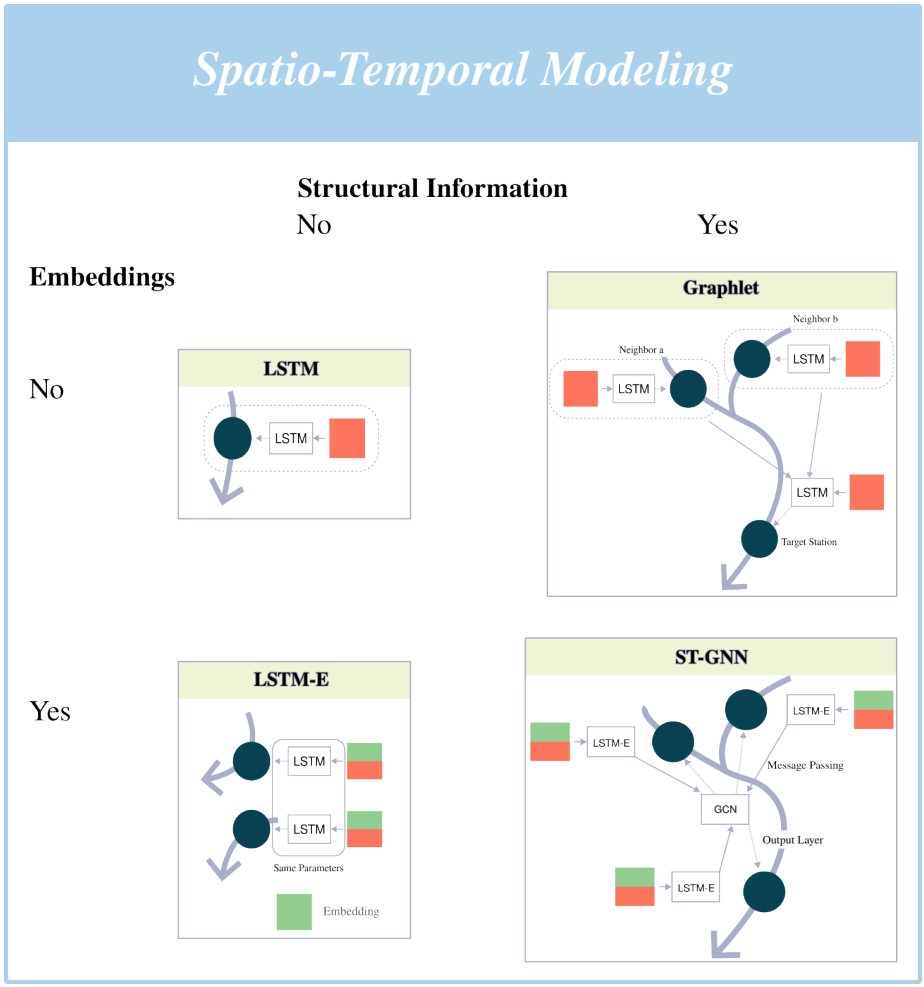
Graphical Abstract

Evaluating the Role of Graphs and Embeddings in Spatio-Temporal Modeling of River Water Temperature

Benjamin Fankhauser, Vidushi Bigler, Kaspar Riesen



Highlights

Evaluating the Role of Graphs and Embeddings in Spatio-Temporal Modeling of River Water Temperature

Benjamin Fankhauser, Vidushi Bigler, Kaspar Riesen

- Three curated river water temperature datasets for fair and robust comparison.
- Systematic evaluation of learnable embeddings and structural information impacts.
- Empirical evidence that structural information consistently improves accuracy.
- Embedding-based methods achieve strong parameter efficiency and scalability.

Evaluating the Role of Graphs and Embeddings in Spatio-Temporal Modeling of River Water Temperature

Benjamin Fankhauser[a,1,*], Vidushi Bigler[b,2], Kaspar Riesen[a,3]

[a] *Institute of Computer Science, University of Bern, Bern, Switzerland*
[b] *Institute for Optimisation and Data Analysis, Bern University of Applied Sciences, Biel, Switzerland*

Abstract

In climate research, modeling river water temperature is a key pattern recognition task for detecting trends and predicting change. As a geospatial time series problem, its predictive performance is highly sensitive to sensor choice and temporal coverage, limiting comparability across studies. As the first major contribution of this work, we curate three river water temperature datasets with a well-defined experimental setup (specifying sensor choice and temporal partitions) to enable fair and robust comparisons across diverse scenarios. Building on this foundation, our second major contribution examines the role of learnable embeddings and structural information – two recently introduced concepts in river water temperature modeling. To this end, we propose and evaluate four models representing all combinations of these components: *LSTM*, *Graphlet*, *LSTM-E*, and *ST-GNN* (a spatio-temporal graph neural network). Our experimental evaluation corroborates the predictive performance reported in prior studies, while revealing substantial variability across water stations. Crucially, we provide statistical evidence that incorporating structural information consistently enhances baseline models across all datasets. Moreover, the marked reduction in learnable parameters – a defining feature of embedding-

*Corresponding author
Email addresses: benjamin.fankhauser@unibe.ch (Benjamin Fankhauser), vidushi.bigler@bfh.ch (Vidushi Bigler), kaspar.riesen@unibe.ch (Kaspar Riesen)
[1]ORCID: 0000-0002-7982-2669
[2]ORCID: 0000-0001-6043-8264
[3]ORCID: 0000-0002-9145-3157

based methods – demonstrates their efficiency and underscores their potential for future river water temperature modeling architectures.
Keywords: Water Temperature Prediction, Spatio-Temporal Modeling, Time Series Forecasting, Graph Neural Networks, Embeddings, Deep Learning, Climate Change

1. Introduction

River water temperature is a critical variable in freshwater ecosystems, influencing diverse ecological and biological processes that affect flora, fauna, agriculture, and drinking water quality [1, 2]. As rivers and their tributaries occur across most habitable regions, understanding and managing their thermal regimes is essential. This paper investigates novel pattern recognition approaches to modeling river water temperature, addressing the need for accurate and adaptable predictive tools. Such models enable precise identification of critical river sections under projected climate scenarios and provide a robust basis for analyzing complex environmental interactions.

The need for such modeling is underscored by the impacts of ongoing climate change [3], which affects air temperature and, in turn, leads to rising water temperatures. For instance, over the past 40 years, river water temperatures in Switzerland have exhibited a consistent warming trend ranging from 0.2 °C to 0.8 °C per decade, with a median of 0.4 °C (evident in Figure 1, where annual changes are estimated by fitting a linear regression to the yearly averages at each water station[4]). Such warming poses a significant threat to aquatic ecosystems, as elevated temperatures reduce dissolved oxygen levels while simultaneously increasing the metabolic rates of aquatic organisms [5]. Over the long term, such conditions can degrade water quality, promote the proliferation of harmful algal blooms, and, in severe cases, precipitate the collapse of entire aquatic ecosystems [6, 7]. Beyond ecological impacts, Switzerland's legal 25 °C threshold

[4]Note that more comprehensive trend analyses are available [4].

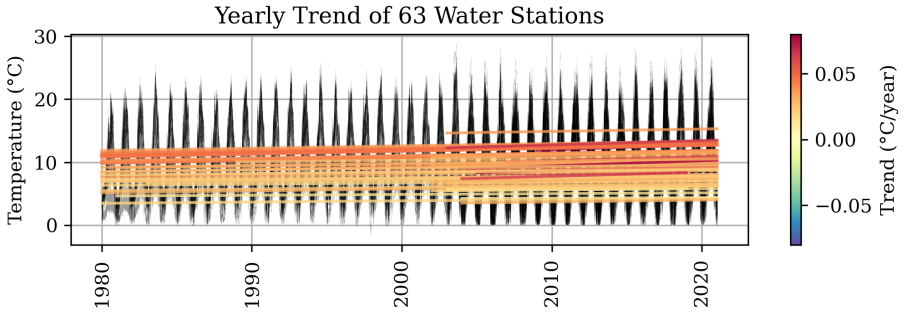


Figure 1: Yearly trends in water temperature of the Swiss River Network [11] since 1980.

for river water temperature is now exceeded annually, forcing nuclear power plants to reduce or suspend operations and threatening the survival of cold-water species such as trout [8]. These developments underscore the urgent need for accurate predictive models capable of capturing complex spatio-temporal dynamics in river systems [9, 10].

Understanding the drivers of these elevated temperatures is essential [12]. Meteorological processes play a primary role: solar radiation heats both the riverbed and the tributaries within the catchment area, while atmospheric exchange processes affect the surface layer – particularly in lakes. In alpine regions, snow and glacier melt during spring and summer also exert substantial influence. In addition, anthropogenic activities further contribute to thermal variation, including altered discharge regimes from hydropower operations, cooling-water discharges from nuclear facilities, and thermal inputs from urban runoff, particularly when rainwater drains from heat-absorbing surfaces such as roofs and asphalt. This interplay of natural and human induced factors, yields a complex system characterized by pronounced nonlinear dynamics.

River water temperature in Switzerland has been monitored by the *Federal Office for the Environment* (FOEN) for over 40 years. Measurements are taken at so-called water stations in 10-minute intervals and reported as daily averages, with sensors regularly inspected and calibrated. To improve monitoring in response to climate change, the FOEN doubled the number of water stations from

approximately 40 to 80 between the years 2000 and 2010. Complementing the federal network, individual cantons – Switzerland's semi-autonomous regional units – operate additional monitoring sites, such as those run by the *Canton of Zurich*. Air temperature data from nearby *MeteoSwiss* stations supports the integration of atmospheric and hydrological observations.

Despite this extensive monitoring, predictive modeling remains somehow fragmented. Many approaches treat monitoring stations independently, neglecting spatial relationships in the river network. Others rely on inconsistent datasets with heterogeneous sensor choice and temporal coverage, limiting reproducibility and comparability across studies. As a result, it is still unclear how recent advances such as learnable embeddings and structural information affect predictive performance and parameter efficiency in this domain. This paper addresses these challenges through two key contributions:

- First, we introduce three curated datasets with a well-defined experimental setup, specifying sensor choice and temporal partitions, and including a novel high-resolution dataset for the Zurich region. These datasets enable robust and fair comparisons across diverse geographical and temporal conditions, using standard performance metrics with particular focus on generalization and parameter efficiency.

- Second, building on this foundation, we systematically evaluate the role of learnable embeddings and structural information in the spatio-temporal modeling of river water temperature. Recognizing that many prior models either treat monitoring stations independently or overlook the spatial relationships between them, we propose a unified framework to assess models representing combinations of embeddings and structural components: a baseline *LSTM*, a structurally enhanced model named *Graphlet*, an embedding-based model termed *LSTM-E*, and a spatio-temporal graph neural network (*ST-GNN*).

By consolidating our preliminary efforts [11, 13, 14] into a unified experimental framework, this work establishes one of the first reproducible benchmarks

for river water temperature modeling.

The remainder of this paper is organized as follows. In Section 2, we discuss related work and position our contribution in the context of existing research. Section 3 describes the proposed and evaluated models in this study. Section 4 provides a detailed description of the datasets. The experimental setup and the empirical evaluation is detailed in Section 5. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Related Work

A variety of methods have been proposed to model river water temperature. This section reviews the current state of the art in this field.

A prominent example is *Air2Stream* [15], a physically inspired model designed to capture the relationship between air temperature and river water temperature, using air temperature and river discharge as primary inputs. The core of the model is a differential equation that describes the heat exchange processes between the atmosphere and the water body. This equation is linearized through a Taylor series expansion, resulting in a tractable form with eight tunable parameters. These parameters are calibrated using observed data, typically via particle-based optimization methods. The performance of Air2Stream is highly sensitive to the choice of hyperparameters and initial conditions, rendering its calibration process potentially unstable and computationally demanding. Despite these limitations, the model has been widely adopted due to its physical interpretability.

Given that modeling river water temperature is inherently a time series problem, various *Long Short-Term Memory* (LSTM) models [16] have been proposed for this task. LSTM networks are a specialized form of *recurrent neural networks* (RNNs) designed to model sequential data, such as time series, by learning temporal dependencies across varying time scales. Unlike standard RNNs, which often struggle to capture long-term patterns due to issues such as vanishing or exploding gradients, LSTMs incorporate an internal architecture composed of

a hidden state and a memory (or cell) state. These states are propagated and updated across time steps through gated mechanisms – namely *input*, *output*, and *forget gates* – that regulate the flow of information and preserve relevant signals over long sequences.

Training LSTM models involves a variant of backpropagation known as *back-propagation through time* (BPTT), in which gradients are computed across multiple time steps to update model parameters [17]. To maintain numerical stability, training is typically performed using fixed-length input windows, over which gradients are accumulated before updates are applied. This design enables LSTMs to mitigate the vanishing gradient problem, making them particularly effective for long-range sequence modeling tasks such as river water temperature prediction. In most applications, LSTMs are used to model the relationship between meteorological variables (most commonly air temperature) and river water temperature.

Qiu et al. [18], for instance, investigate the use of LSTM networks for daily river water temperature forecasting under the influence of climate change and anthropogenic disturbances, such as dam construction. Using data from nine river gauges worldwide, they demonstrate that LSTM models outperform alternative methods in capturing mean daily temperature variations. Moreover, a detailed case study on the *Yangtze River* shows that the LSTM effectively reconstructs natural thermal conditions and quantifies the impact of the *Three Gorges Reservoir*, revealing strong seasonal shifts in water temperature. The study highlights LSTM's potential as a robust tool for water temperature prediction and riverine ecosystem management.

Jia et al. [19] propose a physics-guided machine learning approach that integrates RNNs with domain knowledge from physics-based models to enhance the prediction of river water temperature and streamflow. Their method incorporates a novel loss function to balance prediction accuracy across diverse river segments and leverages sparse training data. Applied to the *Delaware River Basin*, the approach outperforms both state-of-the-art physics-based models and standard LSTM models, and also demonstrates strong generalization across

seasons and varying streamflow conditions.

Bao et al. [20] propose a dynamic graph network in which spatial relationships between river segments are explicitly modeled over time using heat-transfer partial differential equations. This formulation enables the model to effectively capture evolving spatio-temporal dependencies and physical interactions within complex river networks, particularly under irregular spatial configurations and variable flow conditions. Consequently, this model combines process-based physical principles with the adaptability of deep learning to deliver generalizable water temperature predictions even with limited observational data.

Moshe et al. [21] explore an approach to improve hydrologic predictions by incorporating the natural structure of river networks into the modeling framework. Their method, a deep neural network termed *HydroNets*, combines basin-specific components with shared model elements, enabling the joint processing of temporal (atmospheric) data, static features, and upstream activities within a unified architecture. Evaluated on two large basins in India, *HydroNets* demonstrate superior performance and improved sample efficiency.

In addition to water temperature, LSTMs are also successfully employed for predicting water levels and discharge. Kim et al. [22], for instance, address the challenge of accurate water level prediction by combining LSTM networks with complex network analysis. To enhance model accuracy, the complex network method is employed to identify the most informative input data, avoiding irrelevant signals from nearby stations. Results show that the LSTM model augmented with complex network selection achieves superior performance. Zhao et al. [23] propose *ST-Hydro*, a spatio-temporal framework for river discharge prediction that integrates both temporal dynamics and spatial relationships among hydrological gauges. Unlike traditional models that focus on isolated basins, their approach constructs and fuses three types of graphs – hydraulic distance, Euclidean distance, and correlation – to represent geographic and hydrological connections. Experiments on real-world datasets demonstrate that ST-Hydro can effectively predict river discharge and provide early warnings.

		Structural Information	
		without	with
Embeddings	without	LSTM	Graphlet
	with	LSTM-E	ST-GNN

Table 1: Comparison of the four models with respect to whether or not they use structural information and, or embeddings.

3. Methods

In this paper, we focus on the task of water temperature modeling, which involves predicting the estimated water temperature $\hat{w}^{(t)}$ at time step t, given an observed air temperature time series $(a^{(1)}, a^{(2)}, \ldots, a^{(t)})$. The objective is to learn a model function f such that

$$f(a^{(1)}, a^{(2)}, \ldots, a^{(t)}) = \hat{w}^{(t)}, \tag{1}$$

where $\hat{w}^{(t)}$ denotes the model's prediction (which is generally distinct from the actual measured water temperature $w^{(t)}$).

3.1. Embeddings and Structural Information

In this paper, we research four models[5] designed to address the task formulated in Equation (1), which can be distinguished by whether they incorporate *embeddings* and, or whether they incorporate *structural information* (see Table 1 for an overview).

Station-specific models without embeddings, i.e. separate functions f_k for each water station k, work well for small datasets [1] and offer flexibility, as they can be developed, trained, and deployed independently. They also reduce bias by avoiding simultaneous training on the entire dataset, which can overrepresent certain catchments or regions. Yet, these models have key drawbacks: as the number of stations grows, learnable parameters and management overhead

[5]Note that each of these models has been introduced individually in earlier preliminary studies [11, 13, 14], their systematic consolidation and integration, as undertaken in this work, is novel.

increase linearly. They also struggle to generalize to unseen stations and often learn redundant patterns, leading to inefficiency and duplicated effort.

Embedding-based approaches, on the other hand, overcome these issues by using a single model to capture the overall problem structure by means of embeddings. That is, *embeddings* represent station-specific characteristics – either static (e.g., soil type, location or altitude of the station) or learned during training – designed so that stations with similar properties may have similar embeddings. Hence, rather than training a separate function f_k for each water station k, a single shared model f can be used, conditioned on a station-specific embedding vector $\mathbf{e}_k$ as $f(\cdot, \mathbf{e}_k)$.

Models *without structural information*, on the other hand, treat stations as independent (or only weakly linked via shared parameters), relying solely on meteorological and station-specific attributes. This simplifies design, avoids dependence on graph construction, and remains applicable when spatial relationships are unknown. However, it cannot exploit upstream-downstream dependencies, may reduce accuracy in spatially correlated regions, and often leads to redundant learning across nearby stations.

Models *with structural information* represent the river network as a graph $G = (V, E)$, where each water station k corresponds to a vertex $v_k \in V$, and an edge $e_{i,j} = (v_i, v_j) \in E$ indicates water flow between stations i and j. Because information from downstream stations can be valuable, all edges are bidirectional. While this choice reflects the natural flow of water, alternative graph constructions – such as Delaunay triangulation or manually defined connectivity – would also be feasible. These models enable message passing between connected stations, capture spatial dependencies, improves accuracy where upstream conditions influence downstream stations, and support generalization to sparsely observed stations by borrowing information from neighbors. The approach, however, requires accurate connectivity data and can degrade if the graph is noisy or misrepresents true hydrological relationships.

In Table 2, we summarize the main benefits and drawbacks of incorporating embeddings and structural information in river water temperature modeling.

Approach	Advantages	Limitations
Station-specific	Effective for small datasets; independent development, training, and deployment; reduce bias by avoiding simultaneous training on the entire dataset.	Parameters and management overhead grow linearly with the number of stations; cannot generalize to unseen stations; often learn redundant patterns, leading to inefficiency.
Embedding-based	Use a single model to capture overall problem structure; encode station-specific characteristics via embeddings; improve efficiency and generalization.	Require a shared representation space; may underperform if embeddings fail to capture relevant station features.
Without structure	Simple design; independent training and deployment; works without known spatial relationships.	Cannot exploit upstream–downstream dependencies; may reduce accuracy in spatially correlated regions; redundant learning across nearby stations.
With structure	Captures spatial dependencies; improves accuracy when upstream conditions matter; supports generalization to sparsely observed stations.	Requires accurate connectivity data; performance can degrade with noisy or incorrect graphs.

Table 2: Comparison of station-specific and embedding-based approaches with and without structural information.

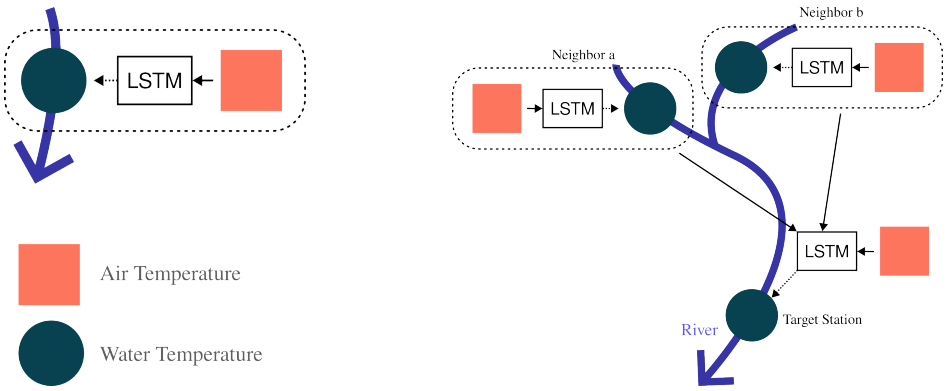
3.2. Station-Specific Models

In our prior work [11], we propose and employ two station-specific models, where the same architecture is trained independently for each station (see Figure 2 for an overview of these two models). The first model is termed *LSTM* and refers to our baseline model, while the second model is termed *Graphlet*, which takes structural information of the neighboring stations into account. Both models are described in details in the following two subsections.

3.2.1. LSTM (Baseline)

The first model can be seen as a baseline and refers to a vanilla LSTM model [16], which takes air temperature as its sole input (see Figure 2a). The model consists of a single LSTM layer, followed by a fully connected layer that maps the final LSTM output to the predicted water temperature $\hat{w}^{(t)}$. Formally, given the input time series $(a^{(1)}, a^{(2)}, \ldots, a^{(t)})$, the LSTM produces the latent representation $\boldsymbol{h}^{(t)}$, which is transformed as

$$\hat{w}^{(t)} = \boldsymbol{W}\boldsymbol{h}^{(t)} + \boldsymbol{b},$$



(a) LSTM (Baseline) (b) Graphlet

Figure 2: Overview of the station-specific models with and without structural information.

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weights and bias of the output layer. The network is trained by minimizing the *Mean Squared Error* (MSE) between the predicted and measured water temperatures.

The architecture is parameterized by the width of the latent space and the depth of stacked LSTM layers, and a separate model f_k is trained independently for each water station k.

3.2.2. Graphlet

This model extends the baseline LSTM model (described in Section 3.2.1) by incorporating spatial structure while remaining station-specific. For each target station k, a graphlet is defined as its 1-hop neighborhood $\mathcal{N}(k)$, consisting of stations directly connected to k (see Figure 2b). Then, the modeling process proceeds in three steps:

1. For each neighboring station $j \in \mathcal{N}(k)$, the baseline model f_j predicts the water temperature $\hat{w}_j^{(t)}$ at time t given the corresponding air temperature time series $(a_j^{(1)}, \ldots, a_j^{(t)})$.
2. The predicted series of temperatures $\{\hat{w}_j^{(1)}, \ldots, \hat{w}_j^{(t)}\}_{j \in \mathcal{N}(k)}$ of each neighboring station, together with the air temperature time series at the target station k, are used as inputs to the station-specific LSTM g_k, which pro-

duces the latent representation $\boldsymbol{h}_k^{(t)}$:

$$\boldsymbol{h}_k^{(t)} = g_k\left(a_k^{(1)}, \ldots, a_k^{(t)}, \{\hat{w}_j^{(1)}, \ldots, \hat{w}_j^{(t)}\}_{j \in \mathcal{N}(k)}\right),$$

where the LSTM g_k is parameterized by width and depth.

3. The latent representation $\boldsymbol{h}_k^{(t)}$ is then transformed by a fully connected output layer to obtain the predicted water temperature $\hat{w}_k^{(t)}$ at the target station k

$$\hat{w}_k^{(t)} = \boldsymbol{W}_k \boldsymbol{h}_k^{(t)} + \boldsymbol{b}_k,$$

where $\boldsymbol{W}_k$ and $\boldsymbol{b}_k$ are the weights and bias of the output layer for station k.

This three-step process is independently repeated for each station k in the network. During training only the additional LSTM g_k and the linear layer are adjusted by minimizing the MSE between the predicted and true water temperatures.

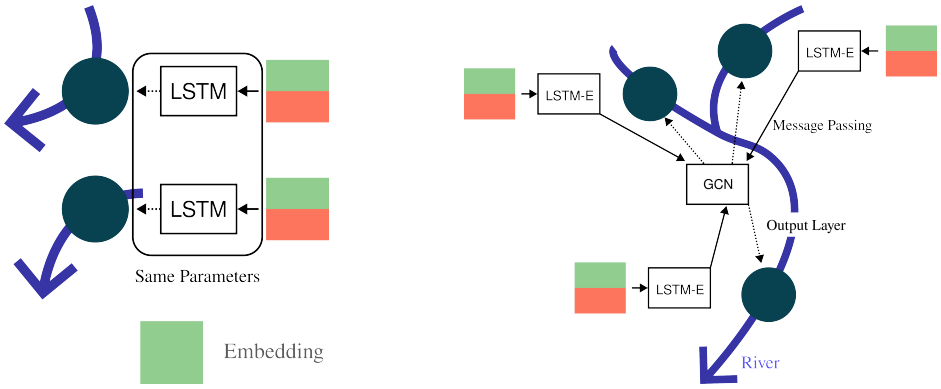
3.3. Embedding-based Models

In our preliminary work [13, 14], we propose and employ two embedding-based models that use a single model $f(\cdot, \mathbf{e}_k)$ for all stations by encoding station-specific characteristics within an embedding space (see Figure 3 for an overview of the two models). The first model is termed *LSTM-E* and extends a standard LSTM with learnable embeddings, while the second model is termed *ST-GNN*, which integrates embeddings and structural information. Both models are described in details in the following two subsections.

3.3.1. LSTM-E

The LSTM-E model (with E denoting *Embedding*), illustrated in Figure 3a, extends a vanilla LSTM by concatenating a trainable, station-specific embedding $\mathbf{e}_k \in \mathbb{R}^n$ to the input at each time step t for station k. Formally, the input to the LSTM at time t is

$$\boldsymbol{x}_k^{(t)} := a_k^{(t)} \,||\, \mathbf{e}_k,$$



(a) LSTM-E (b) ST-GNN

Figure 3: Overview of the embedding-based models.

where $a_k^{(t)} \in \mathbb{R}$ denotes the air temperature at time t for station k, and $||$ denotes concatenation. Due to the concatenation of the embedding $\mathbf{e}_k$ with the air temperature, the embedding can influence all LSTM gates with substantial control. For instance, the input gate $\boldsymbol{i}_k^{(t)}$ is computed as

$$\boldsymbol{i}_k^{(t)} = \sigma\left(\boldsymbol{W}_i a_k^{(t)} + \boldsymbol{V}_i \mathbf{e}_k + \boldsymbol{U}_i \boldsymbol{h}^{(t-1)} + \boldsymbol{b}_i\right),$$

where σ denotes the sigmoid activation function, and $\boldsymbol{W}_i$, $\boldsymbol{V}_i$, $\boldsymbol{U}_i$, and $\boldsymbol{b}_i$ are the trainable weight matrices and bias vector for the input gate, shared across all stations. The computations for the forget, output, and candidate gates follow the same pattern [13] and are used to determine $\boldsymbol{h}_k^{(t)}$.

In a last step, the final latent representation $\boldsymbol{h}_k^{(t)}$ is mapped to the predicted water temperature $\hat{w}_k^{(t)}$ by a linear output layer

$$\hat{w}_k^{(t)} = \boldsymbol{W}\boldsymbol{h}_k^{(t)} + \boldsymbol{b},$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the output layer parameters, again, shared across all stations.

During training, the embeddings $\mathbf{e}_k$ and all network parameters $\boldsymbol{\theta}$ are optimized jointly by minimizing the MSE loss over all stations by computing the

respective gradients $\nabla_{\mathbf{e}_k}\mathcal{L}$ and $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ with

$$\mathcal{L} = \frac{1}{NT}\sum_{k=1}^{N}\sum_{t=1}^{T}\left(\hat{w}_k^{(t)} - w_k^{(t)}\right)^2, \tag{2}$$

where $w_k^{(t)}$ denotes the measured water temperature for station k at time t, while N and T represent the total number of stations and the number of time steps in the prediction window, respectively.

The only station-specific part is the embedding vector $\mathbf{e}_k$. Further, the LSTM-E is parameterized by its latent space width, depth (number of stacked LSTM layers), and the embedding dimension n. Note that this model omits explicit spatial structure and serves as an embedding-only baseline.

3.3.2. ST-GNN

The ST-GNN model, illustrated in Figure 3b, integrates both embedding-based representations and the physical structure of the river network through a spatio-temporal graph neural network [14]. This design enables the use of structure in a single global model. Let $G = (V, E)$ denote the river network where each station k is modeled as vertex $v_k \in V$ and edges $(v_i, v_j) \in E$ modeling connectivity of stations i and j, then the model proceeds in three stages:

1. **Temporal Processing**. The temporal stage consists of an LSTM-E model (excluding its output layer) as detailed in the previous section. Formally, for each station modeled as node $v_k \in V$, the model computes a latent representation at time t as

$$\boldsymbol{h}_k^{(t)} = \text{LSTM-E}(a_k^{(1)}, \ldots, a_k^{(t)}, \mathbf{e}_k).$$

   The collection of all latent representations at time t is denoted by

$$H^{(t)} := \{\boldsymbol{h}_k^{(t)}\}_{v_k \in V}.$$

2. **Spatial Processing**. The latent representations $H^{(t)}$ serve as node fea-

tures in a *Graph Neural Network* (GNN), Formally, the spatial processing is given by

$$\hat{H}^{(t)} = \mathrm{GNN}(H^{(t)}, E),$$

where $\hat{H}^{(t)}$ denotes the updated node feature matrix after message passing along the edges E.

3. **Output Layer**. The output layer transforms the latent representation $\hat{\boldsymbol{h}}_{\boldsymbol{k}}^{(\boldsymbol{t})} \in \hat{H}^{(t)}$ to the predicted water temperatures:

$$\hat{w}_k^{(t)} = \boldsymbol{W}\hat{\boldsymbol{h}}_{\boldsymbol{k}}^{(\boldsymbol{t})} + \boldsymbol{b}$$

During training, all the trainable parameter of the LSTM-E, GNN, and output layer are jointly optimized with the embeddings $\mathbf{e}_k$ with respect of the MSE Loss as defined in Equation (2). The architecture is parameterized by the width of the latent space, the depth of LSTM Layers, the embedding dimension, the type of the GNN (namely, GCN [24], GIN [25], or GraphSAGE [26]), and the amount of GNN layers.

4. Datasets

Properly curated datasets are essential for consistent and meaningful model evaluation, as station-specific characteristics – such as differing sensitivity to air temperature – can introduce substantial variability. Evaluation is further complicated by the heterogeneity of prediction complexity across stations, which can strongly affect aggregate performance metrics. In addition, interannual variability in atmospheric conditions can further bias the results. The latter is particularly problematic when anomalous years fall within the test set, as they may distort assessments of model accuracy and generalization.

To address these challenges, the present paper places a strong emphasis on fairness and reproducibility by providing one of the first systematic evaluations in river water temperature modeling with explicitly defined training, validation and testing periods, transparent and deliberate station selection, and carefully

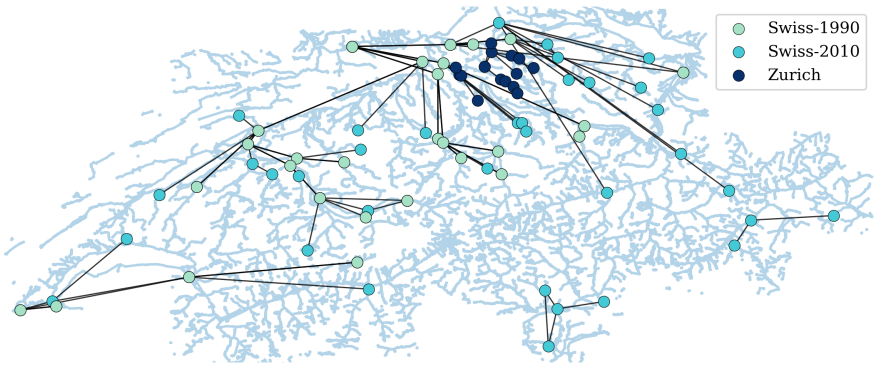


Figure 4: The three curated datasets in front of the Swiss River Network. The stations of Swiss-1990 are a subset of Swiss-2010, and Zurich is located in the north east of Switzerland.

Dataset	Stations	Training	Testing	Years	km²/Station
Swiss-1990	28	1990-2012	2013-2020	30	1596 km^2
Swiss-2010	63	2010-2017	2018-2020	10	1039 km^2
Zurich	15	2009-2019	2020-2022	13	74 km^2

Table 3: Summary of the curated datasets used for model evaluation. Swiss-1990 includes the longest time series, Swiss-2010 covers the largest number of stations, and Zurich offers the highest spatial resolution. All datasets span the full calendar year, from January 1 to December 31.

curated input data. In particular, input features contain no missing values, and while some temperature measurements are unavailable, these are appropriately handled through loss masking during training. The test sets – used for final evaluation only – consist of complete calendar years and are strictly separated from the training data, ensuring robust and interpretable performance comparisons across models.

In particular, three datasets are curated (Table 3 summarizes the major characteristics of the three datasets):

- **Swiss-1990** features the longest temporal series with 30 years in total.
- **Swiss-2010** includes a larger number of stations than Swiss-1990 (namely 63 rather than 28 stations) but shorter measurement periods (ten years in total).

- **Zurich** represents a cantonal river network with a much higher spatial resolution of stations (viz. only $74\,\mathrm{km}^2$ per station rather than more than $1000\,\mathrm{km}^2$ per station).

Figure 4 illustrates the three curated datasets overlaid on the Swiss River Network. The Swiss-1990 stations form a subset of Swiss-2010, while the Zurich dataset covers a high-resolution network in the northeast of Switzerland. Note that the first two datasets have been substantially improved compared to their use in earlier preliminary work [11] (with more thorough curation, cleaning, and preprocessing to ensure consistency and reliability). Furthermore, this work introduces and analyzes the Zurich dataset for the first time – a novel, high-resolution dataset that captures fine-grained hydrological dynamics (the level of spatial detail of this dataset is particularly valuable for advancing the understanding of groundwater inflows and their impact on river water temperature).

To support reproducibility and ease of use, all datasets used in this study are publicly released along with data-loading utilities, enabling seamless integration into machine learning pipelines[6].

5. Experimental Evaluation

The three curated datasets provide a rigorous foundation for evaluating the comparative performance of the proposed models across diverse spatial, temporal, and hydrological contexts. Rather than focusing solely on benchmarking, we also examine how different model architectures complement each other in handling diverse station characteristics and river regimes.

5.1. Evaluation Metrics

To assess predictive performance, we employ the following three evaluation metrics.

[6] `https://swiss-river-network.github.io/data`

1. We measure the *Root Mean Squared Error* (RMSE) formally defined as

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(w^{(t)} - \hat{w}^{(t)})^2} \tag{3}$$

where w is the measured water temperature and $\hat{w}$ is the model's prediction. The RMSE provides a quantitative assessment of the model's prediction error, capturing the root mean squared error across all evaluated time steps. It serves as a key indicator of model accuracy, with lower values indicating better predictive performance.

2. In contrast to the RMSE, the *Mean Absolute Error* (MAE), measures the average magnitude of the prediction errors without squaring them, making it less sensitive to outliers. As such, MAE provides a more balanced view of overall model accuracy when large deviations are not disproportionately penalized. Formally, the MAE is defined as follows:

$$\frac{1}{T}\sum_{t=1}^{T}|w^{(t)} - \hat{w}^{(t)}| \tag{4}$$

3. Lastly, the *Nash–Sutcliffe Model Efficiency Coefficient* (NSE) evaluates the predictive skill of a model by comparing its performance to the mean of the observed data. Values closer to 1 indicate high predictive accuracy, while values less than zero suggest that the model performs worse than simply using the mean as a predictor. Formally, the NSE is defined as follows:

$$1 - \frac{\sum_{t=1}^{T}(w^{(t)} - \hat{w}^{(t)})^2}{\sum_{t=1}^{T}(w^{(t)} - \bar{w})^2}, \quad \bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)} \tag{5}$$

5.2. Training and Hyperparameter Tuning

We use the Adam optimizer as gradient descent based optimizer [27]. Hyperparameter selection is performed via random search using the ASHA scheduler [28]. For the validation set we use an 80%/20% split of the training data and base the model selection on the lowest validation RMSE. Table 4 sum-

Hyperparameter	**Values**	LSTM	Graphlet	LSTM-E	ST-GNN
Epochs[1]	[1, 30]	✓	✓	✓	✓
Learning Rate	[0.001, 0.01]	✓	✓	✓	✓
Batch Size	[32, 256]	✓	✓	✓	✓
Width (latent space)	[16, 128]	✓	✓	✓	✓
Depth (stacked LSTMs)	[1, 3]	✓	✓	✓	✓
Embedding Dimension	[1, 30]			✓	✓
GNN Layers	[1, 7]				✓
GNN Type	{GCN, GIN, GraphSAGE}				✓

Table 4: Random search ranges and options used for model selection. Not all models utilize every listed hyperparameter (✓ indicates the hyperparameter is used by the model). [1]We use early stopping.

marizes the specific ranges and options of the hyperparameters explored. The embedding spaces are learned independently for each dataset.

5.3. Results

Table 5 presents the median, minimum and maximum RMSE values for the four methods across the three datasets. According to the median RMSE, the three novel models are better (or at least as good) as the LSTM baseline method on both Swiss datasets. On the last dataset, Zurich, only Graphlet outperforms the baseline model. The improvements of the Graphlet model on all datasets and the improvement of LSTM-E on the Swiss-2010 dataset are statistically significant (according to the Wilcoxon signed rank test with $p < 0.05$).

A very similar picture emerges when looking at the minimum values of the observed RMSE. Here, the Graphlet method achieves the best result on all three datasets and outperforms the baseline model in every case. According to the maximum value, one of the three models (Graphlet, LSTM-E, ST-GNN) achieves the best result on each of the three datasets. It is interesting to note that on the Zurich dataset, the two embedding-based models cannot beat the baseline model according to the median, but still show better worst-case performance (impacting the NSE metric as will be seen below). Overall, we observe that Graphlet achieves the best result six times, ST-GNN twice, and LSTM-E

Dataset	Method	RMSE Median ↓	RMSE Min ↓	RMSE Max ↓
Swiss-1990	LSTM (Baseline)	.799	.402	1.428
	Graphlet*	**.740**	**.401**	**1.357**
	LSTM-E	.799	.441	1.359
	ST-GNN	.758	.457	1.391
Swiss-2010	LSTM (Baseline)	.792	.380	1.716
	Graphlet*	.764	**.346**	1.718
	LSTM-E*	.769	.395	1.680
	ST-GNN*	**.760**	.373	**1.615**
Zurich	LSTM (Baseline)	.712	.611	1.164
	Graphlet*	**.672**	**.589**	1.161
	LSTM-E	.722	.606	**.874**
	ST-GNN	.740	.602	.921

Table 5: The median, minimal, and maximal RMSE per method and dataset. Bold indicates the best result, underlining indicates the second-best. *Significant improvement over the baseline model by Wilcoxon signed rank test ($p < 0.05$).

once.

Table 6 shows the medians of the MAE and NSE metrics, corroborating the patterns observed in Table 5. This means that, in general, the proposed models are better than the LSTM baseline – On the Swiss-1990 and Swiss-2010 datasets, Graphlet and ST-GNN achieve the highest performance across both metrics. For the Zurich dataset, Graphlet attains the lowest MAE, while LSTM-E achieves the highest NSE.

Based on Tables 5 and 6, we conclude that the embedding-based models (LSTM-E and ST-GNN) achieve predictive performance comparable to their station-specific counterparts (LSTM and Graphlet). Notably, both models substantially reduce the number of learnable parameters. Figure 5 shows the number of learnable parameters for each dataset and method as bar plots on a logarithmic scale. LSTM-E and ST-GNN require two to three orders of magnitude fewer parameters – including embeddings – compared to the station-specific models. This reduction in model complexity is highly beneficial, as it leads to lower computational costs and faster training times.

Due to the diverse landscape and anthropogenic influences across the river

Dataset	Method	MAE Median ↓	NSE Median ↑
Swiss-1990	LSTM (Baseline)	.610	.964
	Graphlet	**.559**	**.966**
	LSTM-E	.595	.960
	ST-GNN	.583	.963
Swiss-2010	LSTM (Baseline)	.612	.968
	Graphlet	.602	.971
	LSTM-E	.609	.968
	ST-GNN	**.592**	**.972**
Zurich	LSTM (Baseline)	.590	.978
	Graphlet	**.527**	.979
	LSTM-E	.576	**.982**
	ST-GNN	.592	.980

Table 6: The medians of the *Mean Absolute Error* (MAE) and *Nash–Sutcliffe Efficiency* (NSE) metrics for each method and dataset. Bold indicates the best result, underlining indicates the second-best.

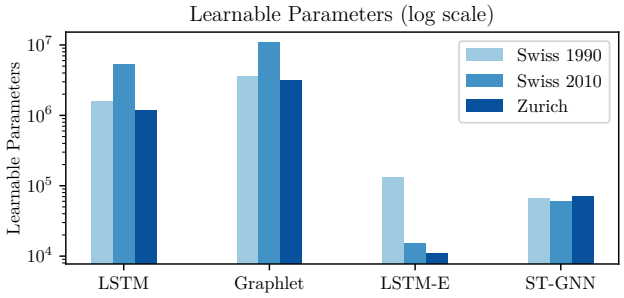
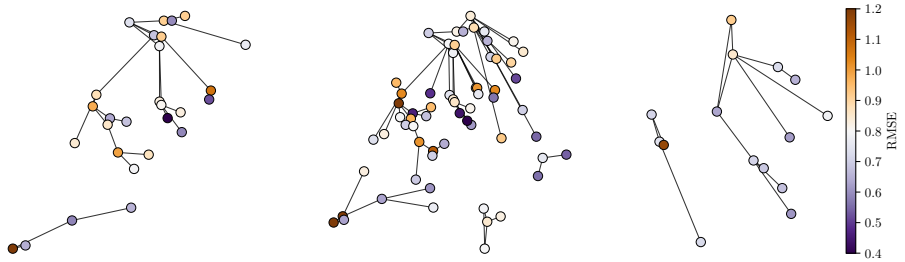


Figure 5: Number of learnable parameters (log scale). The LSTM-E and ST-GNN models use two to three orders of magnitude fewer learnable parameters (including embeddings) compared to the station-specific models.

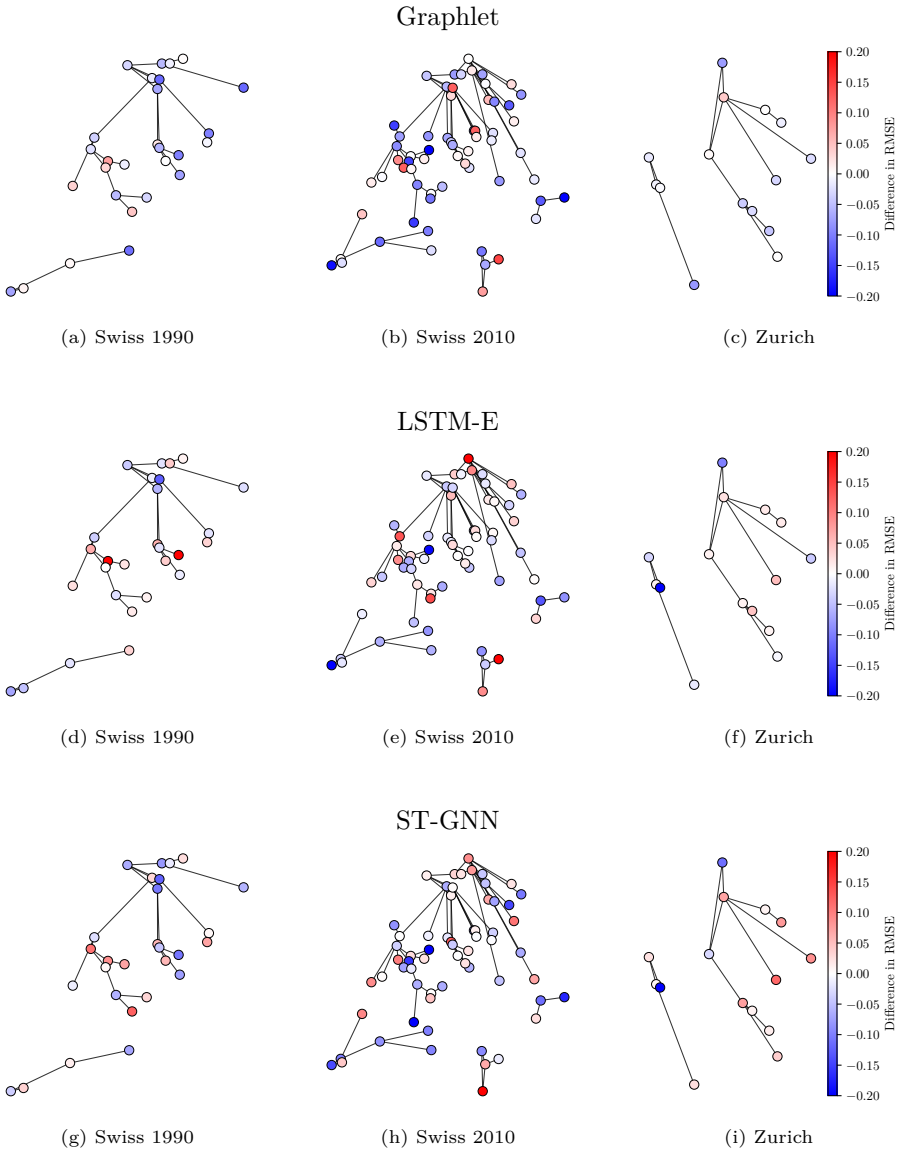


(a) Swiss 1990 (b) Swiss 2010 (c) Zurich

Figure 6: Absolute RMSE results of the LSTM baseline. Lower is better. There is a high variability of the prediction accuracy based on air temperature among the different nodes.

network, the accuracy of predictions varies considerably between stations. Air temperature alone is often insufficient to capture these differences, meaning some stations are inherently harder to model than others. This phenomenon is illustrated in Figure 6 using the baseline LSTM model, showing absolute RMSE values for individual water stations (shown as nodes in the graph). The high variability of the RMSE across stations is clearly visible on all datasets, underscoring the importance of station-level analysis when assessing model performance.

Building on this observation, Figure 7 compares the three proposed models with the LSTM baseline at the station level. Improvements compared to the baseline model are shown in blue, and deteriorations in red. The stronger the color intensity, the greater the deviation from the baseline result. The Graphlet model demonstrates consistently improved performance across most stations across all datasets. The LSTM-E model performs comparably to the baseline, and the ST-GNN further improves upon the performance of the LSTM-E model. The advantage of station-specific models is most apparent at "outlier" stations, exhibiting atypical behavior, often located in alpine source regions, near power plants, or downstream of lakes. In the difference maps between LSTM-E or ST-GNN and the baseline (second and third rows of Figure 7), these stations appear as dark red regions, indicating pronounced performance degradation relative to the baseline.



(a) Swiss 1990 (b) Swiss 2010 (c) Zurich

(d) Swiss 1990 (e) Swiss 2010 (f) Zurich

(g) Swiss 1990 (h) Swiss 2010 (i) Zurich

Figure 7: The per-node RMSE differences between the evaluated models and the baseline. Lower values in blue indicate better accuracy.

5.4. Relation to Prior Work

While the general performance trends reported in prior work could be reproduced, the absolute results obtained here are consistently lower than previously claimed [14]. This discrepancy likely reflects differences in evaluation rigor, suggesting that earlier studies may have benefited from selective reporting or insufficiently controlled validation protocols. These findings highlight the critical need for transparent benchmarking, reproducible workflows, and unambiguous evaluation settings in river water temperature modeling. Furthermore, an exclusive focus on mean performance metrics can obscure substantial variability across stations and hydrological conditions, potentially leading to overestimation of model robustness. The curated datasets and rigorous evaluation framework introduced in this work directly address these issues, establishing a reliable foundation for future model development and comparison.

6. Conclusion and Future Work

Rising river water temperatures, driven by both climate change and anthropogenic activities, pose a growing threat to freshwater ecosystems, water quality, and energy security. This warming trend, now consistently exceeding ecological and regulatory thresholds in Switzerland, underscores the urgent need for accurate, adaptable predictive pattern recognition models capable of capturing complex spatio-temporal dynamics.

This work investigates the role of embeddings and structural information in the geospatial time series task of water temperature modeling. We introduce three curated datasets with clearly defined training and test periods to ensure fair and reproducible evaluation. Notably, the Zurich dataset offers a novel contribution through its high spatial resolution, enabling fine-grained analysis of river network dynamics.

Our comparative study includes two station-specific models (LSTM and Graphlet) as well as two embedding-based models (LSTM-E and ST-GNN). Station-specific models demonstrate a clear advantage when applied to "out-

lier" stations – those exhibiting unique or atypical hydrological behavior. These stations often deviate from broader trends and benefit from specialized modeling that can capture localized patterns more effectively. In contrast, embedding-based methods prove highly efficient across the full dataset. They not only reduce the number of learnable parameters by magnitudes but also offer strong generalization capabilities. The ability to train a single model simplifies both the training pipeline and the deployment process, making these approaches preferable for large-scale applications. In both modeling paradigms, incorporating structural information – specifically, the natural flow of the river network – improves predictive performance in general. In practice, the Graphlet model often outperforms the baseline LSTM, while the ST-GNN generally achieves better results than the embedding-only LSTM-E.

For future research, we see several rewarding avenues that could be pursued. The promising results of embedding-based models suggest that embeddings might play a central role in the next generation of water temperature modeling. This naturally raises the question of transformer-based architectures, which have shown considerable success in various domains [29, 30]. Furthermore, prior work has demonstrated the effectiveness of LSTM-based transformer architectures on water temperature modeling [31].

Despite the overall strength of embedding-based approaches, there remains a performance gap between the Graphlet and the ST-GNN model, particularly on "outlier" stations. This discrepancy suggests that the current embedding space may lack the flexibility required to fully capture the diverse dynamics present in such stations. Addressing this gap presents a compelling research challenge, inviting the exploration of more expressive or adaptive architectures – such as enhanced attention mechanisms or hybrid graph-transformer models [32] – to better accommodate to atypical conditions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Supported by Swiss National Science Foundation (SNSF) Grant Nr. PT00P2_206252. Data are kindly provided by the *Federal Office for the Environment*, *MeteoSwiss*, and *Canton of Zurich*.

References

[1] S. Piccolroaz, E. Calamita, B. Majone, A. Gallice, A. Siviglia, M. Toffolon, Prediction of river water temperature: a comparison between a new family of hybrid models and statistical approaches, Hydrological Processes 30 (21) (2016) 3901–3917.

[2] D. Caissie, The thermal regime of rivers: a review, Freshwater biology 51 (8) (2006) 1389–1406.

[3] P. C. Reid, R. E. Hari, G. Beaugrand, D. M. Livingstone, C. Marty, D. Straile, J. Barichivich, E. Goberville, R. Adrian, Y. Aono, et al., Global impacts of the 1980s regime shift, Global change biology 22 (2) (2016) 682–703.

[4] A. Michel, T. Brauchli, M. Lehning, B. Schaefli, H. Huwald, Stream temperature and discharge evolution in switzerland over the last 50 years: annual and seasonal behaviour, Hydrology and Earth System Sciences 24 (1) (2020) 115–142.

[5] F. T. Dahlke, S. Wohlrab, M. Butzin, H.-O. Pörtner, Thermal bottlenecks in the life cycle define climate vulnerability of fish, Science 369 (6499) (2020) 65–70.

[6] A. Karvonen, P. Rintamäki, J. Jokela, E. T. Valtonen, Increasing water temperature and disease risks in aquatic systems: climate change increases the risk of some, but not all, diseases, International journal for parasitology 40 (13) (2010) 1483–1488.

[7] M. F. Johnson, L. K. Albertson, A. C. Algar, S. J. Dugdale, P. Edwards, J. England, C. Gibbins, S. Kazama, D. Komori, A. D. MacColl, et al., Rising water temperature in rivers: Ecological impacts and future resilience, Wiley Interdisciplinary Reviews: Water 11 (4) (2024) e1724.

[8] K. Matthews, N. Berg, Rainbow trout responses to water temperature and dissolved oxygen stress in two southern california stream pools, Journal of Fish Biology 50 (1) (1997) 50–67.

[9] CH2018, Climate scenarios for switzerland, Tech. rep., National Centre for Climate Services, Zurich (2018).

[10] L. Råman Vinnå, V. Bigler, O. S. Schilling, J. Epting, Multi-fidelity model assessment of climate change impacts on river water temperatures, thermal extremes and potential effects on cold water fish in switzerland, EGUsphere 2025 (2025) 1–44.

[11] B. Fankhauser, V. Bigler, K. Riesen, Graph-based deep learning on the swiss river network, in: International Workshop on Graph-Based Representations in Pattern Recognition, Springer, 2023, pp. 172–181.

[12] R. I. Woolway, M. T. Dokulil, W. Marszelewski, M. Schmid, D. Bouffard, C. J. Merchant, Warming of central european lakes and their response to the 1980s climate regime shift, Climatic Change 142 (2017) 505–520.

[13] B. Fankhauser, V. Bigler, K. Riesen, Leveraging lstm embeddings for river water temperature modeling, in: IAPR Workshop on Artificial Neural Networks in Pattern Recognition, Springer, 2024, pp. 283–294.

[14] B. Fankhauser, V. Bigler, K. Riesen, Spatio-temporal graph neural networks for water temperature modeling, in: Structural, Syntactic, and Sta-

tistical Pattern Recognition, Springer Nature Switzerland, Cham, 2025, pp. 31–40.

[15] M. Toffolon, S. Piccolroaz, A hybrid model for river water temperature as a function of air temperature and discharge, Environmental Research Letters 10 (11) (2015) 114011.

[16] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[17] P. J. Werbos, Backpropagation through time: what it does and how to do it, Proceedings of the IEEE 78 (10) (2002) 1550–1560.

[18] R. Qiu, Y. Wang, B. Rhoads, D. Wang, W. Qiu, Y. Tao, J. Wu, River water temperature forecasting using a deep learning method, Journal of Hydrology 595 (2021) 126016.

[19] X. Jia, J. Zwart, J. Sadler, A. Appling, S. Oliver, S. Markstrom, J. Willard, S. Xu, M. Steinbach, J. Read, et al., Physics-guided recurrent graph model for predicting flow and temperature in river networks, in: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), SIAM, 2021, pp. 612–620.

[20] T. Bao, X. Jia, J. Zwart, J. Sadler, A. Appling, S. Oliver, T. T. Johnson, Partial differential equation driven dynamic graph networks for predicting stream water temperature, in: 2021 IEEE International Conference on Data Mining (ICDM), IEEE, 2021, pp. 11–20.

[21] Z. Moshe, A. Metzger, G. Elidan, F. Kratzert, S. Nevo, R. El-Yaniv, Hydronets: Leveraging river structure for hydrologic modeling, arXiv preprint arXiv:2007.00595 (2020).

[22] D. Kim, H. Han, W. Wang, H. S. Kim, Improvement of deep learning models for river water level prediction using complex network method, Water 14 (3) (2022) 466.

[23] Q. Zhao, Y. Zhu, K. Shu, D. Wan, Y. Yu, X. Zhou, H. Liu, Joint spatial and temporal modeling for hydrological prediction, Ieee Access 8 (2020) 78492–78503.

[24] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 2017, 2017.

[25] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019.

[26] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Advances in neural information processing systems 30 (2017).

[27] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[28] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-Tzur, M. Hardt, B. Recht, A. Talwalkar, A system for massively parallel hyperparameter tuning, Proceedings of Machine Learning and Systems 2 (2020) 230–246.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[30] B. Lim, S. Ö. Arık, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, International journal of forecasting 37 (4) (2021) 1748–1764.

[31] R. S. Padrón, M. Zappa, L. Bernhard, K. Bogner, Extended range forecasting of stream water temperature with deep learning models, EGUsphere 2024 (2024) 1–27.

[32] H. Chen, P. Jiao, M. Du, X. Guo, Z. Zhao, D. Jin, X. Liu, Tgformer: Towards temporal graph transformer with auto-correlation mechanism, Pattern Recognition 170 (2026) 112053.