Multi-Scale Graph Learning for Anti-Sparse Downscaling

Yingda Fan[1], Runlong Yu[1], Janet R. Barclay[2], Alison P. Appling[3], Yiming Sun[1], Yiqun Xie[4], Xiaowei Jia[1]

[1]Department of Computer Science, University of Pittsburgh
[2]U.S. Geological Survey, New England Water Science Center
[3]U.S. Geological Survey, Water Mission Area
[4]Department of Geographical Sciences, University of Maryland
{yif47, ruy59, yis108, xiaowei}@pitt.edu, {jbarclay, aappling}@usgs.gov, xie@umd.edu

Abstract

Water temperature can vary substantially even across short distances within the same sub-watershed. Accurate prediction of stream water temperature at fine spatial resolutions (i.e., fine scales, $\leq$ 1 km) enables precise interventions to maintain water quality and protect aquatic habitats. Although spatiotemporal models have made substantial progress in spatially coarse time series modeling, challenges persist in predicting at fine spatial scales due to the lack of data at that scale. To address the problem of insufficient fine-scale data, we propose a Multi-Scale Graph Learning (MSGL) method. This method employs a multi-task learning framework where coarse-scale graph learning, bolstered by larger datasets, simultaneously enhances fine-scale graph learning. Although existing multi-scale or multi-resolution methods integrate data from different spatial scales, they often overlook the spatial correspondences across graph structures at various scales. To address this, our MSGL introduces an additional learning task, cross-scale interpolation learning, which leverages the hydrological connectedness of stream locations across coarse- and fine-scale graphs to establish cross-scale connections, thereby enhancing overall model performance. Furthermore, we have broken free from the mindset that multi-scale learning is limited to synchronous training by proposing an Asynchronous Multi-Scale Graph Learning method (ASYNC-MSGL). Extensive experiments demonstrate the state-of-the-art performance of our method for anti-sparse downscaling of daily stream temperatures in the Delaware River Basin, USA, highlighting its potential utility for water resources monitoring and management.

Introduction

Predicting water temperature at fine spatial scale (e.g., for discrete stream reaches of 1 km or shorter) can enhance understanding of and planning for the effects of weather and climate changes on fish habitats, as stream water temperatures can vary greatly along stream reaches due to differences in shade, groundwater inflows, and the presence of deep and shallow sections. Stream temperature downscaling aims to predict water temperatures at a fine spatial scale, using insights partly derived from spatially coarse data (e.g., stream reaches of roughly 10 km long) at the same temporal resolution. In this paper, we use both "scale" and "resolution" to describe the density of reach segments modeled per unit area. Our evaluation focuses on daily predictions for the Delaware River Basin, which supplies drinking water to over 15 million people (Williamson et al. 2015) and requires maintaining cool water temperatures to preserve the habitat of aquatic life (Brett 1971; Ravindranath et al. 2016).

Spatiotemporal graphs have shown utility in modeling stream networks (Moshe et al. 2020; Jia et al. 2021b; Sun et al. 2021; Chen, Zwart, and Jia 2022; Topp et al. 2023; Jia et al. 2023; He et al. 2024; Jia et al. 2021a; Chen et al. 2023) as they concisely capture the interactions amongst different stream segments. However, existing graph-based stream models are focused on simulating water temperature at a coarse resolution due to data limitations at finer resolutions. Specifically, collecting water temperature observations requires costly field visits for sensor installation and maintenance, thus limiting the spatial resolution of observations. As a result, existing water temperature datasets in most places can only support coarse-resolution models. These models fail to reflect the local variations in water temperature at a fine resolution (Terry et al. 2022) and thus remain limited in informing local management decisions.

To work around sparsity in fine-scale observations, various techniques leverage coarse-resolution data to inform fine-scale predictions. Multi-resolution learning models, particularly those employing Gaussian processes, effectively integrate data across different scales (Yousefi, Smith, and Álvarez 2019; Hamelijnck et al. 2019). Bayesian deconditioning refines low-resolution spatial fields by incorporating high-resolution information (Chau, Bouabid, and Sejdinovic 2021). Techniques like Generative Adversarial Networks (GANs) and quantile perturbation have been employed for meteorological downscaling (Chaudhuri and Robertson 2020; Tabari et al. 2021). Further, efficient algorithms have been developed to enhance learning accuracy from imprecise, coarse-labeled data (Fotakis 2021), and multiple instance learning has been extended to fine-scale predictions from aggregate observations (Zhang et al. 2020). The combination of consistent supervision and residual feature augmentation has shown promise in advancing super-resolution prediction tasks (Qin, Huang, and Wen 2020). Additionally, multi-resolution data fusion techniques also improve high-resolution land cover mapping (Robinson et al. 2019).

Although existing methods, such as multi-resolution learning, provide valuable frameworks for downscaling and fine-scale prediction, they are primarily designed for gridded data, such as images, and are less suitable for irregularly structured data, such as graphs. Hence, these approaches are not well-adapted to stream temperature prediction models that rely on spatiotemporal graph networks. In contrast, our approach, termed Multi-Scale Graph Learning (MSGL), emphasizes the topological logic inherent in graph networks.

In graph networks, graphs at different scales represent the same physical entities (e.g., stream reaches). We can thus use a cross-scale distance matrix to describe the physical distances and positional relationships between pairs of stream reaches at coarse and fine scales. The proposed MSGL method aims to adaptively learn the dynamic behaviors of nodes and their interactions across different scales.

In detail, MSGL processes data simultaneously at different scales by considering three scale learning tasks. **Coarse-scale graph learning (CSL)** aims to extract patterns of water dynamics at the coarse scale using features $\mathbf{X}_c$ and labels $\mathbf{Y}_c$ derived from the coarse-scale dataset $\mathcal{D}_{\text{coarse}}$. **Cross-scale interpolation learning (CrSL)** is designed to capture the relationships between stream segments at different scales. It takes coarse-scale input features $\mathbf{X}_c$ from the dataset $\mathcal{D}_{\text{coarse}}$ and utilizes an interpolation module to generate fine-scale predictions $\hat{\mathbf{Y}}_f$. **Fine-scale graph learning (FSL)** focuses on predicting the target variable at the fine scale. This task integrates multi-scale information by combining fine-scale features $\mathbf{X}_f$ and latent representations derived from the cross-scale interpolation. The three tasks share a subset of parameters to transform the same set of input features into latent representations, which helps capture water dynamics processes influenced by these features. Inspired by existing optimization algorithms (Sener and Koltun 2018; Mahapatra and Rajan 2020), MSGL uses a Multi-Scale Optimization (MSO) algorithm to balance the contributions of three learning tasks in the training process.

Most multi-resolution methods adopt synchronous strategies, such as combining loss functions (Robinson et al. 2019). A few asynchronous methods initiate with pre-training on low-resolution datasets and end in fine-tuning across multiple resolutions (Ji, Zhong, and Ma 2021). Our work differs in fully capitalizing on directional heat transport and spatially correlated heat sources within watersheds (Ward 1985), which ensure that coarse-scale temperatures are predictive of the mean temperature across fine-scale stream reaches in the same area. Thus, predictions from a coarse-scale model can be remapped to the fine scale to provide simulation labels for pre-training in MSGL, to be followed by fine-tuningon observed but sparse fine-scale data. We refer to this approach as Asynchronous Multi-Scale Graph Learning (ASYNC-MSGL).

We conduct extensive evaluations on real-world watersheds by comparing MSGL and ASYNC-MSGL with other spatiotemporal graph baselines that either focus on fine-scale data or multiple scales. Ablation studies further reveal the contribution of different scale-specific tasks to downscaling. The results highlight the method's superior predictive accuracy and robustness, particularly where data are sparse.

Problem Definition and Background

Graph learning for predicting water temperature

Stream networks consist of multiple interconnected stream segments through which water flows from upstream to downstream. Graphs have been widely used to represent complex interactions among multiple segments in stream networks (Moshe et al. 2020; Jia et al. 2021b; Sun et al. 2021; Chen, Zwart, and Jia 2022; Chen et al. 2021; Topp et al. 2023). A stream network can be represented by a graph $\{\mathcal{V}, \mathcal{E}, \mathbf{A}\}$, where the node set $\mathcal{V}$ contains the set of river segments, the edge set $\mathcal{E}$ contains the edges between upstream and downstream river segments, and the adjacency matrix $\mathbf{A}$ measures the connection strength between each pair of nodes. The graph-based methods have shown good performance in predicting water temperature (Jia et al. 2021b; Topp et al. 2023), streamflow (Moshe et al. 2020; Sun et al. 2021), and other water properties (Li et al. 2024).

These methods remain limited in addressing sparse and imbalanced observations over different segments due to the substantial cost needed for measuring water properties (Harmel et al. 2023). Moreover, the stream networks can be created at different spatial scales based on varying standards for splitting segments. When modeling fine-scale networks, the model needs to capture both local and long-distance dependencies over the graph, which remains challenging for most graph neural network models (Di Giovanni et al. 2023).

Definition of anti-sparse downscaling

The objective is to predict the target variable (daily water temperature) $\mathbf{Y}_f$ in different stream segments at fine spatial resolution. Our method is anti-sparse as it maintains reasonable downscaling performance even when the local labeled data are extremely sparse. We have input drivers for N fine-resolution river segments over T dates. We also have an adjacency matrix $\mathbf{A}_f$ containing the adjacency (inverse of stream distance) between each pair of upstream-downstream segments. In addition to the fine-resolution data, we have input drivers $\mathbf{X}_c$, the observed target variable $\mathbf{Y}_c$, and the adjacency matrix $\mathbf{A}_c$ over M ($M < N$) coarse-resolution stream segments that cover the same watershed in less detail. We also use a cross-scale matrix $\mathbf{D} \in \mathbb{R}^{N \times M}$. Each entry $\mathbf{D}_{ij}$ represents the inverse stream distance between each pair of fine and coarse segments.

The input drivers $\mathbf{X}_f$ and $\mathbf{X}_c$ contain the same set of variables at the same temporal resolution (more details in dataset description below), but differ in spatial resolutions. For example, coarse-resolution data in $\mathbf{X}_c$ are estimated as averages over larger polygons than the fine-resolution data in $\mathbf{X}_f$. Both coarse and fine observations ($\mathbf{Y}_c$ and $\mathbf{Y}_f$) are available only for certain segments on certain dates. Because observations are made at points and then aggregated to segments, we typically have a larger number of observations per segment and a greater observation density at the coarse resolution (Appendix B.1). The proposed work aims to leverage the spatiotemporal patterns learned from coarse-resolution data to aid in the prediction of $\mathbf{Y}_f$ at fine resolution.

Recurrent graph network (RGrN)

As the base model for our work, we use RGrN, which couples temporal recurrence with spatial graph convolution. Previous studies have shown utility of this model in predicting stream water temperature and streamflow (Jia et al. 2021b; Topp et al. 2023; Barclay et al. 2023). RGrN extends the standard LSTM structure by incorporating neighborhood information in state evolution. Specifically, it computes states as $\mathbf{s}^t = \mathbf{fg}^t \odot (\mathbf{s}^{t-1} + \mathbf{A} \cdot g(\mathbf{s}^{t-1})) + \mathbf{ig}^t \odot \bar{\mathbf{s}}^t$, where $\mathbf{s}^t$ contains the states of all stream segments at time t, $\mathbf{A}$ is an adjacency matrix used to convolve states among nearby stream segments, $g(\cdot)$ is a transformation implemented by fully connected layers, $\odot$ is the entry-wise product, and $\mathbf{fg}$, $\mathbf{ig}$, and $\bar{\mathbf{s}}^t$ are the forget gate, input gate, and candidate cell state, respectively, as in the standard LSTM (Hochreiter and Schmidhuber 1997). The state $\mathbf{s}^t$ generates the hidden representation $\mathbf{h}^t$ and predicts the target variable.

Multi-objective learning and optimization

Multi-scale learning methods have been proposed to extract and fuse features at different scales (Jiang et al. 2020; Liu et al. 2021; Zhao, Liu, and Wang 2021). These methods have found success in many computer vision tasks, such as fine-grained visual categorization and image classification (He, Li, and Chen 2017; Zhang et al. 2021; Chen, Fan, and Panda 2021; Gu et al. 2022). One major issue in multi-scale learning is balancing and optimizing the learning process across scales. Building upon the proposed multi-scale learning framework, this work also introduces a multi-scale optimization algorithm (to be discussed later) to adjust the influence of the gradient for learning tasks at different scales. The proposed optimization algorithm extends the principles of the Pareto multi-objective optimization algorithm (Deb and Gupta 2005; Désidéri 2012; Sener and Koltun 2018; Mahapatra and Rajan 2020; Suzuki et al. 2020) and the robust multi-objective optimization algorithm (Roberts et al. 2018; Zhou et al. 2018; Daulton et al. 2022).

Multi-Scale Graph Learning

As shown in Figure 1, the proposed framework contains two major modules, the shared spatiotemporal graph embedding module, abbreviated as **Graph-Embed**, and the scale-specific module. The shared graph embedding module converts input drivers at fine or coarse scale into latent representations:

$$\mathbf{H} = \text{Graph-Embed}(\mathbf{X}, \mathbf{A}; \Theta), \tag{1}$$

where $\mathbf{H}$ represents the graph-embedded feature representation, encompassing comprehensive spatial and temporal information. The input $\mathbf{X}$ and $\mathbf{A}$ can be from either the coarse or the fine spatial resolution (i.e., $\{\mathbf{X}_c, \mathbf{A}_c\}$ or $\{\mathbf{X}_f, \mathbf{A}_f\}$), and the parameters Θ are shared across different spatial resolutions. The embedding function Graph-Embed is implemented using the RGrN model (Jia et al. 2021b).

The representation $\mathbf{H}$ is then passed into the scale-specific prediction module, which conducts three predictive learning tasks to create predictions at different spatial scales. These tasks are designed to complement each other while being optimized simultaneously. Next, we describe the three learning tasks in the scale-specific module.

Synchronous multi-scale graph learning

Coarse-scale graph learning. The **CSL** task is designed to capture broad spatial patterns in water temperature dynamics from $\mathcal{D}_{\text{coarse}}$, with the expectation that fine-scale temperatures fluctuate around those coarse-scale patterns.

Given the input drivers $\mathbf{X}_c$ and adjacency matrix $\mathbf{A}_c$ at the coarse scale, we embed the hidden representations $\mathbf{H}_c$ through $\mathbf{H}_c = \text{Graph-Embed}(\mathbf{X}_c, \mathbf{A}_c; \Theta)$. The final output is generated through fully connected layers as $\hat{\mathbf{Y}}_c = \text{Dense}(\mathbf{H}_c; \Phi_c)$, where $\hat{\mathbf{Y}}_c$ represents the predicted coarse-scale temperatures, and $\Phi_c = \{\mathbf{W}_c, \mathbf{b}_c\}$ represents the parameters for the fully connected layers that are specific to the CSL. In comparison to graphs relying solely on the fine-scale data $\mathcal{D}_{\text{fine}}$, the incorporation of CSL can help stabilize fine-scale predictions around coarse-scale trends, resulting in more robust performance.

Cross-scale interpolation learning. The **CrSL** task interpolates the coarse-scale graph embeddings $\mathbf{H}_c$ to generate fine-scale latent representations $\mathbf{H}_f$ that preserve the graph's structure and reflect the fine-scale distribution as

$$\mathbf{H}_{cr} = \mathbf{D} \cdot \mathbf{H}_c, \tag{2}$$

where $\mathbf{D} \in \mathbb{R}^{N \times M}$ is the cross-scale matrix. Each entry $\mathbf{D}_{ij}$ represents the adjacency (inverse of distance) between each pair of fine-scale node i and coarse-scale node j.

Next, the interpolated representation $\mathbf{H}_{cr}$ is processed through an attention mechanism (Vaswani et al. 2017), and then gets concatenated with the original representation $\mathbf{H}_{cr}$, and normalized to produce the final feature representation $\tilde{\mathbf{H}}_{cr}$, as $\tilde{\mathbf{H}}_{cr} = \text{BatchNorm}\left([\mathbf{H}_{cr}, \text{Multihead-Attn}(\mathbf{H}_{cr})]\right)$. The final output is generated through fully connected layers, as $\hat{\mathbf{Y}}_{cr} = \text{Dense}(\tilde{\mathbf{H}}_{cr}; \Phi_{cr})$, where Φ_{cr} represents the parameters for the fully connected layers that are specific to the cross-scale interpolation task. The task output $\hat{\mathbf{Y}}_{cr}$ is produced at the fine scale.

Equation 2 integrates the cross-scale distance matrix $\mathbf{D}$ into the interpolation task within the implicit representation space, encoding cross-scale information by capturing relative node positions at different scales. This, along with an attention mechanism, enhances the model's understanding of intrinsic relationships between these cross-scale nodes, thereby improving overall performance.

Fine-scale graph learning. The **FSL** aims to directly predict high-resolution labels $\mathbf{Y}_f$ using $\mathbf{X}_f$ while also leveraging the representations $\mathbf{H}_{cr}$ from CrSL. Specifically, the high-resolution input $\mathbf{X}_f$ is first embedded through the shared graph embedding module, as $\mathbf{H}_f = \text{Graph-Embed}(\mathbf{X}_f, \mathbf{A}_f; \Theta)$.

Although $\mathbf{H}_f$ has the same dimensions as $\mathbf{H}_{cr}$, it is important to note that $\mathbf{H}_f$ directly encodes the information from the fine-resolution inputs $\mathbf{X}_f$ and $\mathbf{A}_f$.

To further enrich the representation, we combine the information obtained from both fine resolution and coarse

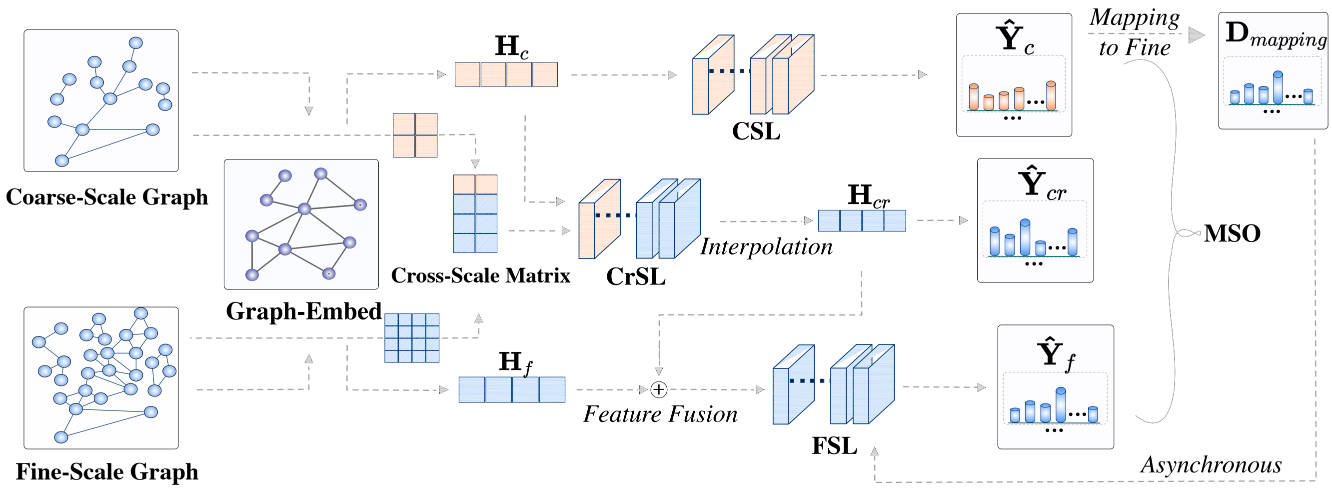

Figure 1: The figure illustrates the architecture of the Multi-Scale Graph Learning (**MSGL**) model, which comprises three parallel training streams: coarse-scale graph learning (**CSL**), fine-scale graph learning (**FSL**), and cross-scale interpolation learning (**CrSL**). The trade-off among these three learning tasks is managed through Multi-Scale Optimization (**MSO**). $\mathbf{H}$ are hidden states. $\hat{\mathbf{Y}}$ are temperature predictions. $\mathbf{c}$, $\mathbf{cr}$, and $\mathbf{f}$ denote the coarse-scale, cross-scale, and fine-scale, respectively. $\mathbf{D}_{mapping}$ is a CSL-only $\hat{\mathbf{Y}}_c$ mapped to fine resolution for use in asynchronous pre-training of MSGL (**ASYNC-MSGL**).

resolution. In particular, concatenating the obtained $\mathbf{H}_f$ with the representation $\mathbf{H}_{cr}$ of the cross-scale task forms a comprehensive feature representation $\mathbf{H}_{f,cr} = [\mathbf{H}_f, \mathbf{H}_{cr}]$. This integration enhances the combination of information from different scales. The combined features are then transformed through a fully connected layer, residual connection, and batch normalization, as $\tilde{\mathbf{H}}_f = \text{BatchNorm}\left(\mathbf{H}_{\text{f,cr}} + \sigma(\mathbf{W}_{\text{res}}\mathbf{H}_{\text{f,cr}} + \mathbf{b}_{\text{res}})\right)$. Then we create the final prediction as $\hat{\mathbf{Y}}_f = \text{Dense}(\tilde{\mathbf{H}}_f; \mathbf{W}_f, \mathbf{b}_f)$, where σ is the non-linear activation function and the parameters $\Phi_f = \{\mathbf{W}_{\text{res}}, \mathbf{b}_{\text{res}}, \mathbf{W}_f, \mathbf{b}_h\}$ are specific to **FSL**.

Multi-scale optimization. Multi-scale optimization (**MSO**) trains the model by iteratively updating the scale-specific parameters Φ_τ ($\tau \in \{c, cr, f\}$) and the shared parameters Θ. We first update Φ_τ using gradient descent.

$$\Phi_\tau = \Phi_\tau - \eta \nabla_{\Phi_\tau} \mathcal{L}_\tau, \tag{3}$$

where η is the learning rate and $\mathcal{L}_\tau$ is the loss function for each task. Both $\mathcal{L}_{cr}$ and $\mathcal{L}_f$ are defined as the mean-squared loss on the fine-scale labels, and $\mathcal{L}_c$ is defined on the coarse-scale labels. The gradient ∇_{Φ_τ} is calculated based on the loss $\mathcal{L}_\tau$ with respect to the parameters Φ_τ.

We next update the shared parameters Θ using information from both scales. We denote by $\nabla_\Theta \mathcal{L}_\tau$, $\tau \in \{c, cr, f\}$, the gradient of the loss function of each scale-specific task with respect to the shared parameters Θ. To optimize multi-scale learning tasks simultaneously without harming any particular task, we utilize the multiple-gradient descent algorithm (MGDA, Désidéri 2012) to adjust the weight of loss gradient for each learning task, as $\{\alpha_\tau\}$, for $\tau \in \{c, cr, f\}$. Finally, we update the shared parameters Θ using the α_τ-weighted sum of gradients, as

$$\Theta = \Theta - \eta \sum_\tau \alpha_\tau \nabla_\Theta \mathcal{L}_\tau. \tag{4}$$

Asynchronous multi-scale graph learning

CSL training, prediction and fine-scale mapping. We introduce an asynchronous MSGL (**ASYNC-MSGL**) method that begins with a standalone coarse-scale graph learning phase. Initially, the model is trained on abundant coarse-scale data collected over large regions using **CSL**. The trained model then generates predictions on the coarse-scale data, which are mapped to the fine scale, creating the labels for the fine-scale data in $\mathcal{D}_{\text{mapping}}$.

To ensure consistency and continuity throughout the multi-task learning process during the $\mathcal{D}_{\text{mapping}}$-based pre-training, the coarse-scale labels in $\mathcal{D}_{\text{mapping}}$ are not the original ground truth labels, but rather the predictions from the trained CSL model. The fine-scale labels are derived by remapping those CSL predictions from the coarse-scale reaches to the spatially corresponding fine-scale reaches.

MSGL pre-training on $\mathcal{D}_{\text{mapping}}$ and fine-tuning on $\mathcal{D}_{\text{groundtruth}}$ dataset. Once the labels are generated, we proceed to pre-training on $\mathcal{D}_{\text{mapping}}$, where the model learns from both coarse-scale and fine-scale inputs. Unlike the fixed periods in the ground truth dataset, $\mathcal{D}_{\text{mapping}}$ covers a broader temporal range as its labels are derived from CSL predictions, not limited by observation data. This extended temporal coverage allows the model to capture longer-term dynamics and improve temporal continuity in predictions, enhancing its robustness against sparse data.

After pre-training, the model is fine-tuned on the $\mathcal{D}_{\text{groundtruth}}$ dataset, which contains the original ground truth labels. This two-stage process—pre-training on simulated data followed by fine-tuning on real data—helps the model align with the true data distribution while maintaining the benefits of the broader scale learning.

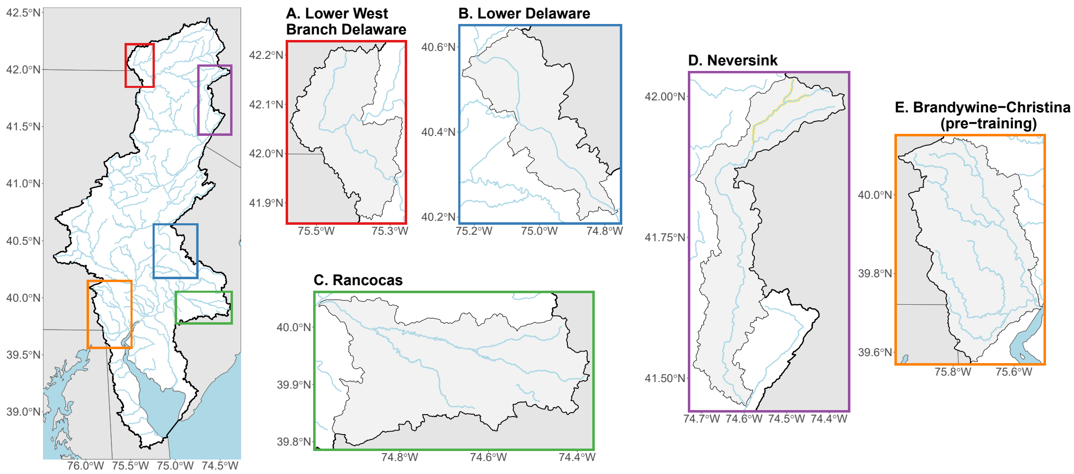

Figure 2: Map of five well-observed watersheds in the Delaware River Basin in the eastern United States.

Experimental Results

Dataset. Figure 2 illustrates the study area, encompassing the Brandywine, Lower Delaware, Neversink, Lower West Branch Delaware, and Rancocas watersheds within the Delaware River Basin, eastern United States. The data and predictions are daily temporal resolution, covering the years 1979-2021. Data were collected at both high and low spatial resolutions (Oliver et al. 2022; Terry et al. 2022; Abatzoglou 2013). Specifically, the coarse-scale river segments were defined by the geospatial fabric used for the National Hydrologic Model (NHM) (Regan et al. 2018) and have an average segment length of 10.5 km in the Delaware River Basin. In contrast, the fine-scale segments were defined by the National Hydrography Dataset (NHD) (U.S. Geological Survey 2019) and have an average segment length of 1.3 km. The input features include slope, elevation, width, daily average air temperature, solar radiation, precipitation, and potential evapotranspiration.

Although the aforementioned watersheds have well-sampled fine-scale and coarse-scale data, fine-scale labels are very limited in most other watersheds. To more accurately evaluate our method in real-world downscaling scenarios, we conduct experiments using training and validation sets with intentionally sparsified fine-scale labels. Meanwhile, labels from testing sets remain intact to accurately assess the downscaling results. For details on dataset characteristics; the partitioning of training, validation, and testing sets based on years; and the use of label masking to simulate data sparsity, please refer to Appendix B.1.

Research questions. We designed evaluations focusing on three core questions:
RQ1. How well does MSGL downscale stream temperature compared to baseline models?
RQ2. How robust is the MSGL model under varying levels of data sparsity?
RQ3. Which is more effective: Cross-Domain Pretraining or Asynchronous MSGL?

Downscaling performance (RQ1)

Baselines. We evaluate MSGL against several baselines, including commonly used single-scale spatiotemporal graph models such as RGrN (Jia et al. 2021b), Spatio-Temporal Graph Convolutional Network (STGCN; Yu, Yin, and Zhu 2018), Diffusion Convolutional Recurrent Neural Network (DCRNN; Li et al. 2018), and Graph Attention Spatial-Temporal Network (GASTN; He 2020), and Multi-Scale (MS) options such as MS-STGCN (Chen 2021) and MS Graph Network (MSGNET; Cai 2024). Appendix D explores additional baseline methods for the special case of downscaling from a regionally trained model to a local watershed.

Performance comparison. Table 1 compares the predictive performance of MSGL with several baseline models across different watersheds using 1% of fine-scale observations randomly selected from the training period. The multi-scale models (MS-STGCN, MSGNET, MSGL, and ASYNC-MSGL) perform substantially better in sparse scenarios than the single-scale models (RGrN, STGCN, GASTN, and DCRNN). Among the multi-scale models, our MSGL and ASYNC-MSGL consistently exhibit superior performance in all four watersheds, demonstrating the robustness and adaptability of our approach across datasets.

We further evaluate the contributions of each component (CSL, CrSL, FSL, and MSO) through an ablation study. The combination of CSL and FSL (”MSGL (w/o CrSL)”) improves performance over individual CSL and FSL but does not achieve the performance of the complete MSGL model. This indicates the importance of CrSL for capturing the relationship across scales. Similarly, the combination of CrSL and FSL (”MSGL (w/o CSL)”) outperforms individual tasks but still falls short compared to the complete MSGL model. This highlights the critical role of CSL in capturing long-distance patterns and general water temperature dynamics at the coarse scale, essential for enhancing the overall predictive accuracy. For the optimization process, the comparison between ”MSGL” and ”MSGL (w/o MSO)”

Category	Method	LOWER DELAWARE RMSE STDEV	LWBRANCH DELAWARE RMSE STDEV	NEVERSINK RMSE STDEV	RANCOCAS RMSE STDEV
Single-scale Baselines	STGCN (2018)	1.653 ± 0.076	2.662 ± 0.107	2.059 ± 0.206	1.984 ± 0.130
	DCRNN (2018)	1.679 ± 0.040	2.487 ± 0.134	2.297 ± 0.120	2.489 ± 0.287
	GASTN (2020)	2.141 ± 0.342	2.525 ± 0.118	2.331 ± 0.153	3.134 ± 0.658
	RGrN (2021)	1.732 ± 0.131	2.648 ± 0.154	1.935 ± 0.090	2.198 ± 0.246
Multi-scale Baselines	MS-STGCN (2021)	1.484 ± 0.117	2.402 ± 0.116	1.937 ± 0.069	1.723 ± 0.263
	MSGNET (2024)	1.492 ± 0.118	2.313 ± 0.125	1.927 ± 0.060	1.813 ± 0.351
Ablation Study	MSGL (CSL)	5.799 ± 2.171	5.465 ± 0.462	4.215 ± 0.150	5.965 ± 0.841
	MSGL (CrSL)	1.578 ± 0.112	2.791 ± 0.144	2.623 ± 0.293	1.863 ± 0.173
	MSGL (FSL)	1.732 ± 0.131	2.648 ± 0.154	1.935 ± 0.090	2.198 ± 0.246
	MSGL (w/o CSL)	1.470 ± 0.056	2.391 ± 0.118	2.249 ± 0.290	1.814 ± 0.225
	MSGL (w/o CrSL)	1.551 ± 0.057	2.593 ± 0.167	2.048 ± 0.256	2.150 ± 0.183
	MSGL (w/o MSO)	**1.391** ± 0.069	2.432 ± 0.108	2.035 ± 0.204	1.743 ± 0.201
Sync vs. Async	MSGL	**1.384** ± 0.082	2.375 ± 0.106	1.919 ± 0.058	**1.450** ± 0.109
	ASYNC-MSGL	**1.360** ± 0.116	**2.179** ± 0.129	**1.617** ± 0.134	**1.482** ± 0.053

Table 1: Root mean squared errors (RMSE, °C, mean and standard deviation of nine replicates) for various methods across four watersheds using 1% random subsets of observed water temperatures for training. Single-scale baselines (RGrN, STGCN, and DCRNN), and multi-scale baselines (MS-STGCN and MSGNET) are described in RQ1-Baselines. The ablation study is described in RQ1-"Performance comparison." Bold values are the best results in each watershed and those not significantly worse (Welch's t-test, p > 0.05).

MSGL	LowerDelaware				LwBranchDelaware				Neversink				Rancocas			
	0.1%	Δ	1%	Δ	0.1%	Δ	1%	Δ	0.1%	Δ	1%	Δ	0.1%	Δ	1%	Δ
Without Pre-training	2.521	—	1.384	—	2.687	—	2.375	—	2.323	—	1.919	—	4.431	—	1.450	—
Pre-trained by Brandywine	2.006	-0.515	1.408	0.024	2.936	0.249	2.495	0.12	2.664	0.341	1.759	-0.16	3.855	-0.576	1.489	0.039
Pre-trained by $\mathcal{D}_{\text{mapping}}$	1.818	-0.703	1.360	-0.024	2.376	-0.311	2.179	-0.196	2.036	-0.287	1.617	-0.302	2.794	-1.637	1.483	0.033

Table 2: Performance of Cross-Domain Pretraining ("MSGL (Pre-trained by Brandywine-Christina)") and Asynchronous MSGL (ASYNC-MSGL, i.e., "MSGL (Pre-trained by $\mathcal{D}_{\text{mapping}}$)") in four watersheds. We simulated data sparsity by using 0.1% or 1% of fine-scale labels for fine-tuning. Dark blue indicates a large decrease ($\Delta < -0.2$), medium blue a moderate decrease ($-0.2 \leq \Delta < -0.1$), light blue a slight decrease ($-0.1 \leq \Delta < 0$), and light red any increase ($\Delta > 0$) in root mean squared error (RMSE, °C) relative to MSGL without Pre-training. Brandywine is one of the watersheds shown in Figure 2.

indicates that MSO substantially enhances performance for the Lower West Branch Delaware, Neversink, and Rancocas watersheds, and provides a slight improvement for Lower Delaware. This suggests that MSO contributes to downscaling across different watersheds and highlights the overall benefit of incorporating MSO in the optimization process.

As for MSGL and ASYNC-MSGL, adding the asynchronous strategy substantially improves performance across most watersheds and two data sparsity levels (Table 1 at 1% of fine-scale labels and Figure 3 at 0.1%), with the Rancocas watershed at 1% the only exception.

Robustness under different sparsity levels (RQ2)

To simulate label sparsity in these unusually well-observed study watersheds, fine-scale observations were randomly masked from all of each watershed's reach-date combinations (Appendix B.3). In Figure 4, we compare robustness under data sparsity among the six spatiotemporal graph models introduced in RQ1-Baselines and our two new methods. Whereas MSGL and ASYNC-MSGL perform similarly to baseline methods when fine-scale labels are abundant (trained on 20%-100% of original observations), they outperform the baseline methods when labels are sparse (0.1%-20% of original observations). Specifically, as the fine-scale labels become sparser, the RMSE of MSGL and ASYNC-MSGL increases more slowly than for other methods. In extremely sparse environments, such as with only 0.1% of observations (which is representative of real-world downscaling tasks), ASYNC-MSGL consistently and substantially outperforms other methods, demonstrating robustness to data sparsity.

Comparison of cross-domain pretraining and Async-MSGL (RQ3)

Cross-domain pretraining is a common pre-training strategy aimed at improving the generalization ability of models. We utilize data from the Brandywine-Christina River Watershed, as shown in Figure 2, for cross-domain pretraining. Results in Table 2 suggest that pretraining on $\mathcal{D}_{\text{mapping}}$ (ASYNC-MSGL) generally provides better perfor-

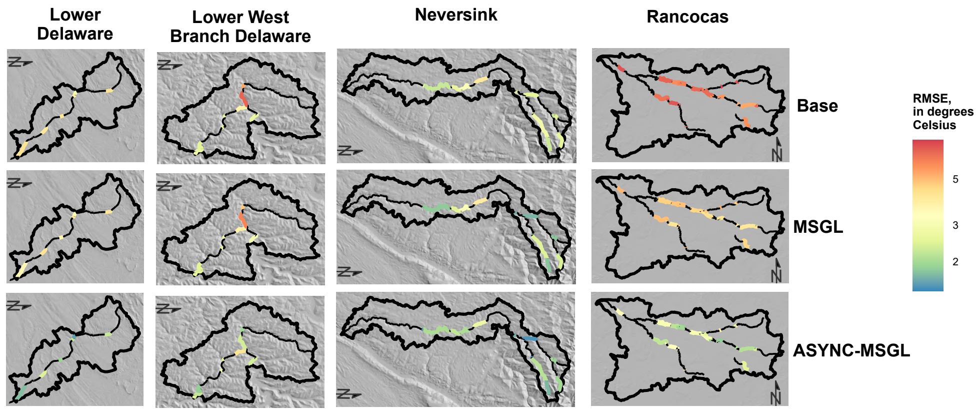

Figure 3: Root mean squared error (RMSE, °C) of methods RGrN (Base), MSGL, and ASYNC-MSGL applied to four watersheds, with 0.1% of the fine-scale labels in each watershed used for training. Reaches of varying lengths are color-coded by average RMSE over all dates. (Map layers from U.S. Environmental Protection Agency and U.S. Geological Survey 2012.)

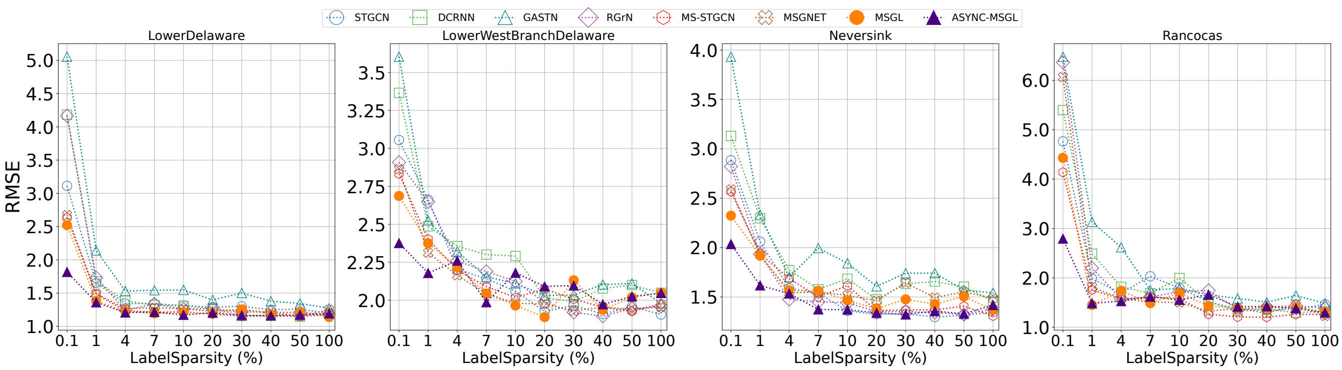

Figure 4: Robustness of spatiotemporal graph models across varying label sparsity levels (training on 0.1%, 1%, 4%, 7%, 10%, 20%, 30%, 40%, 50%, and 100% of all fine-scale reach-days) in four watersheds. Filled markers represent our new methods MSGL and ASYNC-MSGL, whereas open markers represent the methods referenced in section RQ1-Baselines.

mance across different watersheds, particularly in scenarios with very limited fine-scale data (0.1%). Although cross-domain pretraining can substantially benefit downscaling on certain sparse datasets, these improvements are inconsistent; in many cases, it may lead to slight degradation or negligible change in performance. In contrast, our ASYNC-MSGL enhances model performance more consistently across watersheds and data sparsity scenarios.

The effectiveness of any stream temperature downscaling approach hinges on (1) the existence of correlations between coarse- and fine-scale information, such that values predicted at a coarse scale, when remapped to a finer scale, accurately reflect the general patterns of temperature variation, and (2) that the finer-resolution inputs and observations contain additional information about temperature variation not present in the coarse data. The asynchronous strategy further assumes that (3) coarse predictions made from a broader-extent model reflect additional learned information not present in the local data at the fine or coarse resolution. The success of ASYNC-MSGL in these experimental results, even when fine-scale observations are sparse, validates these three assumptions.

Conclusion

We introduce MSGL to address the challenges of fine-scale stream temperature prediction where data are sparse. Existing general multi-resolution approaches fail to consider the graph structure of water temperature data at different scales. Building on a hydrological understanding of stream graph structure, we propose a cross-scale learning task that incorporates positional information between nodes at different scales, enabling the model to capture physical connections across scales. We further enrich MSGL with an asynchronous strategy for pretraining on predictions from a coarse-resolution regional model, $\mathcal{D}_{\text{mapping}}$, to leverage regional knowledge for local prediction. Extensive experiments in the Delaware River Basin demonstrate the superior predictive accuracy of MSGL and ASYNC-MSGL over six baseline single-scale and multi-scale graph models, especially in scenarios of data sparsity. The methods presented here are readily transferable to other cases where data are irregularly structured and sparsely observed at the target resolution, such as prediction of other water quality variables or modeling of meteorological dynamics in complex terrain.

Acknowledgements

YF, RY, and XJ were supported by the National Science Foundation (NSF) under grants 2239175, 2316305, 2147195, and 2425845, U.S. Geological Survey (USGS) awards G21AC10564 and G22AC00266, and the NASA grant 80NSSC24K1061. YX was supported by the NSF under grants 2126474, 2147195, 2425844, 2430978, the Google's AI for Social Good Impact Scholars program. This research was also supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. JRB was supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Environmental System Science Data Management Program, ExaSheds project, under Award Number 89243021SSC000068. APA was supported by the USGS Water Availability and Use Science Program. We thank Lauren Koenig for sharing prepared data and insights into the fine-scale stream networks. We thank Jeremy Diaz, Scott Hamshaw, and several anonymous peers for their thoughtful reviews. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Model inputs, code, and outputs are in Barclay et al. 2024 (https://doi.org/10.5066/P1UP5DXN).

References

Abatzoglou, J. T. 2013. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1): 121–131.

Barclay, J. R.; Fan, Y.; Koenig, L.; Yu, R.; Sun, Y.; Xie, Y.; Jia, X.; and Appling, A. P. 2024. Downscaling and multi-scale modeling of stream temperature in five watersheds of the Delaware River Basin, 1979-2021. In *U.S. Geological Survey data release*.

Barclay, J. R.; et al. 2023. Train, Inform, Borrow, or Combine? Approaches to Process-Guided Deep Learning for Groundwater-Influenced Stream Temperature Prediction. *Water Resources Research*, 59(12): e2023WR035327.

Brett, J. R. 1971. Energetic responses of salmon to temperature. A study of some thermal relations in the physiology and freshwater ecology of sockeye salmon (Oncorhynchus nerkd). *American Zoologist*.

Cai, W. e. a. 2024. MSGNet: Learning Multi-Scale Inter-series Correlations for Multivariate Time Series Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Chau, S. L.; Bouabid, S.; and Sejdinovic, D. 2021. Deconditional downscaling with gaussian processes. In *Advances in Neural Information Processing Systems*, volume 34, 17813–17825.

Chaudhuri, C.; and Robertson, C. 2020. CliGAN: A Structurally Sensitive Convolutional Neural Network Model for Statistical Downscaling of Precipitation from Multi-Model Ensembles. *Water*, 12(12): 3353.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chen, S.; Appling, A.; Oliver, S.; Corson-Dosch, H.; Read, J.; Sadler, J.; Zwart, J.; and Jia, X. 2021. Heterogeneous stream-reservoir graph networks with data assimilation. In *2021 IEEE International Conference on Data Mining (ICDM)*, 1024–1029. IEEE.

Chen, S.; Kalanat, N.; Xie, Y.; Li, S.; Zwart, J. A.; Sadler, J. M.; Appling, A. P.; Oliver, S. K.; Read, J. S.; and Jia, X. 2023. Physics-guided machine learning from simulated data with different physical parameters. *Knowledge and Information Systems*, 65(8): 3223–3250.

Chen, S.; Zwart, J. A.; and Jia, X. 2022. Physics-guided graph meta learning for predicting water temperature and streamflow in stream networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2752–2761.

Chen, Z. e. a. 2021. Multi-Scale Spatial Temporal Graph Convolutional Network for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Daulton, S.; et al. 2022. Robust multi-objective bayesian optimization under input noise. In *International Conference on Machine Learning*. PMLR.

Deb, K.; and Gupta, H. 2005. Searching for robust pareto-optimal solutions in multi-objective optimization. In *Evolutionary Multi-Criterion Optimization*, 150–164. Springer Berlin Heidelberg.

Di Giovanni, F.; Giusti, L.; Barbero, F.; Luise, G.; Lio, P.; and Bronstein, M. M. 2023. On over-squashing in message passing neural networks: The impact of width, depth, and topology. In *International Conference on Machine Learning*, 7865–7885. PMLR.

Désidéri, J.-A. 2012. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 313–318.

Fotakis, D. e. a. 2021. Efficient algorithms for learning from coarse labels. In *Conference on Learning Theory*. PMLR.

Gu, J.; et al. 2022. Multi-scale high-resolution vision transformer for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hamelijnck, O.; Damoulas, T.; Wang, K.; and Girolami, M. A. 2019. Multi-resolution multi-task Gaussian processes. In *Advances in Neural Information Processing Systems*.

Harmel, R. D.; Preisendanz, H. E.; King, K. W.; Busch, D.; Birgand, F.; and Sahoo, D. 2023. A Review of Data Quality and Cost Considerations for Water Quality Monitoring at the Field Scale and in Small Watersheds. *Water*, 15(17).

He, E.; Xie, Y.; Sun, A.; Zwart, J.; Yang, J.; Jin, Z.; Wang, Y.; Karimi, H.; and Jia, X. 2024. Fair Graph Learning Using Constraint-Aware Priority Adjustment and Graph Masking in River Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22087–22095.

He, K. e. a. 2020. Graph attention spatial-temporal network with collaborative global-local learning for citywide mobile traffic prediction. *IEEE Transactions on Mobile Computing*, 21(4): 1244–1256.

He, M.; Li, B.; and Chen, H. 2017. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Ji, J.; Zhong, B.; and Ma, K.-K. 2021. Single image super-resolution using asynchronous multi-scale network. *IEEE Signal Processing Letters*, 28: 1823–1827.

Jia, X.; Chen, S.; Zheng, C.; Xie, Y.; Jiang, Z.; and Kalanat, N. 2023. Physics-guided graph diffusion network for combining heterogeneous simulated data: An application in predicting stream water temperature. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 361–369. SIAM.

Jia, X.; Xie, Y.; Li, S.; Chen, S.; Zwart, J.; Sadler, J.; Appling, A.; Oliver, S.; and Read, J. 2021a. Physics-guided machine learning from simulation data: An application in modeling lake and river systems. In *2021 IEEE International Conference on Data Mining (ICDM)*, 270–279. IEEE.

Jia, X.; et al. 2021b. 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)*. Society for Industrial and Applied Mathematics.

Jiang, K.; et al. 2020. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *6th International Conference on Learning Representations (ICLR 2018)*.

Li, Z.; Liu, H.; Zhang, C.; and Fu, G. 2024. Real-time water quality prediction in water distribution networks using graph neural networks with sparse monitoring data. *Water Research*, 250: 121018.

Liu, J.; et al. 2021. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 105–119.

Mahapatra, D.; and Rajan, V. 2020. Multi-task learning with user preferences: Gradient descent with controlled ascent in Pareto optimization. In *International Conference on Machine Learning*. PMLR.

Moshe, Z.; Metzger, A.; Elidan, G.; Kratzert, F.; Nevo, S.; and El-Yaniv, R. 2020. Hydronets: Leveraging river structure for hydrologic modeling. *arXiv preprint arXiv:2007.00595*.

Oliver, S.; Sleckman, M.; Appling, A.; Corson-Dosch, H.; Zwart, J.; Thompson, T.; Koenig, L.; White, E.; Watkins, D.; Platt, L.; Padilla, J.; and Sadler, J. 2022. Data to support water quality modeling efforts in the Delaware River Basin.

Qin, J.; Huang, Y.; and Wen, W. 2020. Multi-scale feature fusion residual network for single image super-resolution. *Neurocomputing*, 379: 334–342.

Ravindranath, A.; et al. 2016. An environmental perspective on the water management policies of the Upper Delaware River Basin. *Water Policy*.

Regan, R. S.; Markstrom, S. L.; Hay, L. E.; Viger, R. J.; Norton, P. A.; Driscoll, J. M.; and LaFontaine, J. H. 2018. Description of the national hydrologic model for use with the precipitation-runoff modeling system (PRMS). Technical Report 6-B9, US Geological Survey.

Roberts, J. J.; Cassula, A. M.; Silveira, J. L.; Bortoni, E. d. C.; and Mendiburu, A. Z. 2018. Robust multi-objective optimization of a renewable based hybrid power system. *Applied Energy*, 223: 52–68.

Robinson, C.; Hou, L.; Malkin, K.; Soobitsky, R.; Czawlytko, J.; Dilkina, B.; and Jojic, N. 2019. Large Scale High-Resolution Land Cover Mapping With Multi-Resolution Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31.

Sun, A. Y.; Jiang, P.; Mudunuru, M. K.; and Chen, X. 2021. Explore spatio-temporal learning of large sample hydrology using graph neural networks. *Water Resources Research*, 57(12): e2021WR030394.

Suzuki, S.; et al. 2020. Multi-objective Bayesian optimization using Pareto-frontier entropy. In *International Conference on Machine Learning*. PMLR.

Tabari, H.; Paz, S. M.; Buekenhout, D.; and Willems, P. 2021. Comparison of statistical downscaling methods for climate change impact analysis on precipitation-driven drought. *Hydrology and Earth System Sciences*, 25(6): 3493–3517.

Terry, N.; Briggs, M.; Kushner, D.; Dickerson, H.; Baldwin, A.; Trottier, B.; Haynes, A.; Besteder, C.; Glas, R.; Doctor, D.; Gazoorian, C.; Odom, W.; and Benton, J. 2022. Stream temperature, dissolved radon, and stable water isotope data collected along headwater streams in the upper Neversink River watershed, NY, USA.

Topp, S. N.; Barclay, J.; Diaz, J.; Sun, A. Y.; Jia, X.; Lu, D.; Sadler, J. M.; and Appling, A. P. 2023. Stream temperature prediction in a shifting environment: Explaining the influence of deep learning architecture. *Water Resources Research*, 59(4): e2022WR033880.

U.S. Environmental Protection Agency; and U.S. Geological Survey. 2012. National Hydrography Dataset Plus (NHDPlus).

U.S. Geological Survey. 2019. National Hydrography Dataset (ver. 2.1).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ward, J. V. 1985. Thermal characteristics of running waters. In *Perspectives in Southern Hemisphere Limnology: Proceedings of a Symposium*. Springer Netherlands.

Williamson, T. N.; Lant, J. G.; Claggett, P.; Nystrom, E. A.; Milly, P. C.; Nelson, H. L.; Hoffman, S. A.; Colarullo, S. J.; and Fischer, J. M. 2015. Summary of hydrologic modeling for the Delaware River Basin using the Water Availability Tool for Environmental Resources (WATER). Technical Report 2015-5143, U.S. Geological Survey Scientific Investigations Report.

Yousefi, F.; Smith, M. T.; and Álvarez, M. A. 2019. Multi-task learning for aggregated data using Gaussian processes. In *Advances in Neural Information Processing Systems*.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 3634–3640.

Zhang, F.; et al. 2021. Multi-branch and multi-scale attention learning for fine-grained visual categorization. In *MultiMedia Modeling: 27th International Conference, MMM 2021*, volume 27. Springer International Publishing.

Zhang, Y.; Charoenphakdee, N.; Wu, Z.; and Sugiyama, M. 2020. Learning from Aggregate Observations. In *Advances in Neural Information Processing Systems*.

Zhao, Z.; Liu, Q.; and Wang, S. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30: 6544–6556.

Zhou, Q.; et al. 2018. A multi-objective robust optimization approach based on Gaussian process model. *Structural and Multidisciplinary Optimization*, 213–233.