https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

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

Mohamed Saadi[1], Louis Guichard[1,*], Gabrielle Cognot[1], Laurent Labbouz[2], Hélène Roux[1]

[1]Université de Toulouse, Toulouse INP, CNRS, Institut de Mécanique des Fluides de Toulouse, 31400 Toulouse, France
[2]Eaucea, 31000 Toulouse, France
[*]Now at Laboratoire Environnement Ville Société (EVS), 69372 Lyon Cedex 08, France

Correspondence to: Mohamed Saadi (mohamed.saadi@toulouse-inp.fr)

Abstract. Deep-learning models have demonstrated strong performances in reproducing stream temperature dynamics, which is promising for the reconstruction of missing stream temperature records at ungauged locations. However, model accuracy for the range of high, summer stream temperature has been usually overlooked, raising the question of the suitability of using deep-learning methods during this crucial season. In this study, we investigated strategies to improve the performances of a stream-temperature model based on LSTM (Long Short-Term Memory) cells over the highest 10% observed values at 21 stations located in the Garonne river catchment. We quantified the gain in model performance thanks to regional multi-catchment training with static attributes, exploiting hydrologically relevant variables, and further penalizing the errors at extreme temperature values using custom loss functions. Our key results are: (1) Regional multi-catchment training is the best strategy to improve the performances of LSTM models not only over the top 10% values but also over the whole range of observations. (2) The gain in performances was mainly brought by the use of static, catchment and reach attributes. (3) Customizing the loss function to emphasize the model errors on extreme temperature values did not lead to significant gains in test performances. This study further confirms the suitability of well-trained LSTM models for extreme stream temperature values, offering significant advantages for water management at data-sparse regions during summer periods.

1 Introduction

1.1 Challenges in stream temperature monitoring

Stream temperature is a key ecological variable that controls the biogeochemical equilibrium of river waters (Ducharne, 2008; Zhi et al., 2023), the spawning, growth, and health of aquatic species (Alfonso et al., 2021; Armstrong et al., 2021; Bowerman et al., 2018), and their spatial distribution along the river network (Bonacina et al., 2023; Buisson et al., 2008; Comte et al., 2013; Daufresne and Boët, 2007; Maire et al., 2019; Picard et al., 2022). Changes in stream temperature have strong impacts on drinking water production (Delpla et al., 2009), fisheries and related recreational activities (Jones et al., 2013), and can lead to disruption in electric energy production by impacting the cooling of power plants (van Vliet et al.,

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

2012). Despite its ecological and socio-economic importance, existing monitoring networks of stream temperature are too recent (see Ouellet et al., 2020) and have a limited spatial coverage, leaving a significant number of reaches ungauged (Arora et al., 2016; Hannah and Garner, 2015). Furthermore, existing large-sample, nation-wide studies such as Beaufort et al. (2022), Hare et al. (2021) or Segura et al. (2015) illustrate the extent to which the characterization of the temporal variability of stream temperature can be limited due to low temporal resolution or high rates of missing values. For example, among the 1700 French stations collected by Beaufort et al. (2022), only 30 stations had year-round observations covering the whole 9-year period of study (2009-2017). This restricted spatiotemporal monitoring not only limits the understanding of stream temperature variability at a variety of locations with contrasting landscape and climatic settings, but also the quantification and attribution of changes in stream temperature due to climate change or anthropogenic perturbations (Essaid and Caldwell, 2017; Kędra and Wiejaczka, 2018; Macedo et al., 2013; Moatar and Gailhard, 2006; Seyedhashemi et al., 2021; van Vliet et al., 2013).

1.2 Modelling approaches to overcome the monitoring gap: A brief overview

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 (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 due to climate or landscape changes. For example, Caissie et al. (2014), Kurylyk et al. (2015) and Rivière et al. (2020) solved the heat transport equation to represent the heat exchanges at the stream-aquifer interface, either to characterize the nature of heat fluxes (advective vs. diffusive) based on temperature measurements or to quantify the sensitivity of the groundwater thermal regime to changes in stream temperature. Dugdale et al. (2024) used a heat balance model to quantify the impact of increasing riparian shading by tree planting on stream temperature at one single reach in Scotland. At the globe scale, Wanders et al. (2019) implemented a solution for the heat balance equation that accounted for heat advection by streamflow in order to reproduce the thermal regime of large rivers worldwide.

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 the land use). Therefore, process-based approaches adopt several simplifying hypotheses, such as combining the physically based heat balance equation with a statistical approach or using the equilibrium temperature concept (Edinger et al., 1968) to parametrize the heat fluxes at the free surface. For example, Toffolon and Piccolroaz (2015) and Gallice et al. (2015) tested models of combined physically based and statistical methods for respectively three and 29 Swiss stations. Larnier et al. (2010) applied the equilibrium temperature concept to reproduce the daily stream temperature dynamics at one station

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

located in the Garonne river catchment (France). Bustillo et al. (2014) tested several variants of this concept for a station at the Loire river catchment (France) by comparing physics-based against regression methods to estimate the equilibrium temperature (in addition to the heat flux coefficient). Beaufort et al. (2016) extended this approach to the regional scale (the whole Loire catchment, ~10^5 km^2) by coupling a hydrological model with a thermal model, in which they accounted for the geomorphological properties of the river reach, the network topology (Strahler order) and the riparian vegetation. A similar model structure was applied by Seyedhashemi et al. (2023) to simulate past and future thermal and hydrological regimes at the scale of the Loire catchment.

Alternative to the process-based approach, the implicit, fully data-driven approach aims at capturing the correlations between stream temperature and hydro-climatic variables using machine learning techniques (see the reviews in Gallice et al. (2015), Benyahya et al. (2007), and Souaissi et al. (2023)). These techniques have less requirements in terms of data variety than process-based models and once calibrated/trained, their computational cost in forward mode is more affordable. Very basic methods apply linear and logistic regression to estimate stream temperature from air temperature (Ducharne, 2008; Jackson et al., 2018; Mohseni et al., 1998; Stefan and Preud'homme, 1993). In several studies, a pairing of stream temperature and air temperature, by looking at the ratio of the characteristic amplitudes of the two signals or the lag between their phases, can help quantify the importance of groundwater contributions (Briggs et al., 2018; Hare et al., 2021; Johnson et al., 2020) or characterize the effects of dams and ponds (Seyedhashemi et al., 2021). More advanced applications account for hydrological and spatial controls on stream temperature in addition to atmospheric variables (Feigl et al., 2021; Isaak et al., 2014) and/or apply advanced machine-learning and deep-learning techniques to efficiently digest larger and more diverse data (Zhi et al., 2024). 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; Kratzert et al., 2018; Ma et al., 2021; Nearing et al., 2024; Song et al., 2024; Zhi et al., 2021). Sadler et al. (2022) showed the ability of LSTM to predict both streamflow and stream temperature for 101 stations in the United States. Rahmani et al. (2021b) illustrated the excellent performances of LSTM models in reproducing stream temperature for 118 United States catchments, with a median RMSE (root-mean-square errors) at 0.69°C. In particular, they showed an interesting application of a two-stage LSTM model, where a hydrological LSTM-based model provides streamflow simulations to an LSTM-based thermal model. Rahmani et al. (2021a) showed that regionalized LSTM networks are very suitable for transferring information on stream temperature dynamics from monitored to unmonitored locations. Other studies demonstrated successful applications of LSTM networks for more operational stream temperature forecasting (e.g., Qiu et al., 2021; Zwart et al., 2023).

1.3 Research questions and scope of the paper

Our literature overview reveals that LSTM networks offer a very promising avenue for catchment-scale reconstruction of stream temperature for several reasons: (1) They provide state-of-the-art performances at several locations (Rahmani et al., 2021a, b); (2) Their performances improve when trained at several locations at once thanks to their flexibility, which is

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

suitable for regionalization (see Kratzert et al., 2024); (3) They can be physically constrained (Hoedt et al., 2021) and their structure can be physically interpretable (Lees et al., 2022). However, evaluation of LSTM performances for extreme values is limited to rainfall-runoff models (Frame et al., 2022; Nearing et al., 2024), and previous studies in the context of stream temperature modelling did not specifically address LSTM performances for extreme temperature values (Rahmani et al., 2021a, b). Extreme stream temperature values usually occur during the summer season, a highly critical period from a water-management standpoint, as it is concomitant with lower streamflow and higher air temperature values. Accurate predictions at this range are crucial to acceptably anticipate when the temperature values will exceed the legal/recommended threshold for drinking water production (25°C according to the Council Directive 80/778/EEC; https://eur-lex.europa.eu/eli/dir/1980/778/oj/eng, last access: 5 April 2025) or the lethal/sub-lethal thresholds for freshwater species (see Jackson et al., 2018).

Our objective is thus to address this research gap by exploring the performances of a daily LSTM-based model for stream temperature at 21 stations in the Garonne river catchment (~10^4 km^2). This catchment faces several water management challenges especially during the summer season due to high anthropogenic pressure (agriculture, hydro-electricity, and urbanization). Paradoxically, very few studies have addressed stream temperature modelling at this catchment (Beaufort et al., 2022; Larnier et al., 2010). In our implementation, we focused on strategies to improve LSTM performances for high daily stream temperature values (top 10%). Specifically, we compared 18 loss functions, local vs. regional/multi-catchment training, and several combinations of static and dynamic input variables. We aimed at answering the following scientific questions:

- For the aim of reproducing high stream temperature values, what can be gained from increasing the weight of extreme stream temperature values in the loss function used for training?
- What is the added value of combining regional training with static catchment and reach attributes in improving the performances of LSTM for high stream temperature values?

The remaining of this manuscript is organized as follows. Section 2 will present the catchment set, the collected stream temperature dataset, and the dynamic and static variables used to construct the LSTM models. Section 3 will detail the methodological setup and the tested strategies to improve LSTM performances for high stream temperature values. Sections 4 and 5 will show and discuss the results, and Section 6 will summarize the main findings and propose some future research directions.

2 Catchment set

We used pre-processed daily stream temperature (T_w) time series provided by the Haute-Garonne Departmental Council (CD31) at 37 stations located within the Garonne catchment upstream of Toulouse, France (Fig. 1). 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

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

days) of datapoints for model testing (see Sect. 3). We call these 21 stations "test stations". Raw T_w time series span years within the period 1991-2021 and were provided by different sources:

- Data for 12 stations were extracted from the Naiades database (https://naiades.eaufrance.fr/acces-donnees#/temperature, last access: 5 April 2025), which is managed at the national scale by the French Office for Biodiversity (*Office Français de la Biodiversité*). Among these stations, 10/12 are test stations;
- Five stations are part of the alert network (*Réseau des Stations d'Alerte*, RSA), which serves to monitor the water quality at specific locations close to withdrawal points for drinking water production. Among these stations, 3/5 are test stations;
- Data for 20 stations were provided by two departmental federations for fishing: FPeche31 (*Fédération de Pêche de la Haute-Garonne*), FPeche65 (*Fédération de Pêche des Hautes-Pyrénées*), and one additional fishing association: MIGADO (*Migrateurs Garonne Dordogne Charente Seudre*). Among these stations, 8/20 stations are test stations.

This diversity of data providers poses a challenge from a standpoint of data quality and data uncertainty. For example, some of the 21 test stations are located very close to each other but their T_w time series come from different sources. This is the case of three pairs of stations (see Fig. 1 for their location): (1) two stations that are located on the Garonne river at Bazacle (Toulouse) and their T_w data were provided by RSA and MIGADO; (2) two stations that are located on the Garonne river at Valentine (Naiades and MIGADO); and (3) two stations that are located on the Pique river at Cier-de-Luchon (Naiades and MIGADO). Since these stations are still away from each other and their time series show some differences (due to sensor location and exposure), we decided to keep them all since there is no additional information that could help decide which stations to discard.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

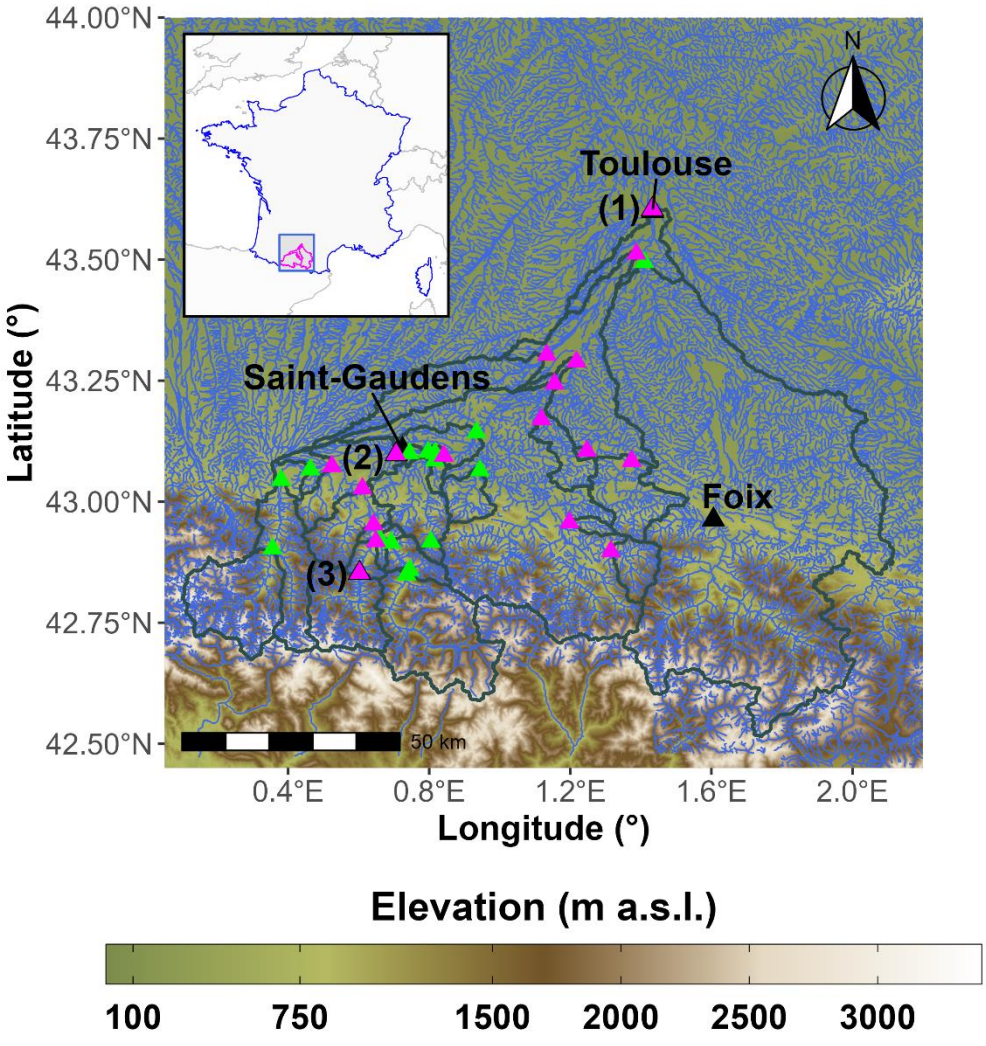

Figure 1: Location of the Garonne catchment and the 37 stream temperature stations used in this study. Test stations (i.e., with more than 2434 daily observations of stream temperature) are shown with magenta triangles, whereas stations used only for the training of regional models are shown in green. Elevation data are extracted from the SRTM GL3 product (90-m resolution). River network is extracted from BD CarTHAgE®. Numbers (1), (2), and (3) indicate the location of the river reaches with a pair of stations close to each other but with stream temperature data from two different sources.

For each T_w station, we extracted hourly air temperature values from the SAFRAN dataset (SAFRAN stands for *Système d'Analyse Fournissant des Renseignements Adaptés à la Neige*; Vidal et al., 2010). This dataset was provided by Météo-France at a resolution of 8 km. From the hourly time series, we calculated daily mean air temperature (T_a) and maximum and minimum air temperature for each day (T_{amx} and T_{amn}).

In addition to station-scale atmospheric forcing, we also collected catchment-scale hydrologically relevant variables to test their added values in improving T_w predictions, especially for the case of summer, high temperatures. A first set of hydrologically relevant variables is composed of catchment-scale precipitation (P) and potential evapotranspiration (PE)

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

time series at the daily time step. First, we delimited the topographic catchment drained by each T_w station using the PyFlwDir library (version 0.5.8; Eilander et al., 2021) with the 30-m SRTM dataset for elevation values (SRTM stands for Shuttle Radar Topography Mission; Farr et al., 2007). We did not choose the national product BD ALTI® because it does not cover the very upstream part of the Garonne catchment located in Spain. Second, we used the catchment polygon to extract spatial averages of precipitation values from the hourly, 1-km COMEPHORE dataset (COMEPHORE stands for *COmbinaison en vue de la Meilleure Estimation de la Précipitation HOraiRE*; Tabary et al., 2012), then we summed the hourly values for each day to obtain the daily time series. Third, we used the catchment polygon to extract spatial averages of air temperature values from the SAFRAN dataset, then we applied a temperature-based formula to compute daily potential evapotranspiration from the daily averages of air temperature (Oudin et al., 2005).

The second set of hydrologically relevant variables is streamflow. We first extracted 18 available streamflow time series within the region of interest from the HydroPortail database (https://hydro.eaufrance.fr/, last access: 14 July 2025; Audouy et al., 2024). However, we found that the set of the 18 streamflow stations did not coincide with the set of T_w stations. To circumvent this problem, we used the daily hydrological model GR6J (Pushpalatha et al., 2011) to simulate streamflow at each T_w station, due to its parsimony and its suitability for low-flow values. We proceeded as follows:

1. For each T_w station, we looked for the closest streamflow station that is located either upstream or downstream following the river reach;
2. We delimited the catchment drained by the streamflow station using the PyFlwDir Python library (Eilander et al., 2021) applied to the 30-m SRTM dataset (Farr et al., 2007);
3. We computed catchment-scale average precipitation and potential evapotranspiration for each streamflow station using the catchment polygon;
4. We calibrated the GR6J hydrological model to reproduce the streamflow at the streamflow station using the corresponding precipitation P_{BV_Q} and potential evapotranspiration PE_{BV_Q}. In other words, we looked for the optimal parameter set $\hat{\theta}$ such as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \mathrm{NSE}\left[\log\left(Q_{sim,BV_Q}(\theta)\right), \log\left(Q_{obs,BV_Q}\right)\right] \quad (1)$$

where θ is the 6-parameter set of GR6J (Pushpalatha et al., 2011), NSE is the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970), $Q_{sim,BV_Q}(\theta) = \mathrm{GR6J}\left(\theta; P_{BV_Q}, PE_{BV_Q}\right)$ is the simulated streamflow using the input precipitation P_{BV_Q} and potential evapotranspiration PE_{BV_Q} for the catchment drained by the streamflow station, and Q_{obs,BV_Q} is the observed streamflow. This calibration has been done using the airGR R library (Coron et al., 2017, 2023). Obtained NSE values on log-transformed streamflow time series ranged from -0.36 to 0.91 (median at 0.77), with only one station having a score below 0.44. This station provides simulated streamflow for a T_w station that is not part of the 21 test stations.
5. We used the optimal set $\hat{\theta}$ to simulate the streamflow at the T_w station:

$$Q_{sim} = \mathrm{GR6J}(\hat{\theta}; P, PE) \quad (2)$$

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

where P and PE are the daily time series of catchment-scale precipitation and potential evapotranspiration computed for the catchment drained by the T_w station.

Finally, we computed a set of static variables that we used to construct regionally trained LSTM models for T_w (Table 1). These static variables were chosen to reflect catchment properties in terms of climate (aridity index, i.e., ratio of average potential evapotranspiration to average precipitation for the catchment), morphology (station elevation, median catchment elevation, catchment area), land use (percent of urban areas, forest cover, and agricultural areas), and soil (percent of sand, clay, and silt). This relatively small dataset covers a rather rich gradient of climatic and landscape characteristics, with cold thermal regimes and steep orographic gradients upstream vs. warm and mild slopes downstream (see Table 1 and Fig. 1). Note that some of these stations are located close to hydroelectric power plants or downstream of a retention pond, underlining a rich mixture of natural and anthropogenically impacted thermal regimes.

Table 1: Summary of dynamic and static variables and their distributions across the catchment set.

Name (unit)	Static/dynamic	Scale	Data source	Min	Median	Max
T_w: Daily stream temperature (°C)	Dynamic	Station	Naiades, RSA, fishing federations	9[a]	11.3[a]	14.2[a]
T_a: Daily air temperature (°C)	Dynamic	Station	SAFRAN (Vidal et al., 2010)	8.4[b]	11.8[b]	14.2[b]
T_{amx}: Daily maximum air temperature (°C)	Dynamic	Station	SAFRAN (Vidal et al., 2010)	12.8[b]	16.8[b]	18.2[b]
T_{amn}: Daily minimum air temperature (°C)	Dynamic	Station	SAFRAN (Vidal et al., 2010)	4.8[b]	7.5[b]	10.2[b]
P: Daily precipitation (mm)	Dynamic	Catchment	COMEPHORE (Tabary et al., 2012)	793[b]	1130[b]	1280[b]
PE: Daily potential evapotranspiration[c] (mm)	Dynamic	Catchment	SAFRAN[c] (Vidal et al., 2010)	491[b]	595[b]	778[b]
Q_{sim}: Daily simulated streamflow (mm)	Dynamic	Station/reach	This study	64[b]	670[b]	988[b]
Aridity index[d] (-)	Static	Catchment	SAFRAN[c] (Vidal et al., 2010) and COMEPHORE (Tabary et al., 2012)	0.41	0.52	0.98
Station elevation (m a.s.l.)	Static	Station/reach	BD ALTI® (25 m ; https://geoservices.ign.fr/bdalti, last access: 14 July 2025)	133	346	788
Median catchment elevation[e] (m a.s.l.)	Static	Catchment	BD ALTI® (25 m; https://geoservices.ign.fr/bdalti, last access: 14 July 2025)	338[e]	1350[e]	1970[e]
Catchment area (km^2)	Static	Catchment	This study	8	737	10^4
Percentage of urban areas[f] (%)	Static	Catchment	CORINE Land cover (100 m; https://land.copernicus.eu/en/products/corine-land-cover/, last access: 14 July 2025)	0	1	3
Percentage of forest cover[g] (%)	Static	Catchment	CORINE Land cover (100 m; https://land.copernicus.eu/en/products/corine-land-cover/, last access: 14 July 2025)	15	43	76
Percentage of agricultural areas[h] (%)	Static	Catchment	CORINE Land cover (100 m; https://land.copernicus.eu/en/products/corine-land-cover/, last access: 14 July 2025)	0	14	83
Sand content[i] (%)	Static	Catchment	Harmonized World Soil Database (version 2.0; FAO and IIASA,	15	37	51

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

			2023)			
Silt content[i] (%)	Static	Catchment	Harmonized World Soil Database (version 2.0; FAO and IIASA, 2023)	16	26	38
Clay content[i](%)	Static	Catchment	Harmonized World Soil Database (version 2.0; FAO and IIASA, 2023)	8	14	22

[a]We excluded the stations with short time series by computing these statistics using the 21 test stations only.
[b]These values represent long-term averages (1997-2022) in °C for temperature and in mm/year for precipitation/potential evapotranspiration.
[c]We estimated potential evapotranspiration using a temperature-based formula (Oudin et al., 2005). Air temperature was extracted from SAFRAN (resolution of 8 km).
[d]Aridity index was computed as the ratio of long-term catchment-average potential evapotranspiration to long-term catchment-average precipitation over the period 1997-2022.
[e]We got similar statistics using the more complete SRTM GL1 dataset (30-m resolution): 341 m, 1360 m, and 1970 m for min, median, and max values.
[f]We took the average over the years 1990 to 2018 of the proportion of the catchment occupied by the following CORINE land-cover classes: classes belonging to 1.1 (Urban fabric), 1.2.1 (Industrial or commercial units), and 1.2.2 (Road and rail networks and associated land).
[g]We took the average over the years 1990 to 2018 of the proportion of the catchment occupied by the CORINE land-cover classes under 3.1 (Forest).
[h]We took the average over the years 1990 to 2018 of the proportion of the catchment occupied by the CORINE land-cover classes under 2 (Agricultural areas).
[i]We took the average over the seven layers (0-20 cm, 20-40 cm, 40-60 cm, 60-80 cm, 80-100 cm, 100-150 cm, and 150-200 cm) using the hwsdr R package (Hufkens and Labs, 2023).

3 Tested strategies to improve the performances of LSTM models over extreme stream temperature values

3.1 Model structure

Our choice of LSTM networks to implement a T_w model is motivated by four main reasons: (1) They demonstrated state-of-the-art performances across several catchments (Rahmani et al., 2021b); (2) They allow for incorporating static in addition to dynamic variables, which is suitable for regionalization (Kratzert et al., 2019; Rahmani et al., 2021a); (3) They are strong spatial extrapolators by performing better in a regionalization mode (Kratzert et al., 2024), which opens the way for reconstructing the missing records at ungauged locations; and (4) They are more flexible than process-based approaches in handling measurement uncertainties and accounting for anthropogenic influences by inferring the influenced T_w behaviour directly from the measurements.

We conducted a preliminary study in which we tested several hyperparameters to see their impact on the LSTM performances (not shown here). We concluded the following:

- The LSTM performances were not significantly sensitive to the number of LSTM layers or the number of LSTM cells per layer. For this reason, we choose to test an LSTM model with only one layer composed of 128 LSTM cells, with a dropout rate at 0.4 to limit the risk of overfitting (Srivastava et al., 2014), and connected to the output layer (1 neuron) via a dense layer of 128 neurons activated using a linear function.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

- The LSTM performances were sensitive to the lookback, which was expected due to the strong differences in T_w dynamics across our catchment set. For this reason, we chose to test three lookbacks (30, 90, 365) to accommodate for T_w dynamics at the monthly, seasonal, and annual scales. Our strategy to choose the final best lookback is explained in Sect. 3.3.
- The LSTM performances were not sensitive to the batch size, which we fixed at 64. In addition, we chose Adam (Kingma and Ba, 2017) as an optimizer with a fixed learning rate at 10^{-4}. The maximum number of epochs was fixed at 10000. However, we implemented an early stopper that halted the training of the LSTM after 100 epochs with no decrease in MSE (mean squared errors) on the validation set by at least 0.001°C^2.

A detailed description of the LSTM equations can be found in Hashemi et al. (2022) or Kratzert et al. (2018). Our LSTM models were implemented and trained using classes and functions from the PyTorch library (Paszke et al., 2017) in a Python development environment (version 3.8).

3.2 Strategies to improve the LSTM performances for extreme stream temperature values

In this part, we summarize the strategies that we tested to look for the best approach to improve the LSTM performances at high T_w values. These strategies include regional multi-catchment training, the inclusion of hydrologically relevant variables, and an adaptation of the loss function to increase the weight of high values in the training phase.

3.2.1 Option 1: Regional training

Due to its flexibility, an LSTM-based model has the ability to learn from a large space of thermal behaviours when trained in a regionalization mode, which may result in better performances for the range of extreme values (Kratzert et al., 2024). Moreover, it offers the possibility of using static, catchment attributes in addition to dynamic variables (see Sect. 3.2.2). 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 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.
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 and compare their performances against the local models.

Note that this methodological setup does not fully examine the LSTM capabilities in spatial extrapolation as done by Rahmani et al. (2021a). To achieve this would have required either holding a subset as pseudo-ungauged for testing, or adopting a leave-one-out approach (Jackson et al., 2018), and these two options are limited by the size of our catchment set

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

and the large number of tests that we already performed. However, our methodological setup will still indicate the extent to which the multi-catchment training helps the regionally trained LSTM models in learning from all the catchment behaviours at once and performing better than the locally trained ones, which is necessary (although not sufficient) for satisfactory spatial extrapolation performances.

3.2.2 Option 2: Choice of the input variables

The most basic T_w models make use of the strong correlation between T_w and air temperature (T_a). Here, we quantify the added value of using catchment-scale hydrological variables especially for the simulation of high T_w values. In addition, we examine the added value of catchment and reach attributes using the regionally trained LSTM models. In summary:

1. We tested four sets of inputs for the local models (Table 2). Comparing models that use T_a against those using $T_{amn} + T_{amx}$ as inputs will show whether there is any added value in using daily extreme air temperature values instead of the daily averages. The added value of hydrological variables will be evaluated by comparing models that use $T_{amn} + T_{amx} + P + PE$ and $T_{amn} + T_{amx} + Q_{sim}$ against those using $T_{amn} + T_{amx}$. 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 ($P + PE$) instead of the more relevant station-scale streamflow (Q_{sim}).
2. We tested seven sets of input variables for the regional models (Table 2). Four of the seven sets are the same sets of input variables that we used for the local models in order to measure the added value of regional, multi-catchment training. In addition, we tested three more sets that use catchment and reach attributes in addition to dynamic inputs (see Table 2). Comparing these sets with the ones using only the dynamic inputs will help us quantify the gain in accuracy brought by the static attributes.

Note that each static attribute is provided to the LSTM model by simply repeating the same temporally fixed value over the whole period. For both sets of local and regional models, all the input data were standardized using moments (mean, standard deviation) computed on the training set.

Table 2: Summary of tested local and regional models. "CatAttrs" refers to static, catchment and reach attributes (see Table 1).

Training scale	Input variables	Models
Local (single catchment)	T_a	$T_w = \text{LSTM}(T_a)$
	$T_{amn} + T_{amx}$	$T_w = \text{LSTM}(T_{amn}, T_{amx})$
	$T_{amn} + T_{amx} + P + PE$	$T_w = \text{LSTM}(T_{amn}, T_{amx}, P, PE)$
	$T_{amn} + T_{amx} + Q_{sim}$	$T_w = \text{LSTM}(T_{amn}, T_{amx}, Q_{sim})$
Regional (multi-catchment)	T_a	$T_w = \text{LSTM}(T_a)$
	$T_{amn} + T_{amx}$	$T_w = \text{LSTM}(T_{amn}, T_{amx})$
	$T_{amn} + T_{amx} + P + PE$	$T_w = \text{LSTM}(T_{amn}, T_{amx}, P, PE)$
	$T_{amn} + T_{amx} + Q_{sim}$	$T_w = \text{LSTM}(T_{amn}, T_{amx}, Q_{sim})$
	$T_{amn} + T_{amx} + CatAttrs$	$T_w = \text{LSTM}(T_{amn}, T_{amx}, CatAttrs)$
	$T_{amn} + T_{amx} + P + PE + CatAttrs$	$T_w = \text{LSTM}(T_{amn}, T_{amx}, P, PE, CatAttrs)$
	$T_{amn} + T_{amx} + Q_{sim} + CatAttrs$	$T_w = \text{LSTM}(T_{amn}, T_{amx}, Q_{sim}, CatAttrs)$

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

3.2.3 Option 3: Adaptation of the loss function

To improve the performances of the LSTM model for a specific range of the target variable, a potential strategy consists in emphasizing the model errors on the range of interest by increasing their weight in the loss function. This is a common strategy for example in the calibration of hydrological models (Garcia et al., 2017; Pushpalatha et al., 2012; Saadi and Furusho-Percot, 2024). We thus tested 18 loss functions that share the following general expression:

$$\mathcal{L}(T_{w,sim}, T_{w,obs}) = \frac{\sum_t \left| g(T_{w,obs,t})^{\mu} - g(T_{w,sim,t})^{\mu} \right|^{\lambda}}{\sum_t \left| g(T_{w,obs,t})^{\mu} - \overline{g(T_{w,obs})^{\mu}} \right|^{\lambda}} \quad (3)$$

where $T_{w,sim,t}, T_{w,obs,t}$ are the simulated and the T_w observed values at the time step t, $g(T_{w,sim,t}), g(T_{w,obs,t})$ their transformed values using the function g, and μ, λ are hyperparameters. The denominator in the loss function of Eq. (3) 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. We tested all the possible combinations from the following configurations:

- Three values for μ: 1, 3, and 5;
- Three values for λ: $\lambda = 2$, which corresponds to using MSE as a metric; $\lambda = 1$ which corresponds to MAE (mean absolute errors); and $\lambda = 4$, which we refer to by M4E (mean errors to the power 4);
- Two functions for g: $g(T) = T$ (no standardized target) and $g(T) = \frac{T-m}{\sigma}$ (standardized target), where m and σ correspond to the mean and the standard deviation of the observed T_w values computed on the training set.

Emphasizing the weight of high T_w values in the loss function can be achieved by increasing the values of λ, μ, and avoiding standardization ($g(T) = T$). Note that the commonly chosen loss function (hereafter, the "reference" loss function) corresponds to $\mu = 1$, $\lambda = 2$ (MSE) and $g(T) = \frac{T-m}{\sigma}$.

3.3 Evaluation framework

We repeated each training configuration (local vs. regional training, input variables, loss functions) three times corresponding to each of the three lookback values (30, 90, and 365). We kept only the lookback value that provided the lowest MSE on the validation set. Thus, we will evaluate a total of 4 × 18 = 72 local models and 7 × 18 = 126 regional models on each of the 21 test stations. To this aim, we computed the mean absolute error MAE (in °C) between the predictions of each model and the observed values during the test period. We computed the MAE score first on the whole test period, then on a restricted period during which observed T_w exceeded the 90th percentile, which corresponds to the 10% highest T_w values observed during the test period.

To evaluate the statistical significance of the differences between two options, say A and B, we applied a one-sided binomial test according to which the null hypothesis is "the proportion of cases where the option B outperforms the option A is lower than or equal to 50%" (Redfern et al., 2016; Saadi et al., 2021). Here, "outperforms" means scoring lower MAE

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

values on the test period. We rejected the null hypothesis when option B scores better performances than A for a sufficiently high number of cases that is defined by the threshold of significance. For example, if the total number of cases is 21 (as 21 stations), we rejected the null hypothesis at the significance level of 5% (resp. of 1%) when option B outperformed option A in more than 15/21 cases (resp. 18/21 cases), in which case we concluded that the use of option B brought statistically significant improvements compared to option A.

4 Results

In this section, we first show the added value of changing the loss function (Sect. 4.1). Then, we show the performances of the local models (Sect. 4.2). Finally, we show the performances of the regionally trained models and the added value of using static, catchment and reach attributes (Sect. 4.3). For information, statistics for the wall-clock time needed to train the LSTM models are given in Table A1 (see Appendix A).

4.1 Effect of the loss function

Unexpectedly, increasing the weight of high T_w values in the loss function deteriorated the performances not only for the reproduction of extreme T_w values (top 10%), but also for the reproduction of the whole range of T_w observations (Fig. 2). 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.
- 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 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 standardized target).

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

Figure 2: Median test performances 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%.**

4.2 Performances of the local models

Using catchment-scale atmospheric forcing (precipitation and potential evapotranspiration) in addition to station-scale atmospheric forcing led to the best results with the local models (Fig. 3). More specifically:

- Over the whole period, the best median performances were obtained with $T_{amn} + T_{amx} + P + PE$ as input variables (median MAE at 0.65°C, MAE without standardization as a training loss function). Over the top 10% range, the best median performances were also obtained with $T_{amn} + T_{amx} + P + PE$ as input variables (median MAE at 0.95°C, MAE with standardization as a loss function).
- The use of the daily min/max air temperature ($T_{amn} + T_{amx}$) instead of daily average air temperature (T_a) did not result in significantly better performances according to the binomial test.
- The use of the hydrological variables ($P + PE$ or Q_{sim}) always led to better results than using only station-scale air temperature ($T_{amn} + T_{amx}$), especially over the top 10% range. However, these improvements were not always statistically significant.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

- The use of simulated streamflow (Q_{sim}) instead of catchment-scale atmospheric forcing ($P + PE$) led to comparable performances both over the whole range and over the top 10% range, suggesting the similarity of catchment-scale forcing or station-scale simulated streamflow in terms of the information content extracted by the LSTM models.

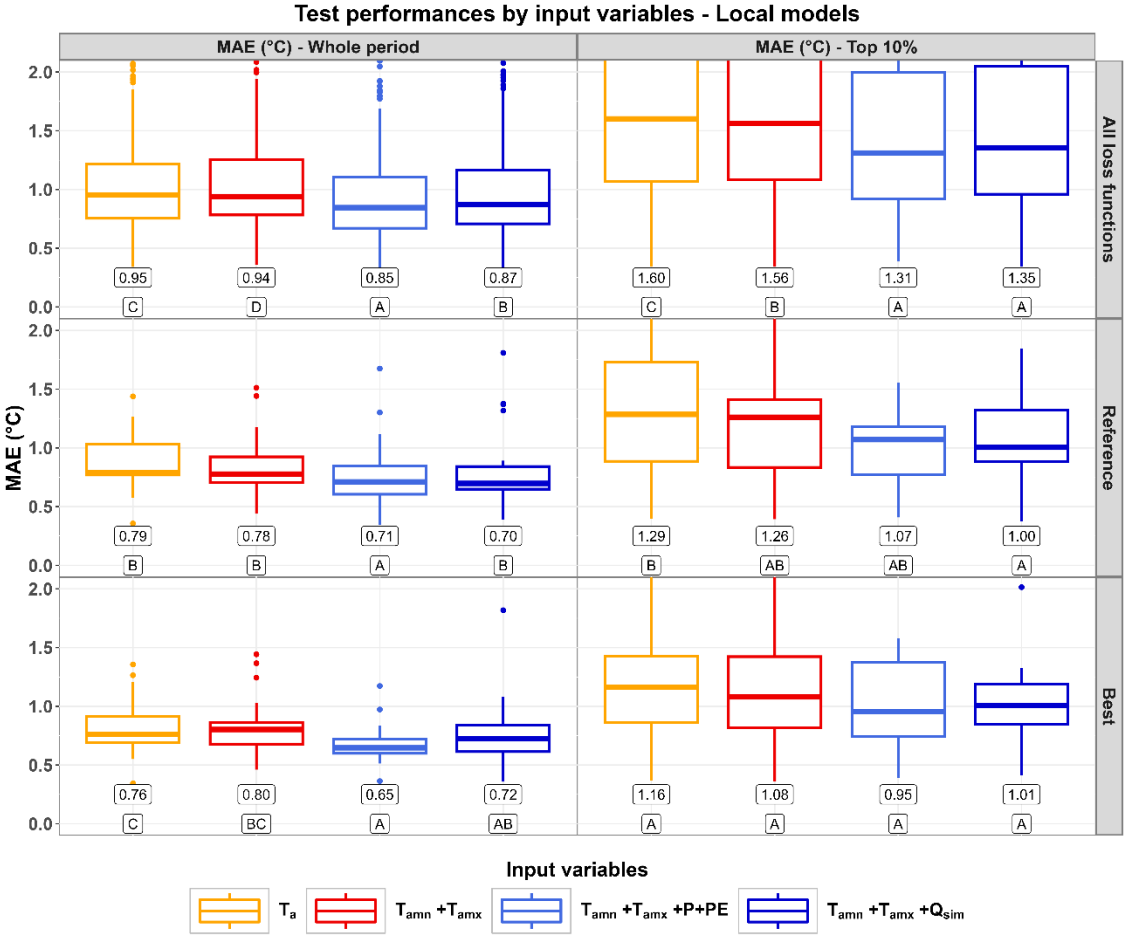

Figure 3: Distributions of the test performances (MAE, in °C) of the local models over the whole test period (left column) and the period corresponding to the highest 10% observed values (right column). The top row shows the distribution across all the loss functions (18 × 21 = 378 points per distribution). 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 (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.

4.3 Performances of the regional models

The use of static, catchment and reach attributes led to the best performances over both the whole test period and on the top 10% range (Fig. 4). More specifically:

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

- Over the test whole period, the best performances were obtained by exploiting catchment-scale atmospheric forcing ($P + PE$) in addition to static attributes (median MAE at 0.56-0.57°C). Over the top 10% range, the best performances (median MAE at 0.73-0.74°C) were obtained by the models that exploited either the catchment-scale atmospheric forcing ($P + PE$) or the simulated streamflow (Q_{sim}) depending on the loss function used for training.
- 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).
- The improvements brought by using the static attributes were almost always statistically significant according to the binomial test. Exceptions are limited to some performance distributions over the top 10% range (namely, models taking as inputs $T_{amn} + T_{amx}$ with vs. without static attributes with the reference loss function, and models taking as inputs $T_{amn} + T_{amx} + P + PE$ with vs. without static attributes with the best loss function; see Fig. 4).

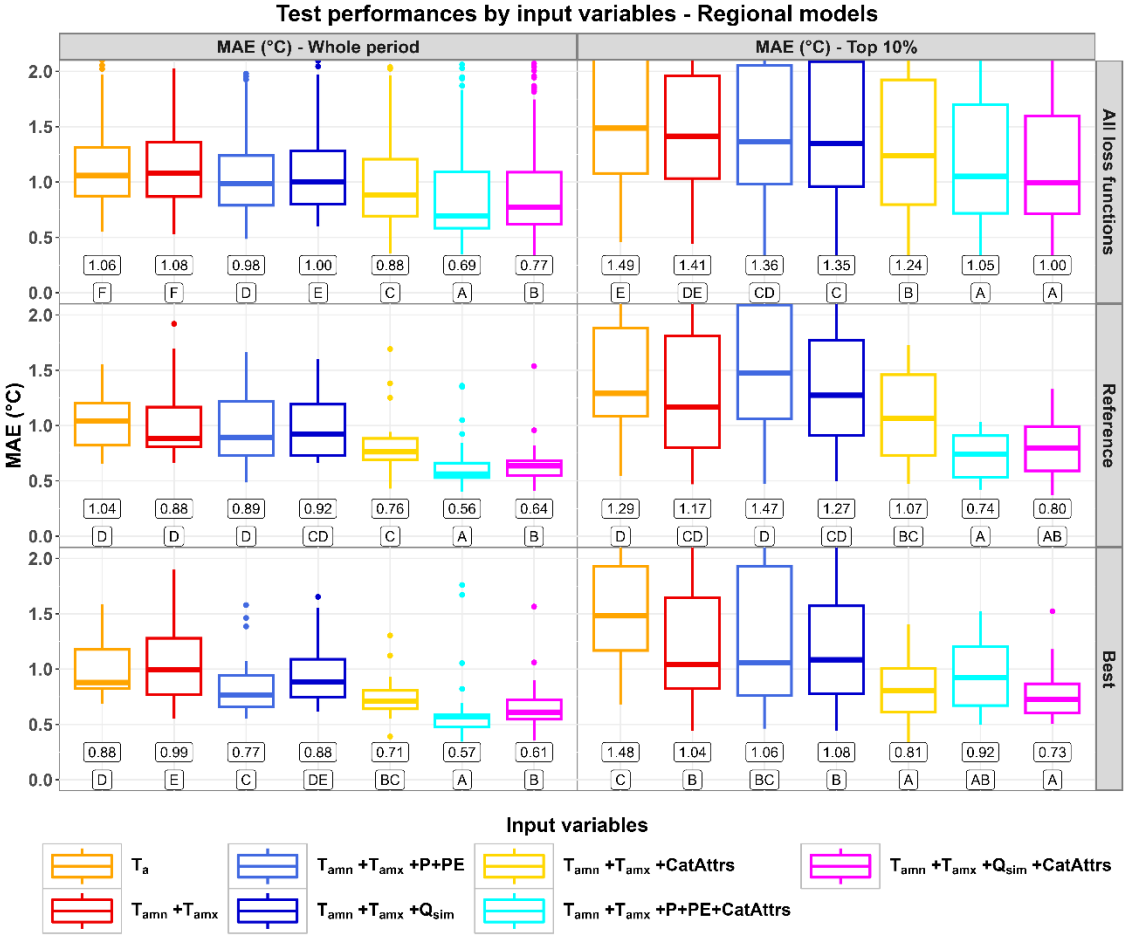

Figure 4: Distributions of the test performances (MAE, in °C) of the regionally trained models over the whole test period (left column) and the period corresponding to the highest 10% observed values (right column). Top row shows the distribution across

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

all the loss functions (18 × 21 = 378 points per distribution). The middle row shows the performances across the reference loss function (MSE with standardization, 21 points per distribution). The bottom row shows the performances with the best loss function (MAE with standardization for the whole period and MSE without standardization for the top 10%, 21 points per distribution). Numerical values under the boxes represent the median value for each distribution. Letters under the numerical values rank the distributions, and are defined such as distributions that share at least one letter are not significantly different according to the binomial test.

To illustrate the improvements brought by the static attributes for regionally trained models, an example of model simulations is shown in Fig. 5 for the Garonne at Valentine (T_w data provided by MIGADO). This station is located at an altitude of 360 m, and drains a catchment area of 2260 km^2 and the median altitude of the catchment is at 1390 m. The T_w observations for this station range between 2°C and 21°C over the whole period of measurements (2007-2021), which a 90% quantile at 17.8°C. Regionally trained models with only dynamic variables as inputs produced satisfactory results over the whole period (MAE at 0.74-0.86°C, Fig. 5a) with a slight improvement thanks to the catchment-scale climatic inputs, but a significant degradation of model performances can be noticed over the top 10% (MAE at 1.18-1.34°C, Fig. 5c). The use of static attributes improved the performances not only over the whole period (MAE down to 0.48°C, Fig. 5b), but also over the highest 10% values (MAE down to 0.48°C, Fig. 5d).

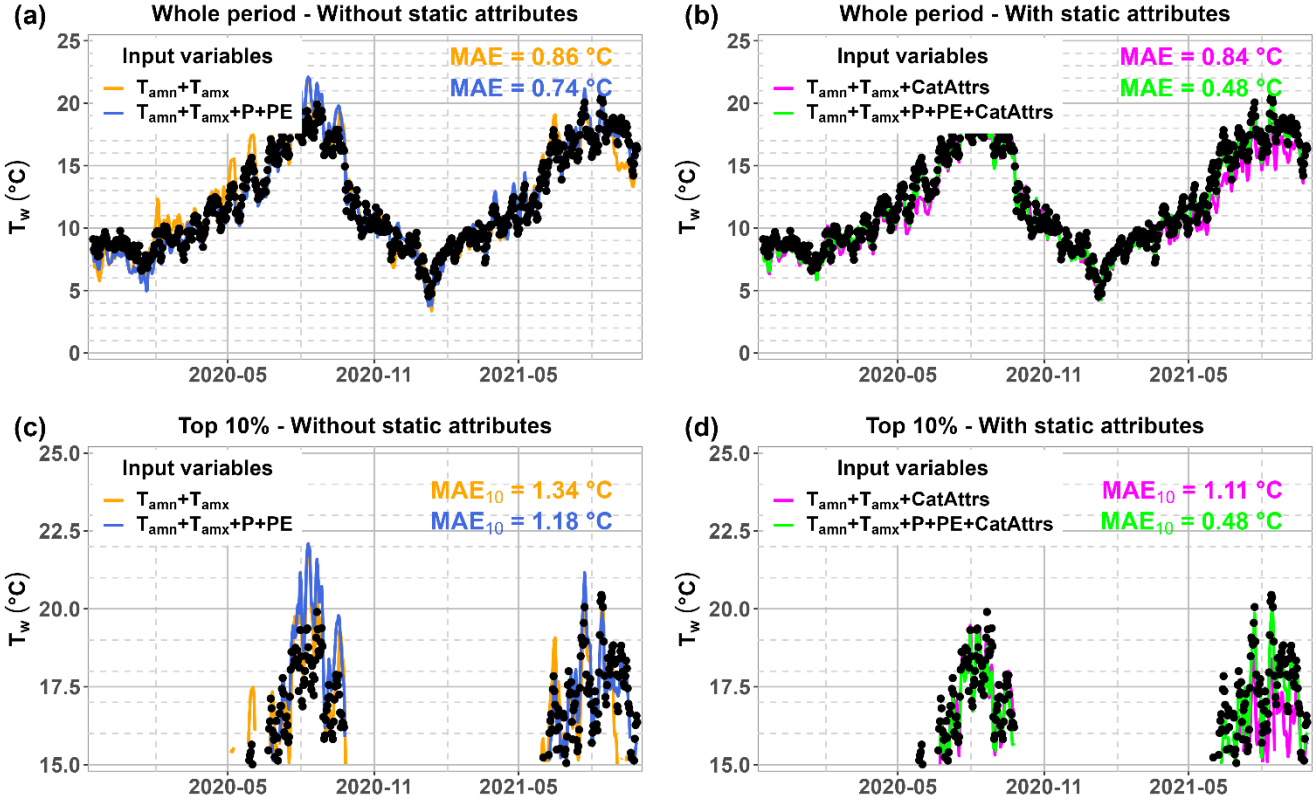

Figure 5: Example of model simulations using LSTM models against observations at the station of the Garonne at Valentine (MIGADO). All models are regionally trained. The two top figures (a and b) show the model simulations and performances (MAE, in °C) over the whole test period, and the two bottom figures (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

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

catchment-scale precipitation and potential evapotranspiration (P+PE). Models that use static attributes among input variables are shown on the right (a and c), while models that use only dynamic variables are shown on the left (b and d).

5 Discussion

To our knowledge, very few studies focused on model performances over periods with extreme or high T_w values (a very close example can be found in Jackson et al., 2018). Our results show that exploiting the static, catchment and reach attributes in addition to dynamic inputs is the best strategy to improve the LSTM performances over the highest 10% values as well as for the whole range of T_w observations (Fig. 4). For example, with the traditional loss function (MSE with standardization), this strategy brought the median MAE performances over the highest 10% range from 1.29°C with only air temperature (T_a, Fig. 3) as input down to 0.74°C with the additional use of catchment-scale forcing and static attributes (Fig. 4). These performances are still slightly worse than the whole-period performances by about 0.2°C on average, highlighting the challenge of simulating extreme values with the LSTM models. The better performances reached with using regionally trained LSTM models confirm their strong capability in learning from a diversity of behaviours (Kratzert et al., 2024) but only if static attributes are provided (see the differences between Fig. 3 and Fig. 4), highlighting the importance of catchment and reach specificities in controlling the thermal regimes (Jackson et al., 2018; Wade et al., 2023).

The obtained performances over the whole period are comparable with previous studies that used LSTM-based models. To set a comparison framework, we also evaluated our LSTM models using RMSE over the test period (see Figs. B1 and B2 in Appendix B). The best median values that we obtained reached 0.72-0.73°C for the whole period thanks to regional training with catchment attributes. These values are comparable to those obtained by Rahmani et al. (2021b), who reached a median RMSE of 0.69°C thanks to more catchment and reach attributes (up to 21 static attributes against 10 in this study) and access to observations of streamflow at the T_w stations. In addition, our results are better than those obtained by Topp et al. (2023), who attempted at a spatiotemporal extrapolation of T_w regimes by testing more sophisticated deep-learning architectures that exploited both spatial and temporal information (RMSE of about 1.64°C). However, our results are still far away from those reported by Feigl et al. (2021), who compared several machine-learning and deep-learning techniques for ten Austrian catchments and obtained an average RMSE of 0.554°C. Their better performances can be attributed either to their methodological setup, in which they also optimized the model structure and the regression technique, or the use of additional features that explicitly encoded seasonality (such as the month of the year). In our study, we limited our analysis to the best overall deep-learning method over a larger sample of catchments following Rahmani et al. (2021b), and we did not apply feature engineering so that we could quantify the upper limit of performances that could be reached with only variables that are explicitly linkable to T_w. However, none of the previous studies evaluated the performances for the range of high T_w. We found that the best RMSE values at the top-10% range were at 0.88°C-0.89°C, which is worse by ~0.16°C on average compared to the whole-period evaluation.

We tested several loss functions to see whether increasing the weight of high T_w values could result in better performances for the top 10% range, following previous works in process-based environmental modelling (see e. g. Jadon et

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

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. Conversely, our results show that it is (slightly) better to use MAE-based loss functions instead of MSE-based ones especially for the reproduction of the overall thermal regime. Nonetheless, we could not obtain statistically significant improvements over the top 10% range following this strategy, which suggests that using the reference loss function (MSE on standardized target) is not a bad choice after all for the simulation of extreme T_w values.

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 capabilities of LSTM models (see the discussion in Hashemi et al., 2022). Second, better input features can be used to construct more robust LSTM models, by explicitly embedding of the month/season of the year (as in Feigl et al., 2021), better streamflow simulations using distributed hydrological models and/or deep-learning techniques (Rahmani et al., 2021b; Seyedhashemi et al., 2023), and other dynamic variables that can reflect changes in landscape and reach-scale anthropogenic influences, namely riparian vegetation and storage facilities (Jackson et al., 2018; Seyedhashemi et al., 2021, 2023). Specifically, streamflow simulations could be enhanced by accounting for snow dynamics, which strongly control thermal regimes especially at headwater locations (Wade et al., 2023). 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 better optimization hyperparameters (e.g., scheduling of the learning rate), training the LSTM on the whole range and then finetuning it on the target range, or designing a custom loss function that is a weighted sum of losses on the whole range and losses on the target range.

6 Conclusion and future research directions

We tested several strategies to improve the performances of LSTM-based thermal models for the specific task of reproducing high T_w values (top 10%). Our tests over a sample of 21 stations in the Garonne river catchment showed that:

1. Regionally training the LSTM model with input features that include static attributes is the best strategy to get a satisfactory reproduction of not only high T_w values but also the whole thermal regime.
2. The added value of emphasizing model errors at extreme T_w values in the loss function used for training is limited compared to the added value of regional, multi-catchment training. However, some improvements can be obtained by using MAE as a loss function (with or without standardization of the target variable) compared to the traditional quadratic loss functions (MSE or NSE).

Future research includes a ranking of local, reach attributes and regional, catchment attributes by quantifying their importance in controlling stream temperature. This will inform management and adaptation strategies by indicating the scale

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

(local or regional) at which they should be implemented. Furthermore, coupling a thermal model with a hydrological model is a promising way for a better, joint reproduction of streamflow and stream temperature dynamics. This coupling will raise the more general question of the best coupling strategy (fully process-based, hybrid, or fully data-driven as in Rahmani et al., 2021b) in terms of both model performances, model interpretability, and extrapolation capabilities. It will not only benefit thermal models but also the characterization of hydrological response thanks to the joint data assimilation of both T_w and streamflow observations.

Appendix A: Training time

Statistics of the wall-clock time spent to train each one of the LSTM models are summarized in Table A1. We trained a total of 4536 local models (18 loss functions, 4 input variables, 3 lookbacks, and 21 stations) and 378 regional models (18 loss functions, 7 input variables, 3 lookbacks) on three computing units with distinct CPU cores (Intel® Xeon® W-2265 with base frequency at 3.5 GHz, Intel® Xeon® W3-2425 with base frequency at 3.0 GHz, and Intel® Xeon® Silver 4216 with base frequency at 2.1 GHz). At each computing unit, we set the number of threads dedicated to each model training at 7. Note that the training was halted once a plateau in the validation performances was reached (100 epochs without a decrease in MSE by 0.001°C^2). Table A1 shows that the training time increases at the first order with the lookback and the amount of data used for training (local vs. regional training). Training a local model takes ~3 min on average for a lookback of 30 days up to ~45 min on average for a lookback of 365 days. Training a regional model takes ~45 min on average for a lookback of 30 days and up to ~12 h for a lookback of 365 days.

Table A1: Statistics of the wall-clock time (in min) needed for the training of the local and regional LSTM models. The number of models from which the statistics were calculated is specified between brackets for each training strategy and each lookback.

Local/regional	Lookback	5th percentile	Median	95th percentile
Locally trained models (4536)	30 (1512 models)	1	3	10
	90 (1512 models)	4	10	45
	365 (1512 models)	15	43	600
Regionally trained models (378)	30 (126 models)	17	45	94
	90 (126 models)	62	159	366
	365 (126 models)	256	681	1621

Appendix B: Test performances using RMSE

To compare our results with previous studies (e.g., Rahmani et al., 2021b), we also evaluated the test performances using RMSE as a criterion. Figures B1 and B2 show the performances of the locally trained (Fig. B1) and the regionally trained (Fig. B2) LSTM models, which are duplicates of Figs. 3 and 4. They 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.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

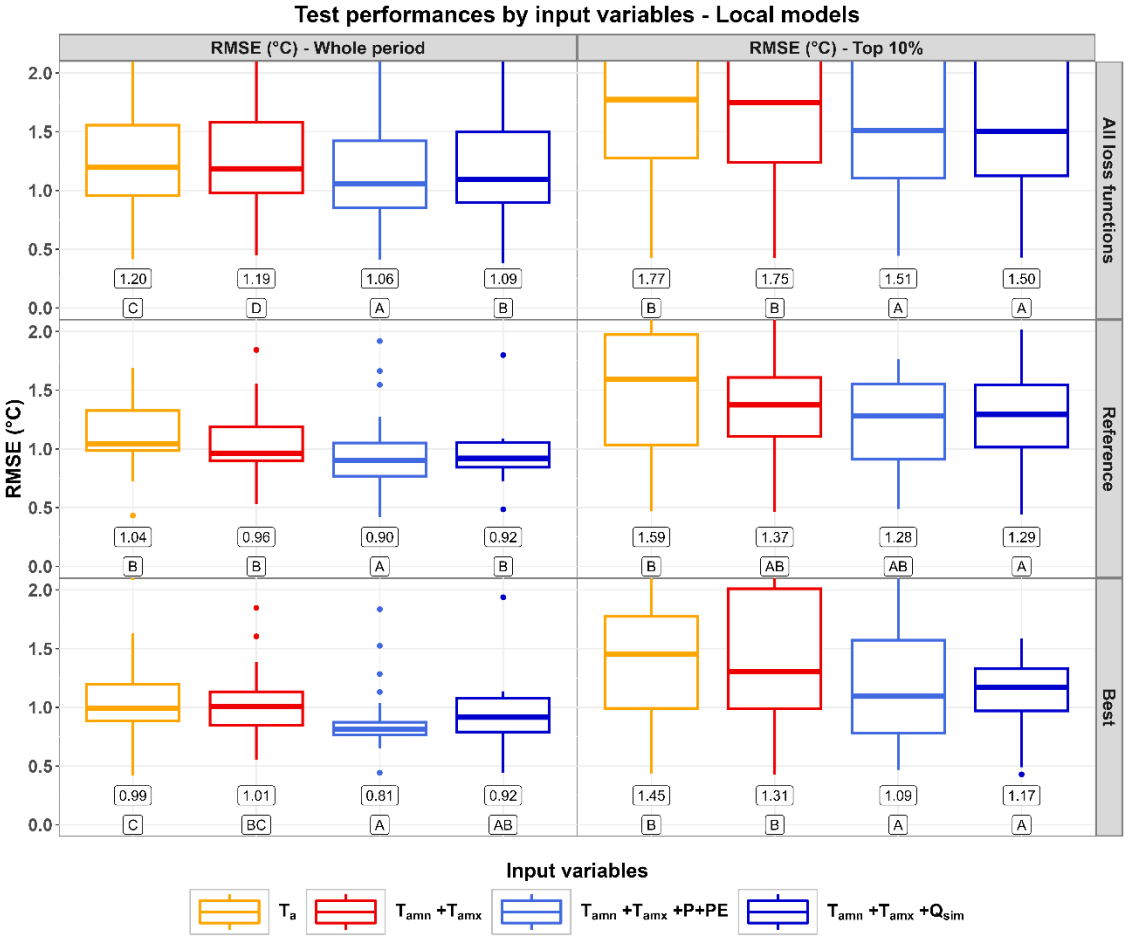

Figure B1: Distributions of the test performances (RMSE, in °C) of the local models over the whole test period (left column) and the period corresponding to the highest 10% observed values (right column). The top row shows the distribution across all the loss functions (18 × 21 = 378 points per distribution). 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 (MAE without standardization for the whole period and M4E 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.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

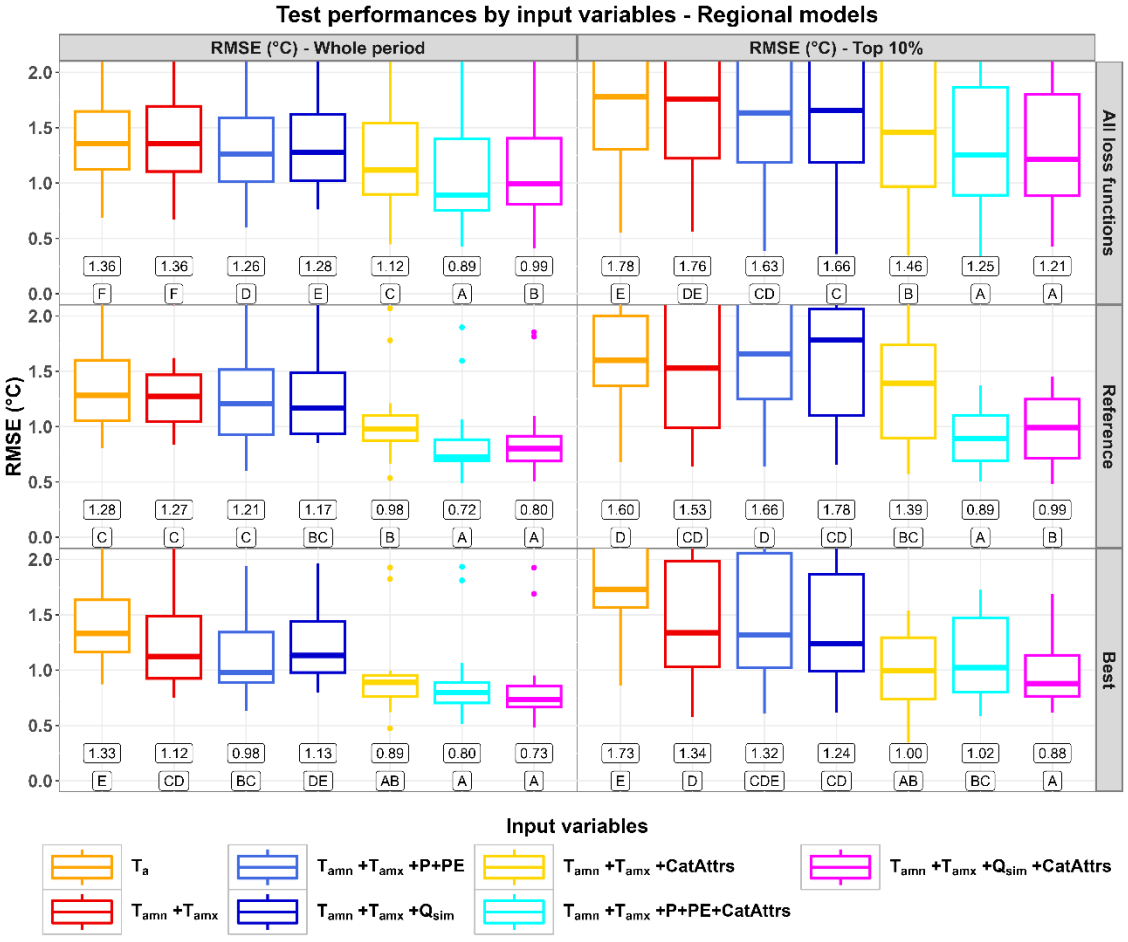

Figure B2: Distributions of the test performances (RMSE, in °C) of the regionally trained models over the whole test period (left column) and the period corresponding to the highest 10% observed values (right column). Top row shows the distribution across all the loss functions (18 × 21 = 378 points per distribution). The middle row shows the performances across the reference loss function (MSE with standardization, 21 points per distribution). The bottom row shows the performances with the best loss function (MSE without standardization for both cases). 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.

Code and data availability

Python scripts used to train the LSTM models, the obtained performances of all tested LSTM models, as well as processed data for a subset of our catchment dataset can be downloaded from https://www.doi.org/10.5281/zenodo.15864784 (Saadi, 2025). Additional R and python libraries that we used were mentioned in the text, and links to download the publicly available datasets were provided in Table 1. Any questions regarding the code or the data can be addressed to the corresponding author (mohamed.saadi@toulouse-inp.fr).

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

Author contribution

LL preprocessed the stream temperature data. MS, LG and GC prepared the input data. LG and GC wrote the initial training scripts and computed the preliminary results during their master's internships supervised by MS and HR. MS rewrote the final training scripts, conceptualized the methodological setup, computed the results and produced the figures shown in this study. MS wrote the original draft and accounted for suggestions and modifications proposed by LG, GC, LL and HR.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to thank Kévin Duplan, Olivier Louis and Vincent Ribot from the Haute-Garonne Departmental Council (CD31) for coordinating the collection of the stream temperature dataset used in this study and for funding the master's internships of Gabrielle Cognot (2023) and Louis Guichard (2024). This work is part of the regional project (*Projet de Territoire*) Garonne-Amont (https://garonne-amont.fr/; last access: 14 July 2025, in French) which aims at tackling the challenges of water management at the upper Garonne catchment by creating an observatory for water resources (quantity and quality) at the catchment scale (among other actions).

References

Alfonso, S., Gesto, M., and Sadoul, B.: Temperature increase and its effects on fish stress physiology in the context of global warming, Journal of Fish Biology, 98, 1496–1508, https://doi.org/10.1111/jfb.14599, 2021.

Armstrong, J. B., Fullerton, A. H., Jordan, C. E., Ebersole, J. L., Bellmore, J. R., Arismendi, I., Penaluna, B. E., and Reeves, G. H.: The importance of warm habitat to the growth regime of cold-water fishes, Nat. Clim. Chang., 11, 354–361, https://doi.org/10.1038/s41558-021-00994-y, 2021.

Arora, R., Tockner, K., and Venohr, M.: Changing river temperatures in northern Germany: trends and drivers of change, Hydrological Processes, 30, 3084–3096, https://doi.org/10.1002/hyp.10849, 2016.

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

Audouy, J.-N., Pitsch ,Stéphanie, Renard ,Benjamin, and and Chaleon, C.: Statistiques hydrologiques en crue : de la Banque HYDRO à l'HydroPortail, LHB, 110, 2317798, https://doi.org/10.1080/27678490.2024.2317798, 2024.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

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

Beaufort, A., Diamond, J. S., Sauquet, E., and Moatar, F.: Spatial extrapolation of stream thermal peaks using heterogeneous time series at a national scale, Hydrology and Earth System Sciences, 26, 3477–3495, https://doi.org/10.5194/hess-26-3477-2022, 2022.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. M. J., and Bobée, B.: A Review of Statistical Water Temperature Models, Canadian Water Resources Journal, 32, 179–192, https://doi.org/10.4296/cwrj3203179, 2007.

Bonacina, L., Fasano, F., Mezzanotte, V., and Fornaroli, R.: Effects of water temperature on freshwater macroinvertebrates: a systematic review, Biological Reviews, 98, 191–221, https://doi.org/10.1111/brv.12903, 2023.

Bowerman, T., Roumasset, A., Keefer, M. L., Sharpe, C. S., and Caudill, C. C.: Prespawn Mortality of Female Chinook Salmon Increases with Water Temperature and Percent Hatchery Origin, Transactions of the American Fisheries Society, 147, 31–42, https://doi.org/10.1002/tafs.10022, 2018.

Briggs, M. A., Johnson, Z. C., Snyder, C. D., Hitt, N. P., Kurylyk, B. L., Lautz, L., Irvine, D. J., Hurley, S. T., and Lane, J. W.: Inferring watershed hydraulics and cold-water habitat persistence using multi-year air and stream temperature signals, Science of The Total Environment, 636, 1117–1127, https://doi.org/10.1016/j.scitotenv.2018.04.344, 2018.

Buisson, L., Blanc, L., and Grenouillet, G.: Modelling stream fish species distribution in a river network: the relative effects of temperature versus physical factors, Ecology of Freshwater Fish, 17, 244–257, https://doi.org/10.1111/j.1600-0633.2007.00276.x, 2008.

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

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

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

Comte, L., Buisson, L., Daufresne, M., and Grenouillet, G.: Climate-induced changes in the distribution of freshwater fish: observed and predicted trends, Freshwater Biology, 58, 625–639, https://doi.org/10.1111/fwb.12081, 2013.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., and Andréassian, V.: The suite of lumped GR hydrological models in an R package, Environmental Modelling and Software, 94, 166–171, https://doi.org/10.1016/j.envsoft.2017.05.002, 2017.

Coron, L., Delaigue, O., Thirel, G., Dorchies, D., Perrin, C., and Michel, C.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling, https://doi.org/10.15454/EX11NA, 2023.

Daufresne, M. and Boët, P.: Climate change impacts on structure and diversity of fish communities in rivers, Global Change Biology, 13, 2467–2478, https://doi.org/10.1111/j.1365-2486.2007.01449.x, 2007.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

Delpla, I., Jung, A.-V., Baures, E., Clement, M., and Thomas, O.: Impacts of climate change on surface water quality in relation to drinking water production, Environment International, 35, 1225–1233, https://doi.org/10.1016/j.envint.2009.07.001, 2009.

Ducharne, A.: Importance of stream temperature to climate change impact on water quality, Hydrology and Earth System Sciences, 12, 797–810, https://doi.org/10.5194/hess-12-797-2008, 2008.

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

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

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

Eilander, D., Van Verseveld, W., Yamazaki, D., Weerts, A., Winsemius, H. C., and Ward, P. J.: A hydrography upscaling method for scale-invariant parametrization of distributed hydrological models, Hydrol. Earth Syst. Sci., 25, 5287–5313, https://doi.org/10.5194/hess-25-5287-2021, 2021.

Essaid, H. I. and Caldwell, R. R.: Evaluating the impact of irrigation on surface water – groundwater interaction and stream temperature in an agricultural watershed, Science of The Total Environment, 599–600, 581–596, https://doi.org/10.1016/j.scitotenv.2017.04.205, 2017.

FAO and IIASA: Harmonized World Soil Database version 2.0, Rome and Laxenburg, 2023.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D.: The Shuttle Radar Topography Mission, Reviews of Geophysics, 45, https://doi.org/10.1029/2005RG000183, 2007.

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

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrology and Earth System Sciences, 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.

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

Garcia, F., Folton, N., and Oudin, L.: Which objective function to calibrate rainfall–runoff models for low-flow index simulations?, Hydrological Sciences Journal, 62, 1149–1166, https://doi.org/10.1080/02626667.2017.1308511, 2017.

Hannah, D. M. and Garner, G.: River water temperature in the United Kingdom: Changes over the 20th century and possible changes over the 21st century, Progress in Physical Geography: Earth and Environment, 39, 68–92, https://doi.org/10.1177/0309133314550669, 2015.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

Hare, D. K., Helton, A. M., Johnson, Z. C., Lane, J. W., and Briggs, M. A.: Continental-scale analysis of shallow and deep groundwater contributions to streams, Nat Commun, 12, 1450, https://doi.org/10.1038/s41467-021-21651-0, 2021.

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

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

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochreiter, S., and Klambauer, G.: MC-LSTM: Mass-Conserving LSTM, in: Proceedings of the 38th International Conference on Machine Learning, International Conference on Machine Learning, 4275–4286, 2021.

Hufkens, K. and Labs, B.: hwsdr: Interface to the "HWSD" Web Services, 2023.

Isaak, D. J., Peterson, E. E., Ver Hoef, J. M., Wenger, S. J., Falke, J. A., Torgersen, C. E., Sowder, C., Steel, E. A., Fortin, M., Jordan, C. E., Ruesch, A. S., Som, N., and Monestiez, P.: Applications of spatial statistical network models to stream data, WIREs Water, 1, 277–294, https://doi.org/10.1002/wat2.1023, 2014.

Jackson, F. L., Fryer, R. J., Hannah, D. M., Millar, C. P., and Malcolm, I. A.: A spatio-temporal statistical model of maximum daily river temperatures to inform the management of Scotland's Atlantic salmon rivers under climate change, Science of The Total Environment, 612, 1543–1558, https://doi.org/10.1016/j.scitotenv.2017.09.010, 2018.

Jadon, A., Patil, A., and Jadon, S.: A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting, in: Data Management, Analytics and Innovation, Singapore, 117–147, https://doi.org/10.1007/978-981-97-3245-6_9, 2024.

Johnson, Z. C., Johnson, B. G., Briggs, M. A., Devine, W. D., Snyder, C. D., Hitt, N. P., Hare, D. K., and Minkova, T. V.: Paired air-water annual temperature patterns reveal hydrogeological controls on stream thermal regimes at watershed to continental scales, Journal of Hydrology, 587, 124929, https://doi.org/10.1016/j.jhydrol.2020.124929, 2020.

Jones, R., Travers, C., Rodgers, C., Lazar, B., English, E., Lipton, J., Vogel, J., Strzepek, K., and Martinich, J.: Climate change impacts on freshwater recreational fishing in the United States, Mitig Adapt Strateg Glob Change, 18, 731–758, https://doi.org/10.1007/s11027-012-9385-3, 2013.

Kędra, M. and Wiejaczka, Ł.: Climatic and dam-induced impacts on river water temperature: Assessment and management implications, Science of The Total Environment, 626, 1474–1483, https://doi.org/10.1016/j.scitotenv.2017.10.044, 2018.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://doi.org/10.48550/arXiv.1412.6980, 2017.

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

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

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

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

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

Larnier, K., Roux, H., Dartus, D., and Croze, O.: Water temperature modeling in the Garonne River (France), Knowl. Managt. Aquatic Ecosyst., 04, https://doi.org/10.1051/kmae/2010031, 2010.

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

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

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

Macedo, M. N., Coe, M. T., DeFries, R., Uriarte, M., Brando, P. M., Neill, C., and Walker, W. S.: Land-use-driven stream warming in southeastern Amazonia, Philosophical Transactions of the Royal Society B: Biological Sciences, 368, 20120153, https://doi.org/10.1098/rstb.2012.0153, 2013.

Maire, A., Thierry, E., Viechtbauer, W., and Daufresne, M.: Poleward shift in large-river fish communities detected with a novel meta-analysis framework, Freshwater Biology, 64, 1143–1156, https://doi.org/10.1111/fwb.13291, 2019.

Moatar, F. and Gailhard, J.: Water temperature behaviour in the River Loire since 1976 and 1881, Comptes Rendus Geoscience, 338, 319–328, https://doi.org/10.1016/j.crte.2006.02.011, 2006.

Mohseni, O., Stefan, H. G., and Erickson, T. R.: A nonlinear regression model for weekly stream temperatures, Water Resources Research, 34, 2685–2692, https://doi.org/10.1029/98WR01877, 1998.

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

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

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., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

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

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

Ouellet, V., St-Hilaire, A., Dugdale, S. J., Hannah, D. M., Krause, S., and Proulx-Ouellet, S.: River temperature research and practice: Recent challenges and emerging opportunities for managing thermal habitat conditions in stream ecosystems, Science of The Total Environment, 736, 139679, https://doi.org/10.1016/j.scitotenv.2020.139679, 2020.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A.: Automatic differentiation in PyTorch, in: NIPS-W, 2017.

Picard, C., Floury, M., Seyedhashemi, H., Morel, M., Pella, H., Lamouroux, N., Buisson, L., Moatar, F., and Maire, A.: 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, 2022.

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

Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, Journal of Hydrology, 420–421, 171–182, https://doi.org/10.1016/j.jhydrol.2011.11.055, 2012.

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

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

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C.: 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, 2021b.

Redfern, T. W., Macdonald, N., Kjeldsen, T. R., Miller, J. D., and Reynard, N.: Current understanding of hydrological processes on common urban surfaces, Progress in Physical Geography: Earth and Environment, 40, 699–713, https://doi.org/10.1177/0309133316652819, 2016.

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

Saadi, M.: 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, 2025.

Saadi, M. and Furusho-Percot, C.: Which range of streamflow data is most informative in the calibration of an hourly hydrological model?, Hydrological Sciences Journal, 69, 1–20, https://doi.org/10.1080/02626667.2023.2277835, 2024.

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

Sadler, J. M., Appling, A. P., Read, J. S., Oliver, S. K., Jia, X., Zwart, J. A., and Kumar, V.: Multi-Task Deep Learning of Daily Streamflow and Water Temperature, Water Resources Research, 58, e2021WR030138, https://doi.org/10.1029/2021WR030138, 2022.

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

Segura, C., Caldwell, P., Sun, G., McNulty, S., and Zhang, Y.: A model to predict stream water temperature across the conterminous USA, Hydrological Processes, 29, 2178–2195, https://doi.org/10.1002/hyp.10357, 2015.

Seyedhashemi, H., Moatar, F., Vidal, J.-P., Diamond, J. S., Beaufort, A., Chandesris, A., and Valette, L.: Thermal signatures identify the influence of dams and ponds on stream temperature at the regional scale, Science of The Total Environment, 766, 142667, https://doi.org/10.1016/j.scitotenv.2020.142667, 2021.

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

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

Souaissi, Z., Ouarda, T. B. M. J., and St-Hilaire, A.: Non-parametric, semi-parametric, and machine learning models for river temperature frequency analysis at ungauged basins, Ecological Informatics, 75, 102107, https://doi.org/10.1016/j.ecoinf.2023.102107, 2023.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, 15, 1929–1958, 2014.

Stefan, H. G. and Preud'homme, E. B.: Stream Temperature Estimation from Air Temperature, JAWRA Journal of the American Water Resources Association, 29, 27–45, https://doi.org/10.1111/j.1752-1688.1993.tb01502.x, 1993.

Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., Laurantin, O., Merlier, C., and Soubeyroux, J.-M.: A 10-year (1997–2006) reanalysis of Quantitative Precipitation Estimation over France: methodology and first results, IAHS Publication, 351, 255–260, 2012.

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

Toffolon, M. and Piccolroaz, S.: 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, 2015.

Topp, S. N., Barclay, J., Diaz, J., Sun, A. Y., Jia, X., Lu, D., Sadler, J. M., and Appling, A. P.: Stream Temperature Prediction in a Shifting Environment: Explaining the Influence of Deep Learning Architecture, Water Resources Research, 59, e2022WR033880, https://doi.org/10.1029/2022WR033880, 2023.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, https://doi.org/10.1002/joc.2003, 2010.

van Vliet, M. T. H., Yearsley, J. R., Ludwig, F., Vögele, S., Lettenmaier, D. P., and Kabat, P.: Vulnerability of US and European electricity supply to climate change, Nature Clim Change, 2, 676–681, https://doi.org/10.1038/nclimate1546, 2012.

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

https://doi.org/10.5194/egusphere-2025-3393
Preprint. Discussion started: 30 July 2025

Wade, J., Kelleher, C., and Hannah, D. M.: Machine learning unravels controls on river water temperature regime dynamics, Journal of Hydrology, 623, 129821, https://doi.org/10.1016/j.jhydrol.2023.129821, 2023.

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

Zhi, W., Feng, D., Tsai, W.-P., Sterle, G., Harpold, A., Shen, C., and Li, L.: 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, 2021.

Zhi, W., Ouyang, W., Shen, C., and Li, L.: Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers, Nat Water, 1, 249–260, https://doi.org/10.1038/s44221-023-00038-z, 2023.

Zhi, W., Appling, A. P., Golden, H. E., Podgorski, J., and Li, L.: Deep learning for water quality, Nat Water, 2, 228–241, https://doi.org/10.1038/s44221-024-00202-z, 2024.

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