda}$$

with the LSTM parameters acting only on $\widehat{\mathrm{y_n}}$. Among the values we tested, the configuration $\mu = 1, \lambda = 2$, and $\mathrm{g}(\mathrm{x}) = \mathrm{x}$ gives

$$\mathcal{L}^*(\mu = 1, \lambda = 2, \mathrm{g}(\mathrm{x}) = \mathrm{x}) = \sum_{\mathrm{n}} |\mathrm{y_n} - \widehat{\mathrm{y_n}}|^2 = \mathrm{MSE}$$

This means that the baseline MSE is already included within our tests. The performances of this baseline compare well with the other loss functions, as can be seen in Figure 2 of the manuscript. We did not test the effect of accounting for inter-catchment differences in the temperature range in our loss functions, and we added a sentence by the end of the Discussion section to underline this limitation:

"Finally, our tests of the loss functions are still exploratory at this stage, and fully analysing the potential of this strategy in improving the learning process of LSTM networks for extreme values can include (1) better optimization hyperparameters (e.g., scheduling of the learning rate), (2) training the LSTM on the whole range and then finetuning it on the target range, and (3) designing a custom loss function that is a weighted sum of losses over the whole range and losses over the target range. In the case of regional training, our tested loss functions can also be improved by accounting for differences in T_w ranges between catchments, which can improve the model performances especially for the cases where the LSTM models perform poorly (Kratzert et al., 2019)."

Comment 12: *"Evaluation relies on MAE, which biases the study towards MAE-based loss functions and does not adequately reflect extreme-value performance. Since the research objective is specifically focused on extremes, an evaluation metric more sensitive to high*

values (e.g. RMSE, quantile-based metrics, or extreme value scores) would be more appropriate."

Authors' response: First, we computed two sets of MAE-based scores: (1) a MAE score over the whole test period, which (as pointed out by the Referee) does not adequately reflect extreme-value performance, and (2) a MAE score restricted to the highest 10% values of the test period, which *does* reflect extreme-value performance. These performances are shown in Figures 2 to 5 and highlight the degradation of model performances over the top 10% range, suggesting that it is difficult for the model to consistently score good performances over the whole range of observations. Note that we chose MAE over RMSE because MAE is more interpretable (average of absolute errors) than RMSE (square root of the average of quadratic errors).

Second, we also evaluated the test performances (over the whole range and over the top 10%) using an evaluation metric that is more sensitive to high values: the RMSE. Appendix B of the original manuscript version already shows model performances using RMSE regarding the effect of input selection and local vs. regional training, but they may not respond to the Referee's comment regarding the bias of the study towards MAE-based loss functions. For this reason, we added a replicate of Figure 2 (Figure R2 of the present answer) to Appendix C of the revised manuscript (Figure C1 of the revised manuscript version), which illustrates that MAE-based loss functions also rank as the best in terms of RMSE over the test period (see Figure R2).

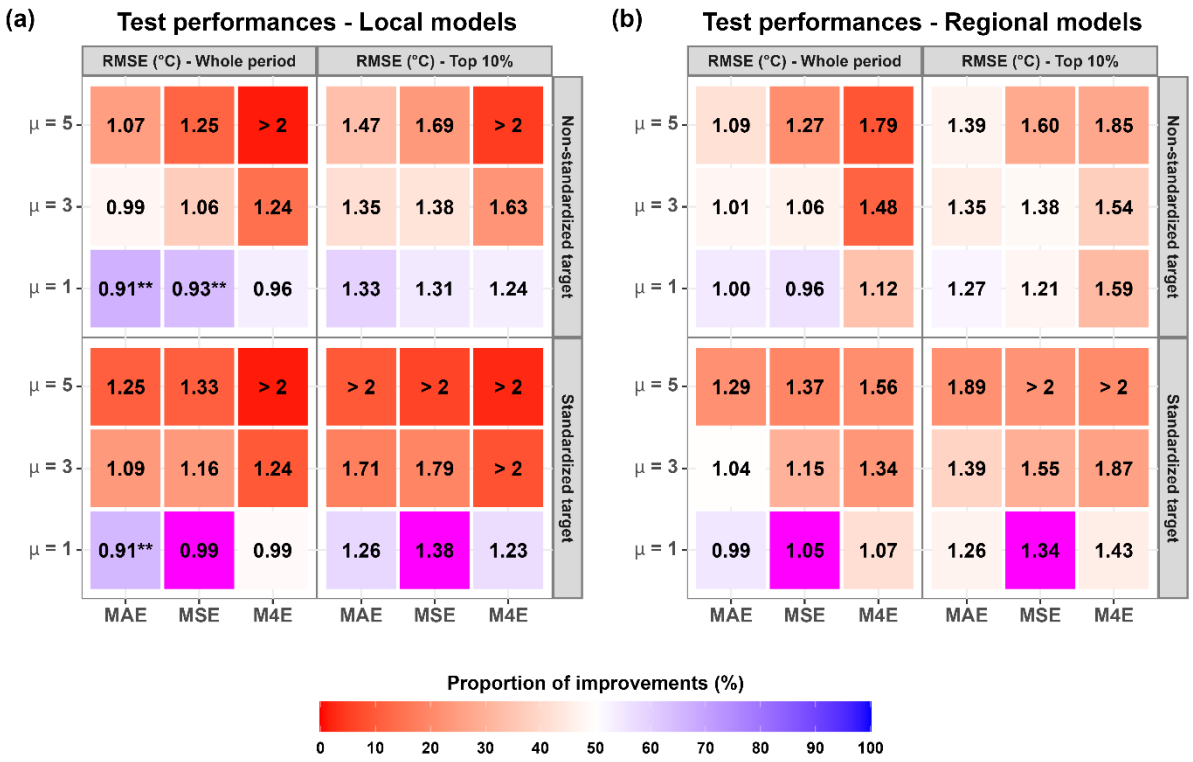

Figure R2: Median test performances in terms of RMSE (in °C) of the local models (a) and the regional models (b) according to the loss function used for training. Colours indicate the proportion of cases for which the use of the loss function yielded better results than the reference loss function (MSE 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%.**

In addition, we modified the summary of Appendix B (now Appendix C in the revised version) to highlight that conclusions regarding the ranking of loss functions did not significantly change with RMSE as a criterion instead of MAE:

"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. 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 a different performance criterion. 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 13: *"As currently presented, the main result is that custom loss functions did not improve performance. However, this may reflect the design of the evaluation rather than a fundamental limitation."*

Authors' response: We disagree with the Referee's comment: This is not *the* main result of the study. The custom loss functions did improve the test performances when compared to the baseline, reference loss function; For instance, in Figures 2 and R2, using MAE (with $\mu = 1$, with/without standardization) significantly improved the results when compared with MSE. However, increasing the weight of high stream temperature values in the training phase unexpectedly degraded not only the performances over the top 10% range but also the overall test performances. Despite these improvements, a better strategy consists in training the LSTM over multiple catchments with the static attributes included in the set of input variables: This is the main result of the study.

Nevertheless, we already stated in the Discussion section that our setup might be held responsible for not succeeding in improving the test performances over the (extreme) 10% range of the observations, and that possibly alternative optimization hyperparameters and/or alternative training procedures should be tried:

"Finally, our tests of the loss functions are still exploratory at this stage, and fully analysing the potential of this strategy in improving the learning process of LSTM networks for extreme values can include (1) better optimization hyperparameters (e.g., scheduling of the learning rate), (2) training the LSTM on the whole range and then finetuning it on the target range, and (3) designing a custom loss function that is a weighted sum of losses over the whole range and losses over the target range."

We also refer to our response to Comment 12 regarding possible biases of interpretation in favour of MAE-based loss functions.

3.5 Technical corrections

Comment 14: *"Abstract: Key results (1) and (2) appear redundant since regional modelling is inherently linked to extended static inputs. Please clarify."*

Authors' response: Regional training refers to the use of data from multiple catchments to train one LSTM model. These data may or *may not* include static attributes. One of the added values of our study is quantifying the contribution of these static attributes compared to information already contained in dynamic attributes (see, for example, Yu et al., 2024). This is why we emphasized this result in the Abstract. More precisely, Figure 4 shows results of regionally trained LSTM models with three pairs of sets of input variables that help us quantify the added value of static attributes:

- T_{amn}+T_{amx} vs. T_{amn}+T_{amx}+CatAttrs;
- T_{amn}+T_{amx}+P+PE vs. T_{amn}+T_{amx}+P+PE+CatAttrs; and
- T_{amn}+T_{amx}+Q_{sim} vs. T_{amn}+T_{amx}+Q_{sim}+CatAttrs.

Figure 4 shows that

- Looking at all loss functions, median MAE values over the whole test period for regionally trained LSTM models with only dynamic variables as inputs were at 1.08°C, 0.98°C, and 1.00°C respectively for T_{amn}+T_{amx}, T_{amn}+T_{amx}+P+PE, and T_{amn}+T_{amx}+Q_{sim}. By adding static attributes (CatAttrs), these median performances decreased down to 0.88°C, 0.69°C, and 0.77°C, respectively. Over the extreme values (top 10%), median MAE values decreased from 1.41°C, 1.36°C, and 1.35°C to 1.24°C, 1.05°C and 1.00°C thanks to the additional use of static variables.
- These gains are more important when looking at selected loss functions, namely the MSE baseline ("Reference") and the "best" loss function in the sense of the best median MAE across all sets of input variables. Thanks to static attributes, median whole-period performances went from 0.89°C with the input variables T_{amn}+T_{amx}+P+PE to 0.56°C with the input variables T_{amn}+T_{amx}+P+PE+CatAttrs with the baseline MSE as a loss function. In terms of median performances over the top 10% values, median performances went from 1.47°C to 0.74°C.
- By comparing Figure 4 with Figure 3, which shows the performances of locally trained models, we can see that regional training actually deteriorated the overall performances of LSTM models that did not use static attributes, highlighting the key importance of these attributes in regional training.

This last result is highlighted in Section 4.3 that comments the performances of regionally trained LSTM models:

"In general, simply training the LSTM at the regional scale led to deteriorated median performances, as can be seen by comparing regionally trained models without static attributes with their counterparts in Fig. 3 (all loss functions and reference loss function)."

Comment 15: *"Abstract: The phrase "well-trained LSTM" is vague—better to define relative to baseline approaches."*

Authors' response: We have reformulated that phrase as follows:

"This study further confirms the suitability of regionally trained LSTM models that exploit static attributes for the reproduction of extreme stream temperature values, offering significant advantages for water management at data-sparse regions during summer periods."

Comment 16: *"Line 126: Why do you need exactly a minimum of 2434 daily observations for 1 test year?"*

Authors' response: The answer to this question is given in the phrase immediately following! Reading Line 125-127 of the original manuscript version:

"Among these stations, only 21 stations have more than 2434 daily observations of T_w, which is required to ensure a minimum of one year (365 days) of datapoints for model testing (see Sect. 3)."

In Section 3, we explained that for each station, 70% of the data is dedicated for model training, 15% for model validation, and 15% for model testing. If we want to guaranty at least 365 datapoints for model testing, we need at least 365/0.15 = 2433.33 or 2434 days. We modified the sentence in question by adding more details:

"Among these stations, only 21 stations have more than 2434 daily observations of T_w, which is required to ensure a minimum of one year (365 days) of datapoints for model testing (see Sect. 3.2; 15% of the available T_w datapoints are dedicated to model testing, therefore we require a minimum of 365/0.15 = 2433.33 or 2434 datapoints in total). We call these 21 stations "test stations" since they are the only stations with enough datapoints to allow for robust model testing (see Sect. 3.2.2 for more details)."

Comment 17: *"Line 127: "We call these 21 stations test station" – please clarify whether this refers to an ML-style train/validation/test split. Overall, it is not entirely clear to me how you split the data, especially not in which situations you split the time series or split by stations? Please state this more clearly."*

Authors' response: We call these test stations because they are the only stations on which LSTM models are tested, since they satisfy the requirement of minimum data availability. More details cannot be provided in this section that is dedicated to present the dataset. Instead, these details should be looked for in the methodology section (namely, Section 3). We modified this line to refer to Section 3.2.2 of the revised manuscript where all these methodological choices are clarified:

"We call these 21 stations "test stations" since they are the only stations with enough datapoints to allow for robust model testing (see Sect. 3.2.2 for more details)."

In Section 3.2.2 of the revised manuscript, we explain the difference between locally trained and regionally trained models, and we provide more details on our methodological setup that better clarify the reason why we chose to call these stations "test stations". Since we have to test all models (local or regional) on the same datapoints, we chose only stations that had enough datapoints, i.e., a minimum of 2434 datapoints (see our response to Comment 16), for model testing, hence the name "test stations". More precisely,

1. for each of these 21 stations, a local model was trained on 70% of the data, validated on 15% of the data, and tested on the remaining 15% of the data. In total, 21 local models were trained (for each configuration of loss function × set of input variables × lookback value).
2. Then, we trained one regional model (again, for each configuration of loss function × set of input variables × lookback value) over a collection/concatenation of training data from all the 21 stations, to which we added 70% of the data from the remaining non-test stations (16/37). We validated this regional model over a collection of validation data from all the 21 stations, to which we added 30% of the data from each record of the remaining non-test stations. In other words, each non-test station contributes with 70% of its data to the training data collection for the regional model, and with 30% of its data to the validation data collection. Finally, the regional model is tested over the same datapoints as the local models.

Section 3.2.2 of the revised manuscript re-states these clarifications as follows:

"In this regard, for each test station (21 in total), we compared two different sets of models:

1. ***The set of local models trained using the first 70% of the available T_w records and their corresponding input, dynamic variables only at the station of interest. In this case, half of the remaining T_w observations (i.e., 15% of the available records) were used for validation and the remaining records (i.e., 15% of the available records) were kept for test. Note that for these stations, 15% of the available records span at least one-year worth of daily observations.***
2. ***The set of regional models trained using data from all the 37 stations. In this case, we constructed the training set by concatenating 70% of the available T_w and their corresponding input variables from each station. The validation set was***

constructed using 15% of observations at the test stations (21/37) and the whole remaining 30% of observations at the non-test stations (16/37). Finally, the remaining 15% of observations at the test stations were used to test the regional models, thus enabling a comparison of locally and regionally trained models on the same datapoints. Although 30% of the observations at the non-test stations were used in the validation set against 15% from the test stations, datapoints from the test stations still constituted up to 78% of the validation set due to the lower availability of T_w records at non-test stations."

Comment 18: *"Table 1: Consider presenting mean and range (min–max) values per train/test group instead of medians only, which are not necessarily more robust here."*

Authors' response: Table 1 of the manuscript already includes the range, i.e. min and max values, computed using the whole set of 37 stations (except for the stream temperature statistics, for which non-test stations were excluded because they did not have enough datapoints to provide robust statistics). We believe that providing the range + the median values is enough to get a concise and informative description of the distribution of the features of our catchment set. Table 1 already contains a lot of information, and adding more statistics (i.e., mean) per each group of train/test data would only burden Table 1 and make its reading unnecessarily more challenging without significantly improving the description of the dataset.

4 Detailed response to Referee #2's comments

4.1 General and specific comments

General comment: *"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."

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

Regarding the novelty of our results, we agree that improving the LSTM performances with static attributes and regional training is not novel (e.g., Kratzert et al., 2024). However, our setup quantifies the contribution of including static attributes in comparison with just training the LSTM at many stations, which highlights the mechanisms by which LSTM models perform better when trained regionally (in this regard, see e.g. Yu et al., 2024). In fact, a simple regional training of LSTM models is detrimental in terms of test performances (for instance, compare the results shown in Figure 3 with those shown in Figure 4 with only dynamic attributes). Only when static attributes are included that regional training becomes more efficient than local training.

Regarding the specific comments and suggestions, below we provide a detailed answer.

Comment 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."*

Authors' response: Catchment-scale potential evapotranspiration (PE) is one of the required inputs to the hydrological model GR6J that we used to reconstruct streamflow (Q_{sim}) at the location of stream temperature stations. In the PE formula by Oudin et al. (2006), PE is (almost) linearly dependent on catchment-average air temperature, meaning that the use of PE as input to the LSTM model provides almost the same information content as that of catchment-average air temperature. Furthermore, we wanted to compare (1) the option of letting the LSTM model decide which information to extract from precipitation (P) and PE that is most relevant to the reproduction of stream temperature, against (2) the option of restraining this hydrological input to the simulated streamflow Q_{sim} only, which somewhat represents a hydrologically digested version of P and PE, and also a more relevant hydrological variable (since it is at the station scale) than P and PE (which are at the catchment scale). Figures 3 and 4 show that these two options provide comparable or similar performances, highlighting their equivalence for the reproduction of daily stream temperature values.

As for the use of the day of the year, we intentionally excluded any features explicitly based on time (day or month of the year), which were used e.g. by Feigl et al. (2021). We fear that the inclusion of these time-based features might overshadow the importance of more physically relevant variables (namely station-scale air temperature and hydrological variables), knowing that stream temperature is a variable that features a strong seasonal variability. In addition, these time-based features are generally not used as forcing by process-based models, which would hinder a fair comparison between LSTM-based models and process-based models, or could suggest that the better performances of LSTM-based models are attributable to the "excessive" reliance on feature engineering. We clarified these choices in the revised manuscript version, as we stated by the end of Section 3.2.3 of the revised manuscript that:

"Finally, we avoided using time-based features (month or day of the year; Feigl et al., 2021) so that the performances of the tested LSTM models remain comparable to process-based models that do not benefit from feature engineering. In addition, knowing the strong seasonality of T_w and of some of the input variables (T_a, T_{amn}, T_{amx}, and PE), the use of time-based features would be redundant information-wise and would likely lead to gains in predictive performances high enough to overshadow the contribution of the more physically relevant variables used in our setup."

Comment 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."*

Authors' response: We believe that the Referee's comment refers to static attributes, as including the values of dynamic variables is unpractical. For static attributes, we believe that Table 1 provides a concise and fairly informative description of the richness of our dataset using the range (min and max values) and the median values of the geographical and climatic features of our catchment set. Adding a fourth appendix to show the values for each one of the 37 stations would only unnecessarily increase the length of the manuscript. Note that these data are provided in the Zenodo repository (https://zenodo.org/records/15864784, file "`data/2024-09-09_ListStationsTw_StaticDesc_v03.TXT`") accompanying the manuscript submission that contains all necessary scripts used to run our experiments (Saadi, 2025).

Comment 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."*

Authors' response: To test the contribution of an additional static attribute we would have to re-run all the 378 (7 × 18 × 3) regional models, which is computationally very expensive (see Appendix B of the revised manuscript for wall-clock times needed for training). Adding static attributes that are computed from the target variable that we want to reproduce by the LSTM is, in our sense, would result in an implicit data contamination (or leakage), thus resulting in (a priori) better performing but less robust LSTM models. In addition, this will substantially hinder any application of the regionally trained models at ungauged locations, where stream temperature records (hence any statistic of stream temperature needed for that kind of model applications) is absent.

Comment 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."

Authors' response: We agree with the Referee that it would be useful to do a more detailed assessment of the effect of hyperparameters (number of layers, number of cells per layer, batch size, etc.), but we already found in our preliminary tests that these have little effect on model performances. Thus, we made choices that are in line with previous studies (Hashemi et al., 2022; Kratzert et al., 2019; Rahmani et al., 2021a, 2021b). These choices are also in line with the best choices found by Kraft et al. (2025).

In our response to Comment 9 made by Referee #1 (see Section 3.3 of the present answer), we made additional, computationally expensive tests using a higher learning rate of 10^{-3} and a dropout rate at 0.1 with the following configurations:

- Two values for the number of layers (NL): 1 and 2;
- Two values for the number of cells per layer (HDN_SZ): 128 and 256;
- Two values for the batch size (BTH_SZ): 64 and 256;
- Four loss functions corresponding to $\mu = 1$, $\lambda = 1, 2$, with and without standardization of the target variable;
- Six input-variable sets: T_{amn}+T_{amx}, T_{amn}+T_{amx}+CatAttrs, T_{amn}+T_{amx}+P+PE, T_{amn}+T_{amx}+P+PE+CatAttrs, T_{amn}+T_{amx}+Q_{sim}, and T_{amn}+ T_{amx}+Q_{sim}+CatAttrs.
- Three values for lookback: 30, 90, and 365.
- All models are trained regionally.

This amounted to training 2 × 2 × 2 × 4 × 6 × 3 = 576 models using the same methodological choices as in our main study. Note that we kept only the best lookback value based on the MSE of the validation period, hence showing the test performances for only 192 models. Figure R1 of Section 3.3 shows that tuning any of the three hyperparameters (number of layers, number of cells per layer or the batch size) has negligible effect on model performances over the whole test period (Figures R1a to R1c) and over the top 10% values of the test period (Figures R1d to R1f).

Finally, implementing the hyperparameter tuning experiment by Kraft et al. (2025) is computationally challenging in our case, noting that we tested 18 loss functions (compared to one loss function in Kraft et al. (2025)), several choices of the input variables, and local vs. regional training.

In the revised manuscript version, we cited the study of Kraft et al. (2025) as part of the very recent literature that applied LSTM models, and also to highlight their method in embedding static features for regional training, both in Section 1.2:

"Among these techniques, models based on LSTM (Long Short-Term Memory; Hochreiter and Schmidhuber, 1997) have demonstrated excellent performances in predicting not only stream temperature but also several other dynamic, environmental variables (Arsenault et al., 2023; Kraft et al., 2025; Kratzert et al., 2018; Ma et al., 2021; Nearing et al., 2024; Song et al., 2024; Zhi et al., 2021)."

and in Section 3.2.3 of the revised manuscript to highlight their assessment of different strategies of embedding static attributes as inputs to LSTM models:

"Note that to feed the LSTM model with the static attributes, we opted for a simple integration strategy (see, e.g., Hashemi et al., 2022) in which we repeated the value of each static attribute at each time step to match the length of the dynamic attributes, then we concatenated the columns of the static attributes to those of the dynamic attributes (for each catchment). This strategy compared well against a separate processing of static attributes from dynamic ones using an entity-aware (EA) variant of LSTM networks (Kratzert et al., 2019), and better strategies to encode the static attributes as well as the dynamic variables as inputs to LSTM models have been recently intercompared by Kraft et al. (2025)."

Comment 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."*

Authors' response: We agree with the Referee's comment regarding the inherent randomness in terms of model performances due to the randomly assigned initial values of model parameters prior to training. However, we have several reasons not to make a detailed assessment of the effect of this randomness on our conclusions, as requested by the Referee.

First, there is no reason for this randomness to introduce a bias in favour of one of the options that we tested. For example, in Figure 2 of the revised manuscript, for each loss function, median statistics were computed from $4 \times 21 = 84$ locally trained models (Figure 2a) and $7 \times 1 = 7$ regionally trained models (Figure 2b), meaning that the variability in LSTM performances due to random initialization is already assessed for each loss function thanks to these repetitions induced by several choices of input variables, under the (most likely valid) assumption that the effect of model initialization is independent from the choice of the input variables. The same can be said for the comparison of the different choices of input variables for locally trained models and regionally trained models shown in Figures 3 and 4 of the manuscript.

Second, we used a binomial test to assess the significance of the differences in performances between the different sets of options (Fidal and Kjeldsen, 2020; Saadi et al., 2021). This binomial test does not specifically look at the magnitude of improvements, but at the number of times an option A (say a loss function or a set of input variables) performs better than another option B. When option B performs systematically better than option A (hence a significant binomial test), there is strong reason to believe that option B brings significant improvements compared to option A. We applied this statistical test to compare all our options, and letters in Figures 3 and 4 for example show that the use of hydrological variables and static attributes significantly improves the LSTM performances (in the sense of the binomial test).

Third, it would take us months to provide, for each model setup, a detailed assessment of the variability in terms of model performances due to random initial parameter values (cf. the wall-

clock times needed to train local and regional models in Appendix B of the revised manuscript version). In a preliminary work, we attempted at getting an order of magnitude of the changes in model performances due to this random initialization of model parameters. We retrained 100 times a regionally trained model on 26 stations with a sequence length of 16 days. Note that these choices are way different than the final setup used in our study. Test results for 10 stations show that the standard-deviation can be up to 0.07°C, with fluctuations (i.e., difference between min and max performances) reaching up to 0.35°C in MAE, as can be seen in Figure R3. The improvements that the static attributes brought in in terms of median MAE are way larger than these values, as can be seen in Figure 4 of the original manuscript.

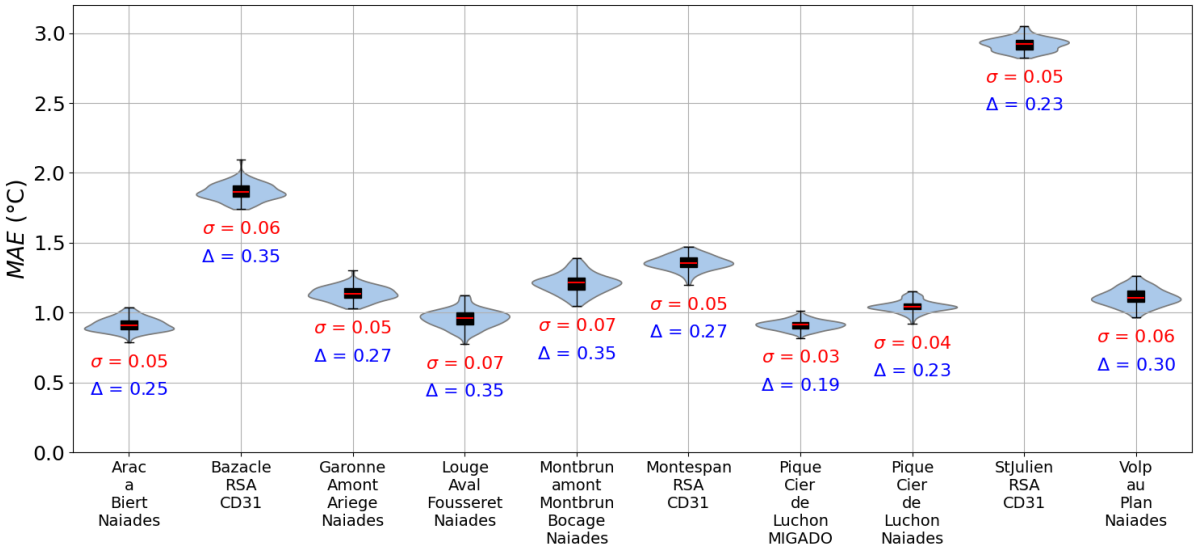

Figure R3: Effect of random parameter initialization on model performances using an ensemble of 100 regionally trained models. For each distribution, values in red refer to the standard deviation, and values in blue refer to the range (max-min).

Comment 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."*

Authors' response: We believe that our choice of using 30% of the observations from the non-test stations instead of 15% adds only to the robustness of our regionally trained models. Furthermore, the total number of datapoints from the 21 test stations is 91705 days, of which 13755 (15% of 91705) were used for validation. The total number of datapoints from the non-test stations is 12626, of which 3788 (30% of 12626) were used for validation. This means that in our configuration, datapoints from the test stations constituted up to 78% (13755/(13755+3788)) of the total number of points on which the regional models were validated. Lowering the contribution of the non-test stations from 30% to 15% would only lead to increasing the dominance of test stations in the validation data from 78% to 88%, and unnecessarily decreasing the total number of points on which the regional models are validated. Finally, the length of available records of stream temperature in non-test stations ranges from 73 to 2365 days, and these datapoints do not span a continuous year for at least half of the non-test stations.

We modified Lines 240-244 of the original manuscript version to emphasize that the test stations still dominate the validation set. Now this part of the text reads:

"[…] The set of regional models trained using data from all the 37 stations. In this case, we constructed the training set by concatenating 70% of the available T_w and their corresponding input variables from each station. The validation set was constructed using 15% of observations at the test stations (21/37) and the whole remaining 30% of observations at the non-test stations (16/37). Finally, the remaining 15% of observations at the test stations were used to test the regional models, thus enabling a comparison of locally and regionally trained models on the same datapoints. Although 30% of the observations at the non-test stations were used in the validation set against 15% from the test stations, datapoints from the test stations still constituted up to 78% of the validation set due to the lower availability of T_w records at non-test stations."

Comment 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."*

Authors' response: Although this is not necessary to follow the methodological setup, we moved Section 3.2.3 of the original manuscript to 3.2.1, Section 3.2.1 to 3.2.2, and Section 3.2.2 to 3.2.3. The introductory paragraph of Section 3.2 now summarizes this part as follows:

"In this part, we summarize the strategies that we tested to look for the best approach to improve the LSTM performances at extreme, high T_w values. We define these values as the daily (average) T_w values exceeded less than 10% of the time. Our tested strategies include an adaptation of the loss function to increase the weight of extreme values in the training phase (Sect. 3.2.1), regional multi-catchment training (Sect. 3.2.2), and the inclusion of hydrologically relevant variables and static attributes (Sect. 3.2.3)."

Comment 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

Authors' response: Looking at the "Loss" column, we believe that the Referee is not computing exactly the same loss function as ours: Instead of computing the ratio of the sums of errors $\frac{\sum_t \left|T^{\mu}_{w,obs,t} - T^{\mu}_{w,sim,t}\right|^{\lambda}}{\sum_t \left|T^{\mu}_{w,obs,t} - \overline{T^{\mu}_{w,obs}}\right|^{\lambda}}$, the Referee computed point-wise fractions of errors at each point, i.e., $\frac{\left|T^{\mu}_{w,obs,t} - T^{\mu}_{w,sim,t}\right|^{\lambda}}{\left|T^{\mu}_{w,obs,t} - \overline{T^{\mu}_{w,obs}}\right|^{\lambda}}$, and of course these ratios are higher for points closer to the mean of observations than points farther from the mean, like the extreme values. An alternative, more rigorous way of evaluating the relative importance of datapoints in the training phase is to compute the magnitude (absolute value) of the sensitivity of the loss function to the simulated stream temperature value, which gives in our case:

$$\left|\frac{\partial \mathcal{L}}{\partial T_{w,sim,t}}\right| = \frac{\lambda\mu T_{w,sim,t}^{\mu-1}\left|T_{w,sim,t}^{\mu} - T_{w,obs,t}^{\mu}\right|^{\lambda-1}}{\sum_t \left|T_{w,obs,t}^{\mu} - \overline{T_{w,obs}^{\mu}}\right|^{\lambda}}$$

Let us consider the time steps H (for high) and L (for low) where a high, extreme stream temperature $T_{w,obs,H}$ and a low stream temperature $T_{w,obs,L}$ are observed. The ratio of the absolute sensitivity values of the loss function to model simulations at high vs. low stream temperature observation time steps can be written as

$$r_{H/L} = \frac{\left|\frac{\partial \mathcal{L}}{\partial T_{w,sim,H}}\right|}{\left|\frac{\partial \mathcal{L}}{\partial T_{w,sim,L}}\right|} = \left(\frac{T_{w,sim,H}}{T_{w,sim,L}}\right)^{\mu-1}\left|\frac{T_{w,sim,H}^{\mu} - T_{w,obs,H}^{\mu}}{T_{w,sim,L}^{\mu} - T_{w,obs,L}^{\mu}}\right|^{\lambda-1}$$

We can see that the denominator plays only the role of a normalizing constant in the loss function, and does not affect the relative importance of model simulations during the training phase because it cancels out in $r_{H/L}$. If we consider the general case, $r_{H/L}$ is proportional to the errors of the simulations when $\lambda > 1$ (i.e., for MSE and M4E in our case). But to understand the weight of extreme vs. usual values (i.e., values closer to the mean), we can consider a simple case where we have a constant shift Δ between the observations and the simulations, i.e., that $T_{w,sim,t} = T_{w,obs,t} + \Delta$ for all time steps t. In this case, normally, if the loss function is somewhat "egalitarian", the ratio should be close to 1; if it emphasizes extreme large values, it should be much higher than 1, and if it emphasizes low values, it should be much lower than 1. The expression of the ratio becomes

$$r_{H/L} = \left(\frac{T_{w,obs,H} + \Delta}{T_{w,obs,L} + \Delta}\right)^{\mu-1}\left|\frac{\left(T_{w,obs,H} + \Delta\right)^{\mu} - T_{w,obs,H}^{\mu}}{\left(T_{w,obs,L} + \Delta\right)^{\mu} - T_{w,obs,L}^{\mu}}\right|^{\lambda-1}$$

A numerical application for $\Delta = 1°C$, $T_{w,obs,L} = 15°C$ and $T_{w,obs,H} = 30°C$ illustrates the evolution of this ratio in Table R1. We can see that with $\mu = 1$, there is in this simple case no overweighting of the extreme value $T_{w,obs,H}$ compared to the low value $T_{w,obs,L}$. As λ and especially as μ increases, the relative importance of the extreme value compared to the average value skyrockets.

Table R1: Values for the ratio $r_{H/L}$ with $\Delta = 1°C$, $T_{w,obs,L} = 15°C$ and $T_{w,obs,H} = 30°C$ computed for all combinations of λ and μ values tested in our study.

	μ = 1	μ = 3	μ = 5
λ = 1 (MAE)	1	4	14
λ = 2 (MSE)	1	15	211
λ = 4 (M4E)	1	218	47269

In summary, the denominator does not theoretically impact the relative importance of the extreme values in the loss function and plays only the role of a normalizing constant. In our opinion, it's important to keep this denominator because (1) it helps interpret the loss function, and (2) without this denominator the learning rate should be modified to account for the large magnitudes of errors with larger values of μ (and λ). We added a sentence in the revised manuscript to justify this choice:

"For an intuitive interpretation of this function, the denominator in the loss function of Eq. (1) standardizes the values of the loss function by comparing the performances of the LSTM model to a "dummy" model that gives the average value of the transformed observations $\overline{g(T_{w,obs})^{\mu}}$ as a prediction for all the time steps. This denominator plays also the role of a normalizing constant without which the learning rate should be

adapted to account for the larger error magnitudes in the numerator induced by higher values of λ and especially μ. This last hyperparameter μ results in magnitudes of $g(T_w)$ that are higher at extreme values than at mild values, inducing larger errors at (and thus more emphasis on) extremely high values."

Comment 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."*

Authors' response: Exactly! We give more explanations regarding the role of μ (and λ) in our answer to Comment 8.

In the manuscript, we added a sentence to give an interpretation of the role of μ in Equation 3 as suggested by the Referee:

"For an intuitive interpretation of this function, the denominator in the loss function of Eq. (1) standardizes the values of the loss function by comparing the performances of the LSTM model to a "dummy" model that predicts the average value of the transformed observations $\overline{g(T_{w,obs})}^{\mu}$ for all the time steps. This denominator plays also the role of a normalizing constant without which the learning rate should be adapted to account for the larger error magnitudes in the numerator induced by higher values of λ and especially μ. This last hyperparameter μ results in magnitudes of $g(T_w)$ that are higher at extreme values than at mild values, inducing larger errors at (and thus more emphasis on) extremely high values."

Comment 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."*

Authors' response: In response to Comment 9, we added a sentence that helps understand the role of μ in the loss function. Please see our answer to Comment 9 and also Comment 8 for a numerical illustration of the role of μ (and also λ) in the loss function.

Comment 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."*

Authors' response: First, the total number of cases is not 21, but 21 times the number of sets of input variables, which is 4 for the local models, and 7 for the regional models. This is why in Figure 2, we represented the percentage of cases for which the use of a loss function improves on the use of the reference loss function (MSE on standardized target). Anyway, we added the number of cases when commenting Figure 2 to better understand the results of the statistical test. This part now reads:

"[…] In detail:

- ***The best performances over the whole period were systematically obtained using MAE (with or without standardization) as a loss function in the training: median test MAE reached 0.72°C for the local models (all configurations of input variables combined, Fig. 2a) and 0.77°C for the regionally trained models (Fig. 2b). In comparison to the reference loss function (MSE on standardized target), these improvements with MAE as a loss function were statistically significant only in the case of locally trained models, for which MAE applied to standardized and non-standardized target resulted in better performances than the reference loss function for 58/84 and 54/84 cases, respectively (recall that the number of cases is the number of the test stations times the number of sets of input***

variables, see Table 2). For the regional models, MSE without standardization is the only loss function that resulted in statistically better performances, with better scores than the reference loss function for 85/147 cases, which is higher in this case than the 5%-significance threshold (84/147).

- ***Compared to the whole-range performances, model performances were systematically lower on the top 10% range. For this range, the best median performances were obtained using MAE/M4E with standardization for the local models (1.07°C, Fig. 2a) or using MSE without standardization for the regional models (0.98°C, Fig. 2b). However, these improvements were not statistically significant at the 5%-level in comparison with the reference loss function (MSE with standardization): For the local models, MAE with standardization resulted in better performances compared to the reference loss function for only 48/84 cases, which is below the 5%-significance threshold (51/84); For the regional models, MSE without standardization gave better scores than the reference loss function for only 70/147 cases."***

4.2 Minor comments and technical corrections

Comment 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."

Authors' response: We thank the Referee for this relevant and timely suggestion. We added a citation of the work of Padrón et al. (2025) by the end of Section 1.2 to cite examples oriented towards forecasting tasks:

"Other studies demonstrated successful applications of LSTM networks for the more operational task of stream temperature forecasting (e.g., Padrón et al., 2025; Qiu et al., 2021; Zwart et al., 2023)."

Comment 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."*

Authors' response: In Line 381 of the manuscript, these values represent the min-max values for the station in question (Garonne at Valentine). The values min-max of Table 1 represent the min-max of the long-term average of stream temperature. In other words:

- We computed the long-term average of stream temperature for each station, i.e., the average of all records. This provides a sample of 37 values (for 37 stations);
- Since some stations have a high rate of missing values, the averages may not be representative. For this reason, we excluded the 16 non-test stations, which have a number of available stream temperature observations lower than 2434 days;
- We finally calculated the min, max and median values using the remaining 21 values.

To add these clarifications, we modified the title of Table 1 as follows:

"Table 1: Summary of dynamic and static variables and their distributions across the catchment set. Min, median, and max values were computed from the set of 37 stations, except for the long-term averages of daily stream temperature."

Additionally, at the bottom of Table 1, we added the following note regarding the values for daily stream temperature:

"aWe first computed, for each station, the long-term average of stream temperature values using the whole time series, which gave 37 values. We then excluded the stations with short time series by computing min, median, and max statistics using the 21 test stations only."

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

Authors' response: We now mention in those lines that 15% of the available records represent at least one-year worth of observations:

"[…] The set of local models trained using the first 70% of the available T_w records and their corresponding input, dynamic variables only at the station of interest. In this case, half of the remaining T_w observations (i.e., 15% of the available records) were used for validation and the remaining records (i.e., 15% of the available records) were kept for test. Note that for these stations, 15% of the available records span at least one-year worth of daily observations."

Comment 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."*

Authors' response: Yes, the bottom row of Figures 3 and 4 corresponds to the best loss function defined as the one with the best median value over all sets of input variables. We now clarified this in the captions of Figures 3, 4, C2, and C3 of the revised manuscript. For example, the caption of Figure 3 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 distribution 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, 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."

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

Authors' response: We thank the Referee for underlining this typo, which we have corrected in the revised version.

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

Authors' response: We now reduced the size of the black dots to improve the visualization of Figure 5 of the manuscript (see Figure R4).

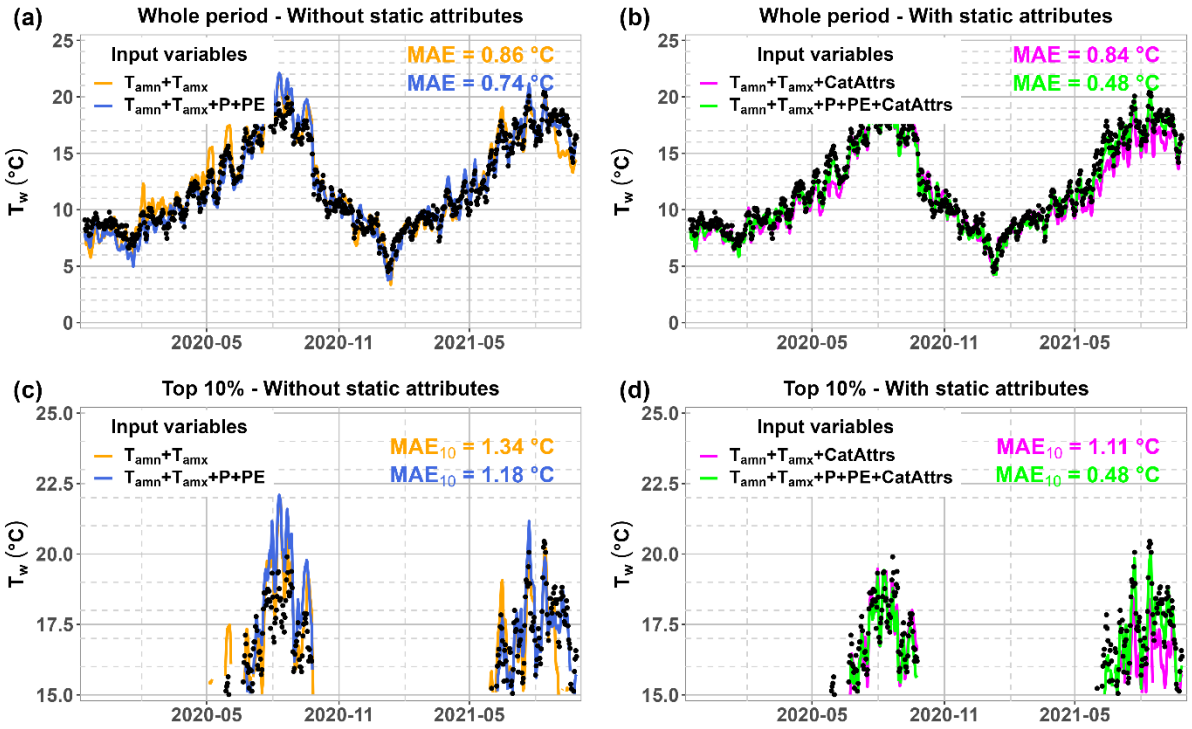

Figure R4: Revised Figure 5 of the manuscript, which shows an example of model simulations using LSTM models against observations at the station of the Garonne at Valentine (MIGADO). All models are regionally trained. The two top figures (a and b) show the model simulations and performances (MAE, in °C) over the whole test period, and the two bottom figures (c and d) zoom in on the highest 10% of the observations during the test period. Input variables include station-scale daily minimum and maximum air temperatures (T_{amn}+T_{amx}) in addition to catchment-scale precipitation and potential evapotranspiration (P+PE). Models that use static attributes among input variables are shown on the right (b and d), while models that use only dynamic variables are shown on the left (a and c).

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

Authors' response: We thank the Referee for mentioning this typo, which we have corrected in the revised manuscript version.

5 Cited References

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. Hydrol. Earth Syst. Sci. 27, 139–157. https://doi.org/10.5194/hess-27-139-2023

Beaufort, A., Moatar, F., Curie, F., Ducharne, A., Bustillo, V., Thiéry, D., 2016. River Temperature Modelling by Strahler Order at the Regional Scale in the Loire River Basin, France. River Res. Appl. 32, 597–609. https://doi.org/10.1002/rra.2888

Bustillo, V., Moatar, F., Ducharne, A., Thiéry, D., Poirel, A., 2014. A multimodel comparison for assessing water temperatures under changing climate conditions via the equilibrium temperature concept: case study of the Middle Loire River, France. Hydrol. Process. 28, 1507–1524. https://doi.org/10.1002/hyp.9683

Caissie, D., 2006. The thermal regime of rivers: a review. Freshw. Biol. 51, 1389–1406. https://doi.org/10.1111/j.1365-2427.2006.01597.x

Caissie, D., Kurylyk, B.L., St-Hilaire, A., El-Jabi, N., MacQuarrie, K.T.B., 2014. Streambed temperature dynamics and corresponding heat fluxes in small streams experiencing seasonal ice cover. J. Hydrol. 519, 1441–1452. https://doi.org/10.1016/j.jhydrol.2014.09.034

Dugdale, S.J., Hannah, D.M., Malcolm, I.A., 2017. River temperature modelling: A review of process-based approaches and future directions. Earth-Sci. Rev. 175, 97–113. https://doi.org/10.1016/j.earscirev.2017.10.009

Dugdale, S.J., Malcolm, I.A., Hannah, D.M., 2024. Understanding the effects of spatially variable riparian tree planting strategies to target water temperature reductions in rivers. J. Hydrol. 635, 131163. https://doi.org/10.1016/j.jhydrol.2024.131163

Edinger, J.E., Duttweiler, D.W., Geyer, J.C., 1968. The Response of Water Temperatures to Meteorological Conditions. Water Resour. Res. 4, 1137–1143. https://doi.org/10.1029/WR004i005p01137

Feigl, M., Lebiedzinski, K., Herrnegger, M., Schulz, K., 2021. Machine-learning methods for stream water temperature prediction. Hydrol. Earth Syst. Sci. 25, 2951–2977. https://doi.org/10.5194/hess-25-2951-2021

Fidal, J., Kjeldsen, T.R., 2020. Operational comparison of rainfall-runoff models through hypothesis testing. J. Hydrol. Eng. 25, 04020005. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001892

Gallice, A., Schaefli, B., Lehning, M., Parlange, M.B., Huwald, H., 2015. Stream temperature prediction in ungauged basins: review of recent approaches and description of a new physics-derived statistical model. Hydrol. Earth Syst. Sci. 19, 3727–3753. https://doi.org/10.5194/hess-19-3727-2015

Hashemi, R., Brigode, P., Garambois, P.-A., Javelle, P., 2022. How can we benefit from regime information to make more effective use of long short-term memory (LSTM) runoff models? Hydrol. Earth Syst. Sci. 26, 5793–5816. https://doi.org/10.5194/hess-26-5793-2022

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

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

Kraft, B., Schirmer, M., Aeberhard, W.H., Zappa, M., Seneviratne, S.I., Gudmundsson, L., 2025. 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

Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin. Hydrol. Earth Syst. Sci. 28, 4187–4201. https://doi.org/10.5194/hess-28-4187-2024

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G., 2019. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrol. Earth Syst. Sci. 23, 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Kurylyk, B.L., MacQuarrie, K.T.B., Caissie, D., McKenzie, J.M., 2015. Shallow groundwater thermal sensitivity to climate change and land cover disturbances: derivation of analytical expressions and implications for stream temperature modeling. Hydrol. Earth Syst. Sci. 19, 2469–2489. https://doi.org/10.5194/hess-19-2469-2015

Leach, J.A., Kelleher, C., Kurylyk, B.L., Moore, R.D., Neilson, B.T., 2023. A primer on stream temperature processes. WIREs Water 10, e1643. https://doi.org/10.1002/wat2.1643

Ma, Y., Montzka, C., Bayat, B., Kollet, S., 2021. Using Long Short-Term Memory networks to connect water table depth anomalies to precipitation anomalies over Europe. Hydrol. Earth Syst. Sci. 25, 3555–3575. https://doi.org/10.5194/hess-25-3555-2021

Michel, A., Schaefli, B., Wever, N., Zekollari, H., Lehning, M., Huwald, H., 2022. Future water temperature of rivers in Switzerland under climate change investigated with physics-based models. Hydrol. Earth Syst. Sci. 26, 1063–1087. https://doi.org/10.5194/hess-26-1063-2022

Moore, R.D., Sutherland, P., Gomi, T., Dhakal, A., 2005. Thermal regime of a headwater stream within a clear-cut, coastal British Columbia, Canada. Hydrol. Process. 19, 2591–2608. https://doi.org/10.1002/hyp.5733

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T.Y., Weitzner, D., Matias, Y., 2024. Global prediction of extreme floods in ungauged watersheds. Nature 627, 559–563. https://doi.org/10.1038/s41586-024-07145-1

Niemeyer, R.J., Cheng, Y., Mao, Y., Yearsley, J.R., Nijssen, B., 2018. A Thermally Stratified Reservoir Module for Large-Scale Distributed Stream Temperature Models With Application in the Tennessee River Basin. Water Resour. Res. 54, 8103–8119. https://doi.org/10.1029/2018WR022615

Oudin, L., Andréassian, V., Loumagne, C., Michel, C., 2006. How informative is land-cover for the regionalization of the GR4J rainfall-runoff model? Lessons of a downward approach. IAHS Publ. 307, 246–255.

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

Padrón, R.S., Zappa, M., Bernhard, L., Bogner, K., 2025. 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

Picard, C., Floury, M., Seyedhashemi, H., Morel, M., Pella, H., Lamouroux, N., Buisson, L., Moatar, F., Maire, A., 2022. Direct habitat descriptors improve the understanding of the organization of fish and macroinvertebrate communities across a large catchment. PLOS ONE 17, e0274167. https://doi.org/10.1371/journal.pone.0274167

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., Andréassian, V., 2011. A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. J. Hydrol. 411, 66–76. https://doi.org/10.1016/j.jhydrol.2011.09.034

Qiu, R., Wang, Y., Rhoads, B., Wang, D., Qiu, W., Tao, Y., Wu, J., 2021. River water temperature forecasting using a deep learning method. J. Hydrol. 595, 126016. https://doi.org/10.1016/j.jhydrol.2021.126016

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., Shen, C., 2021a. 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

Rahmani, F., Shen, C., Oliver, S., Lawson, K., Appling, A., 2021b. Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. Hydrol. Process. 35, e14400. https://doi.org/10.1002/hyp.14400

Rivière, A., Flipo, N., Goblet, P., Berrhouma, A., 2020. Thermal reactivity at the stream–aquifer interface. Hydrogeol. J. 28, 1735–1753. https://doi.org/10.1007/s10040-020-02154-6

Saadi, M., 2025. Scripts for the paper “Which strategy to improve the performances of an LSTM-based model for extreme stream temperature values?” https://doi.org/10.5281/zenodo.15864784

Saadi, M., Oudin, L., Ribstein, P., 2021. Physically consistent conceptual rainfall–runoff model for urbanized catchments. J. Hydrol. 599, 126394. https://doi.org/10.1016/j.jhydrol.2021.126394

Seyedhashemi, H., Moatar, F., Vidal, J.-P., Thiéry, D., 2023. Past and future discharge and stream temperature at high spatial resolution in a large European basin (Loire basin, France). Earth Syst. Sci. Data 15, 2827–2839. https://doi.org/10.5194/essd-15-2827-2023

Song, Y., Chaemchuen, P., Rahmani, F., Zhi, W., Li, L., Liu, X., Boyer, E., Bindas, T., Lawson, K., Shen, C., 2024. Deep learning insights into suspended sediment concentrations across the conterminous United States: Strengths and limitations. J. Hydrol. 639, 131573. https://doi.org/10.1016/j.jhydrol.2024.131573

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

Toffolon, M., Piccolroaz, S., 2015. A hybrid model for river water temperature as a function of air temperature and discharge. Environ. Res. Lett. 10, 114011. https://doi.org/10.1088/1748-9326/10/11/114011

van Hamel, A., Brunner, M.I., 2024. Trends and Drivers of Water Temperature Extremes in Mountain Rivers. Water Resour. Res. 60, e2024WR037518. https://doi.org/10.1029/2024WR037518

van Vliet, M.T.H., Franssen, W.H.P., Yearsley, J.R., Ludwig, F., Haddeland, I., Lettenmaier, D.P., Kabat, P., 2013. Global river discharge and water temperature under climate change. Glob. Environ. Change 23, 450–464. https://doi.org/10.1016/j.gloenvcha.2012.11.002

Wanders, N., van Vliet, M.T.H., Wada, Y., Bierkens, M.F.P., van Beek, L.P.H. (Rens), 2019. High-Resolution Global Water Temperature Modeling. Water Resour. Res. 55, 2760–2778. https://doi.org/10.1029/2018WR023250

Yu, Q., Jiang, L., Schneider, R., Zheng, Y., Liu, J., 2024. Deciphering the Mechanism of Better Predictions of Regional LSTM Models in Ungauged Basins. Water Resour. Res. 60, e2023WR035876. https://doi.org/10.1029/2023WR035876

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., Li, L., 2021. From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? Environ. Sci. Technol. 55, 2357–2368. https://doi.org/10.1021/acs.est.0c06783

Zwart, J.A., Oliver, S.K., Watkins, W.D., Sadler, J.M., Appling, A.P., Corson-Dosch, H.R., Jia, X., Kumar, V., Read, J.S., 2023. Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. JAWRA J. Am. Water Resour. Assoc. 59, 317–337. https://doi.org/10.1111/1752-1688.13093