Hydrol. Earth Syst. Sci., 29, 2521–2549, 2025
https://doi.org/10.5194/hess-29-2521-2025

Hydrology and Earth System Sciences
Open Access
EGU

Machine learning in stream and river water temperature modeling: a review and metrics for evaluation

Claudia Rebecca Corona[1] **and Terri Sue Hogue**[1,2]

[1]Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, CO 80401, United States
[2]Hydrologic Science and Engineering Program, Colorado School of Mines, Golden, CO 80401, United States

Correspondence: Claudia Rebecca Corona (claudia.corona@mines.edu)

Received: 21 November 2023 – Discussion started: 27 November 2023
Revised: 4 February 2025 – Accepted: 17 March 2025 – Published: 17 June 2025

Abstract. As climate change continues to affect stream and river (henceforth stream) systems worldwide, stream water temperature (SWT) is an increasingly important indicator of distribution patterns and mortality rates among fish, amphibians, and macroinvertebrates. Technological advances tracing back to the mid-20th century have improved our ability to measure SWT at varying spatial and temporal resolutions for the fundamental goal of better understanding stream function and ensuring ecosystem health. Despite significant advances, there continue to be numerous stream reaches, stream segments, and entire catchments that are difficult to access for a myriad of reasons, including but not limited to physical limitations. Moreover, there are noted access issues, financial constraints, and temporal and spatial inconsistencies or failures with in situ instrumentation. Over the last few decades and in response to these limitations, statistical methods and physically based computer models have been steadily employed to examine SWT dynamics and controls. Most recently, the use of artificial intelligence, specifically machine learning (ML) algorithms, has garnered significant attention and utility in hydrologic sciences, specifically as a novel tool to learn undiscovered patterns from complex data and try to fill data streams and knowledge gaps. Our review found that in the recent 5 years (2020–2024), more studies using ML for SWT were published than in the previous 20 years (2000–2019), totaling 57. The aim of this work is threefold: first, to provide a concise review of the use of ML algorithms in SWT modeling and prediction; second, to review ML performance evaluation metrics as they pertain to SWT modeling and prediction to find the commonly used metrics and suggest guidelines for easier comparison of ML performance across SWT studies; and, third, to examine how ML use in SWT modeling has enhanced our understanding of spatial and temporal patterns of SWT and examine where progress is still needed.

1 Introduction

Water temperature in a stream or river plays a vital role in nature and society, regulating dissolved oxygen concentrations (Poole and Berman, 2001), biochemical oxygen demand rates, and chemical toxicities (Cairns et al., 1975; Patra et al., 2015). Additionally, SWT is an important indicator of cumulative anthropogenic impacts on lotic environments (Risley et al., 2010). Observations of SWT changes over time can reveal the effects of streamflow regulation, riparian alteration (Johnson and Jones, 2000), and large-scale climate change (Barbarossa et al., 2021) on local ecosystems. From an ecological standpoint, SWT strongly influences (Ward, 1998) the health, survival, and distribution of freshwater fish (Ulaski et al., 2023; Wild et al., 2023), amphibians (Rogers et al., 2020), and macroinvertebrates (Wallace and Webster, 1996). As climate change progresses, SWT will be an increasingly critical proxy for ecosystem health and function both locally and nationally.

1.1 SWT modeling in the 21st century

Technological advances since the turn of the 20th century have improved our ability to measure SWT in an affordable and dependable manner at varying spatial and temporal resolutions (Benyahya et al., 2007; Dugdale et al., 2017). Despite significant advances in the last 100 years, many stream reaches, stream segments, and entire catchments remain dif-

ficult to access for a myriad of reasons (Ouellet et al., 2020), including but not limited to the following: physical limitations, i.e., streams may be in private property, remote, or dangerous-to-access areas; financial constraints, i.e., access may be limited by monetary resources or lack thereof; and temporal limitations such as uncertainties and inconsistencies in the continuity of measurements or unforeseen equipment loss or failure (Webb et al., 2015; Isaak et al., 2017). In response to these limitations, statistical methods and physically based computer models have been steadily employed over the last few decades to support the advancement of scientific understanding of stream form and function as well as subsequent implications for water management (Cluis, 1972; Caissie et al., 1998; DeWeber and Wagner, 2014; Isaak et al., 2017).

Aided by the continued development of computers and the internet, physical and statistical computer models have gained prominence outside of academia and are more commonly being used by stakeholders and local groups to address a myriad of hydrology challenges (Maheu et al., 2016; Liu et al., 2018; Tao et al., 2020; Rogers et al., 2020). At the same time, the problem-solving success of machine learning (ML), which falls under the umbrella of artificial intelligence, has become increasingly popular in hydrologic sciences in the last few years (DeWeber and Wagner, 2014; Xu and Liang, 2021). Artificial intelligence (AI) describes technologies that can incorporate and assess inputs from an environment, learn optimal patterns, and implement actions to meet stated objectives or performance metrics (Xu and Liang, 2021; Varadharajan et al., 2022). As a subset of AI, the goal of ML algorithms and models is to learn patterns from complex data (Friedberg, 1958). A global call to better predict and prepare for near- and far-future hydrologic conditions has led researchers in the last few decades to use ML algorithms to model hydrologic processes at various temporal and spatial scales (Poff et al., 1996; Solomatine et al., 2008; Cole et al., 2014; Khosravi et al., 2023). For example, a type of ML called artificial neural networks (ANNs) have been used since the 1990s in many subfields of hydrology, such as streamflow predictions (Karunanithi et al., 1994; Poff et al., 1996), rainfall-runoff modeling (Hsu et al., 1995; Shamseldin, 1997), subsurface flow and transport (Morshed and Kaluarachchi, 1998), and flood forecasting (Thirumalaiah and Deo, 1998). For SWT modeling, however, the use of ML algorithms such as ANNs has only recently garnered interest (Zhu and Piotrowski, 2020).

1.2 Study objective

The current work includes an extensive literature review of studies that used ML algorithms and models for river and SWT modeling, hindcasting, and forecasting. The intent of this review is twofold: (1) to introduce ML for hydrologists who have modeling experience and are interested in pursuing ML use for their SWT studies and (2) to provide a broad overview of machine learning applications in SWT. For ML experts, we think that this review could also prove useful as a reference for how ML has been applied in the field of SWT modeling and where improvement is needed. Overall, this article aims to serve as a bridge between hydrologists and machine learning experts. Our review includes papers cited by Zhu and Piotrowski (2020), who previously conducted a study of ANNs used in SWT modeling; however, we provide a comprehensive examination of peer-reviewed journals that use any type of artificial intelligence or ML algorithm to model or evaluate river or SWT. This review's first objective is to provide a concise review of ML algorithm use in SWT modeling. Secondly, our goal is to examine the ML performance evaluation metrics used in SWT modeling and find the most-used metrics and suggest guidelines for clearer comparison of ML performance. The third objective is to discuss the community's use of ML to address physical system understanding in SWT modeling. Overall, this review aims to serve as a critical assessment of the state of SWT understanding given the increasing popularity of ML use in SWT modeling.

2 Overview: stream water temperature model types

2.1 SWT statistical (also stochastic or empirical) models

In the 1960s, considerable interest grew in the prediction of SWT, particularly in the western United States (US) due to increased awareness of environmental quality issues (Ward, 1963; Edinger et al., 1968; Brown, 1969). The creation of large dams, daily release of heated industrial effluents, growing agricultural waste discharge, and forest clear-cutting could influence downstream SWT. However, the extent of such influence remained poorly understood and difficult to test at large spatial and temporal scales (Brown, 1969). From the 1960s to the 1970s, understanding of the relationship between SWT and ambient air temperature (AT) was solidified, and scientists began to increasingly use statistical methods to examine the air–water relationships in stream environments (Morse, 1970; Cluis, 1972). Statistical (also stochastic or empirical) models are governed by empirical relations between SWT and their predictors, which require fewer input data. An example of such progress took place in Canada, where researchers created an autoregressive model to calculate mean daily SWT fluctuations using 6 months of data from the summer and winter months of 1969 (Cluis, 1972). Cluis (1972) further said that their model was transferrable to other streams of comparable size. The use of statistical methods in SWT modeling became increasingly common in the latter half of the 20th century due in large part to minimal data requirements (Benyahya et al., 2007). For example, scientists in Europe used limited data and statistics to examine the influence of atmospheric and topographic factors on the tem-

perature of a small upland stream (Smith and Lavis, 1975). In Australia, scientists interested in finding limits for reaches of streams downstream from thermal discharges found a simple method that could predict SWT based solely on site altitude and AT or upstream SWT (Walker and Lawson, 1977). In Canada, SWT was predicted using a stochastic approach, which included the use of Fourier series, multiple regression analysis, Markov processes, and a Box–Jenkins time series model (Caissie et al., 1998). In the 21st century, statistical methods continue to be a prominent tool used for SWT modeling and prediction (Ahmadi-Nedushan et al., 2007; Chang and Psaris, 2013; Segura et al., 2015; Detenbeck et al., 2016; Siegel and Volk, 2019; Ulaski et al., 2023; Fuller et al., 2023). We refer the reader to Benyahya et al. (2007) for a comprehensive review of SWT statistical models and approaches.

2.2 SWT physically based (also process-based, deterministic, mechanistic) models

While statistical methods can be straightforward to use and require minimal in situ data for first analysis (Benyahya et al., 2007), limitations and uncertainty with regards to SWT predictions are possible, specifically when trying to understand the controls of the energy transfer mechanisms responsible for trends (Dugdale et al., 2017). To address these shortcomings and with the introduction of personal computers in the late 1960s (Dawdy and Thompson, 1967), researchers developed computer models and software programs that tried to address the more fundamental hydrology questions founded in physics and natural processes (Theurer et al., 1984; Bartholow, 1989). One example of such progress was an SWT prediction one-dimensional computer model that used a simplified energy conservation equation to predict SWT for the upper reaches of the Columbia River in the Pacific Northwest of the US during July 1966 (Morse, 1970). These models are described as being physically based or process-based (alternatively called "deterministic" or "mechanistic" models).

Due to the continued lack of sufficient in situ observations and resources with which to undertake field studies in SWT science (Dugdale et al., 2017), physically based models became increasingly used. From the end of the 20th century through the present, they are considered one of the best available options in generating predictions of SWT, particularly at a localized scale (Dugdale et al., 2017). Physically based models became useful enough that government agencies introduced their own models to encourage uniformity. In the 1980s, the US Geological Survey (USGS) introduced a physically based model that simulated SWT called SNTemp (Theurer et al., 1984; Bartholow, 1989). A few years later, the US Environmental Protection Agency (EPA) introduced SHADE-HSPF for similar purposes (Chen et al., 1998a, b). Where available, academic scientists coupled field measurements with physically based numerical models. For example, scientists in Minnesota created a numerical model, called MNSTREM, based on a finite-difference solution of the nonlinear equation to predict SWT at 1 h increments for the Clearwater River (Sinokrot and Stefan, 1993). Similarly, academic scientists in Canada introduced CEQUEAU, a water balance type of model which incorporated vegetation and soil characteristics to solve for SWT (St-Hilaire et al., 2000). Physically based models became commercially available in the 2000s, one example being the MIKE suite of models, which were created to solve the heat and advection–dispersion equation to simulate both surface and subsurface water dynamics, created by the DHI consulting group (Jaber and Shukla, 2012; Loinaz et al., 2013). In addition to the models mentioned, over a dozen more physically based models were created and used between 1990 and 2017 (Dugdale et al., 2017). For a more detailed review of physically based SWT models, we refer the reader to Dugdale et al. (2017).

2.3 Artificial intelligence models in SWT modeling

Initial discussion of artificial intelligence can be traced back to 1943, when McCulloch and Pitts presented a computer model that functioned like neural networks of the brain (McCulloch and Pitts, 1943). In 1958, R.M. Friedberg published *A Learning Machine: Part 1* in IBM's Journal of Research and Development, one of the first to describe the concept of "machine learning". Friedberg hypothesized that machines could be taught how to learn such that they developed the capability to improve their own performance to the point of completing tasks or meeting objectives (Friedberg, 1958). Sixty years later, ML has grown as a field of study in academia and as an area of great interest in society, the latter due in large part to the popularity of large language models (a type of machine learning that we will not discuss here), such as ChatGPT (OpenAI, Inc., 2025), Copilot (Microsoft, Inc., 2025), and Gemini (Google, 2025).

In the last decade, computing advances in AI have started to offer several advantages for using machine learning (ML) in hydrology that are comparable to physically based models (Cole et al., 2014; Rehana and Rajesh, 2023). In contrast to traditional physically based models, the code underlying ML models is generally open-source and publicly available, allowing for near-real-time accessible advances and user feedback, whereas the source code for some physically based models may be inaccessible to the public due to being privately managed (MIKE suite of models), or the model software may be publicly available but could take years to publish updates (USGS MODFLOW, Simunek's HYDRUS). One advantage that has made ML increasingly appealing includes its ability to learn directly from the data (i.e., data-driven), which can be useful when the underlying physics are not fully understood or are considered too complex to model accurately.

Additionally, ML models are more efficient in making predictions compared to the time-intensive solvers of physically based models. ML models can also handle the challenge of

scalability, which means managing large datasets and seamlessly deploying across various computer platforms and applications (Rehana and Rajesh, 2023). Air2stream, a hybrid statistical–physically based SWT model (Toffolon and Piccolroaz, 2015; Piccolroaz et al., 2016), initially outperformed earlier ML models such as Gaussian process regression (Zhu et al., 2019a). However, in the last few years, Air2stream has had its performance matched and even exceeded by recent neural network models (Feigl et al., 2021; Rehana and Rajesh, 2023)

Finally, with computer processing power improving and the emergent field of quantum computing, there is a strong belief that using ML and by extension AI in science applications will drive innovation to the point where natural patterns and insights not currently apparent in physical modeling will be uncovered (Varadharajan et al., 2022). Thus, while physically based models are considered invaluable for their interpretability and grounding in established physics, ML models have the potential for growth in various fields of hydrology, where they can be used to first complement and eventually lead as powerful tools for prediction, optimization, and understanding in increasingly complex and data-rich environments.

For this review, we differentiate between traditional ML and newer ML, where the former includes approaches that have been used for decades in hydrologic modeling, i.e., cluster analysis, support vector machine, and shallow neural networks. We define newer ML as that introduced in hydrologic modeling in recent years, such as the deep learning long short-term memory NNs, extreme learning machine, and ML hybridizations. The following sections provide an overview of ML types and learning techniques. Finally, we assume that readers have a very basic understanding of the differences between machine learning types such as supervised, semi-supervised, and unsupervised learning and refer the reader to Xu and Liang (2021) for a nice overview.

2.3.1 Traditional ML algorithms

K-nearest neighbors

K-nearest neighbors (K-nn) is a versatile supervised ML algorithm (Fix and Hodges, 1952; Cover and Hart, 1967) used to solve nonparametric classification and regression problems. The K-nn algorithm uses proximity between data points to make classifications or evaluations about the grouping of any given data point. K-nn gained popularity in the 2000s due to its simplicity in implementation and understanding, making it readily accessible to hydrologic researchers and practitioners. For example, St.-Hilaire et al. (2011) used various K-nn model configurations to model SWT for the Moisie River in northern Quebec, Canada, finding that the best K-nn model required prior-day SWT data and day of year (DOY), an indicator of seasonality. Advantages of K-nn include its non-assumptions of the underlying distribution of the data, allowing it to handle nonlinear complexities without requiring a solid model structure as is the case for some physical models (St-Hilaire et al., 2011). There are some disadvantages of K-nn: it is computationally intensive and may require extensive cross-validation; performance can be affected by irrelevant and/or redundant features; and due to its high memory and computational needs it is impractical for large-scale applications, i.e., scalability issues (Acito, 2023). For example, Heddam et al. (2022b) compared K-nn with other ML algorithms, finding that K-nn was outperformed by other MLs such as least-squares support vector machine and neural networks. The use of K-nn may still be reasonable for simple, local cases but we suggest other MLs for more complex or larger use cases.

Cluster analysis and variants

Cluster analysis is a category of unsupervised ML methods used to create groups from an unlabeled dataset. Clustering methods use distance functions such as Euclidean distance, Manhattan distance, Minkowski distance, cosine similarity, and others to group data into clusters (Irani et al., 2016). The analysis separates data into groups of maximum similarity, while also trying to minimize the similarity from group to group (Xu and Liang, 2021). In SWT modeling, studies have used cluster analysis to try a reduction of a dataset prior to assessment (Voza and Vuković, 2018) and/or to find spatiotemporal patterns in a dataset (Krishnaraj and Deka, 2020). Another popular clustering technique is discriminant analysis, which tries to find parameters that are most significant for temporal differentiation between rendered periods (Voza and Vuković, 2018). *K*-means, a type of unsupervised ML, is a clustering algorithm that finds k number of centroids in the dataset and distributes each respective data value to the nearest cluster while keeping the smallest number of centroids possible (Krishnaraj and Deka, 2020). Krishnaraj and Deka (2020) used *K*-means to organize spatial grouping for water quality monitoring stations for dry and wet regions along the Ganges River basin in India to identify whether pollution patterns could be discerned.

While cluster analysis and discriminant analysis are generally used to reduce datasets, another technique, the principal component analysis (PCA) (or factor) test, is applied to assess dominant factors in datasets. Mathematically, principal component analysis (PCA) is a statistical unsupervised ML technique that uses an orthogonal transformation (a linear transformation that preserves lengths of vectors and angles) to convert a set of variables from correlated to uncorrelated (Krishnaraj and Deka, 2020). Using PCA, Krishnaraj and Deka (2020) found that certain water quality parameters (dissolved oxygen, sulfate, electrical conductivity) were more dominant in the dry season compared to the wet season (total dissolved solids, sodium, potassium, sodium, chlorine, chemical oxygen demand), data which could be used to cater the monitoring program to the important parameters.

SWT was not a dominant parameter, likely in part because the SWT values of large downstream rivers like the Ganges are generally less variable due to their larger volume and stronger thermal buffer.

Support vector machine and regression

Support vector machine (SVM) is a supervised learning technique used for classification, regression, and outlier detection. The aim of SVM is to find a hyperplane (or the decision surface) in an N-dimensional space (N is the number of features) that best separates labeled categories, or support vectors (Cortes and Vapnik, 1995). One of the advantages of SVM is that it seeks to minimize the upper bound of the generalization error instead of the training error (Cortes and Vapnik, 1995). A big disadvantage is that it does not perform well with large datasets due to the likelihood of greater noise, which would cause support vectors to overlap, making classification difficult. For a more detailed explanation of SVM, we refer the reader to Cortes and Vapnik (1995) and Xu and Liang (2021). In the last few decades, SVM has been coupled with other ML models to find the best-performing models for short-term water quality predictions (Lu and Ma, 2020) and daily SWT modeling (Heddam et al., 2022b). For example, Heddam (2022b) used least-squares SVM (LSSVM), a variant of SVM which takes a linear approach (instead of quadratic-like SVM), to reach a solution (Suykens and Vandewalle, 1999).

A version of SVM used for regression tasks is support vector regression (SVR). SVR attempts to minimize the objective function (composed of loss greater than a specified threshold) and a regularization term (Rehana, 2019; Hani et al., 2023). For further detail on SVR, we refer the reader to Rehana (2019) and Hani et al. (2023). Using historical data, SVR has been compared with other ML models that evaluate SWT variability due to climate change (Rehana, 2019), finding temperature increases less pronounced in the SVR model. Jiang et al. (2022) compared SVR to other ML models to forecast SWT in cascade-reservoir-influenced rivers. For the cascade-reservoir-operation-influenced study, SVR was outperformed by random forest (RF) and gradient boosting (Jiang et al., 2022). Focusing on 78 catchments in the mid-Atlantic and Pacific Northwest hydrologic regions of the US, researchers used SVR and an ML algorithm called XGBoost to predict monthly SWT (Weierbach et al., 2022), finding that SVR significantly outperformed traditional statistical approaches such as multi-linear regression (MLR) but did not outperform XGBoost. In addition, the SVR models had the highest accuracy for SWT across different catchments (Weierbach et al., 2022). In Quebec, Canada, a comparison of four ML models that estimated hourly SWT showed that SVR outperformed by RF (Hani et al., 2023).

A lesser-known form of SVM is its extended form, called relevance vector machine (RVM). RVM is a form of supervised learning that uses a Bayesian framework to solve classification and regression problems (Tipping, 2001). Locally weighted polynomial regression (LWPR) is a form of supervised ML (Moore et al., 1997) used for learning continuous nonlinear mappings from real-valued (i.e., functions whose values are real numbers) inputs and real-valued outputs. LWPR works by adapting the model locally to the respective data points, assigning different weights to different data points based on data point proximity to the target (Moore et al., 1997). This type of regression is best employed when the variance around the regression line is not constant, thereby suggesting heteroscedasticity.

Gaussian process regression and generalized additive models

Gaussian process regression (GPR) is a type of nonparametric supervised learning method used to solve regression problems. As a Bayesian approach, GPR assumes a probability distribution over all functions that fit the data. GPR is specified by a mean function and covariance kernel function which reflect prior knowledge of the trend and level of smoothness of the target function (Xu and Liang, 2021). One of GPR's advantages is the model's ability to calculate empirical confidence intervals, allowing the user to consider refitting predictions to areas of interest in the function space (Grbić et al., 2013). For more details on GPR, we refer the reader to Xu and Liang (2021). Grbić et al. (2013) used GPR for SWT modeling of the river Drava, Croatia, where model no. 1 estimated the seasonal component of SWT fluctuations and model no. 2 estimated the shorter-term component (Grbić et al., 2013). A separate study for the river Drava used three variations of GPR to model SWT, finding that GPR was outperformed by the physically based, stochastically calibrated model, Air2stream (Zhu et al., 2019). More recently, Majerska et al. (2024) used GPR to simulate SWT for a nonglaciated arctic catchment, Fuglebekken (Spitsbergen, Svalbard). Using GPR and another model, the authors identified a diurnal warming trend of 0.5–3.5 °C per decade through the summer season, implying a warming thermal regime in the Fuglebekken catchment (Majerska et al., 2024).

Generalized additive models (GAMs) are a type of semiparametric, nonlinear model with a wide range of flexibility, allowing the model to analyze data without assuming relations between inputs and outputs (Hastie and Tibshirani, 1987). Where GPR uses a probabilistic approach, GAM uses smoothing functions (i.e., splines) to model the relationship between a predictor variable and response variable. GAMs have been used to model SWT for the Sainte-Marguerite River in eastern Canada (Laanaya et al., 2017; Souaissi et al., 2023; Hani et al., 2023). Hani et al. (2023) used GAMs to identify potential thermal refuge areas for Atlantic salmon in two tributary confluences using sub-hourly observations.

Decision trees and classification and regression trees

Decision trees (DTs) are a nonparametric, supervised learning technique. DTs can make predictions or decisions based on a set of input features and are likely to be more accurate where the problem can be solved in a hierarchical sequence of decisions (Breiman, 2001). Classification and regression trees (CARTs) are a specific type of algorithm that builds decision trees, where the internal node in the tree splits the data into two branches (sub-nodes) based on the specified decision rule (Loh, 2008). While CART can quickly find relationships between data, it is prone to overfitting and can be statistical unstable, where a small perturbation in the training data could negatively affect the output of the tree (Hastie et al., 2001; Xu and Liang, 2021). For a detailed explanation of DT and CART, we refer the reader to Hastie et al. (2001), Loh (2008), and Xu and Liang (2021). An SWT modeling study comparing the output of three model versions of DT, GPR, and feed-forward neural networks for multiple sites found that DTs could perform similarly to GPR and feed-forward neural networks when detailed statistics of air temperature, day of year, and discharge were included (Zhu et al., 2019a). However, when comparing daily SWT results from DTs with gradient boosting (GB) or random forest (RF), DTs generally underperform (Anmala and Turuganti, 2021; Jiang et al., 2022). Recent studies have compared CART with other ML algorithms to model water quality parameters (including SWT), finding that CART underperformed due to overfitting compared to RF (Souaissi et al., 2023) and extreme learning machine (ELM) (Heddam et al., 2022a). To combat the problem of overfitting that can occur using decision trees, the idea of using multiple trees by bootstrap aggregation (i.e., bagging) has gained interest.

Random forests and XGBoost

RF and XGBoost have been used to predict daily SWT prediction for Austrian catchments,with results showing minor differences in model performance, with a median RMSE difference of 0.08 °C between tested ML models (Feigl et al., 2021). Using RF and XGBoost along with four other ML models, Jiang et al. (2022) estimated daily SWT below dams in China, finding day of year, streamflow flux, and AT to be most influential in the prediction of SWT (Jiang et al., 2022). Weierbach et al. (2022) used XGBoost and SVR to predict SWT at monthly timescales for the Pacific Northwest region of the US, showing that an ensemble XGBoost outperformed all modeling configurations for spatiotemporal predictions in unmonitored basins, with AT identified as the primary driver of monthly SWT. Zanoni et al. (2022) used RF and a deep learning model to develop regional models of SWT and other water quality parameters, with RF performance comparatively less effective at detecting nonlinear relationships than the deep learning model, though both models identified AT as most influential (Zanoni et al., 2022).

Souaissi et al. (2023) tested the performance of RF and XGBoost with nonparametric models for the regional estimation of maximum SWT at ungaged locations in Switzerland, finding no significant differences between the ML performance and the nonparametric model performances, which was attributed to the lack of a large dataset. Hani et al. (2023) used four supervised ML models – MARS, GAM, SVM, and RF – to model potential thermal refuge area (PTRA) at an hourly time step for two tributary confluences of the Sainte-Marguerite River in Canada. RF had the highest accuracy at both locations in terms of hourly PTRA estimates and modeling SWT (Hani et al., 2023). Wade et al. (2023) conducted a CONUS-scale study using 410 USGS sites with 4 years of daily SWT and discharge to examine maximum SWT. They used RF to estimate max SWT and thermal sensitivity (Wade et al., 2023), finding that AT was the most influential control followed by other properties (watershed characteristics, hydrology, anthropogenic impact).

2.3.2 Traditional artificial neural networks (ANNs)

An artificial neural network (ANN) is a type of ML algorithm inspired by biological neural networks in the brain (McCulloch and Pitts, 1943; Hinton, 1992). ANNs learn from data provided and improve on their own to progressively extricate higher-level trends or relationships within the given dataset (Hinton, 1992). Currently, ANNs are capable of data classification, pattern recognition, and regression analysis. Considered robust, ANNs can undergo supervised, unsupervised, semi-supervised, and reinforcement learning. The first study that utilized ANNs specifically for SWT modeling was published around the year 2000. The work was done by researchers interested in hindcasting SWT for a river in Canada for a 41-year period dating back to 1953 (Foreman et al., 2001). Since 2000, various types of ANNs have been increasingly used to model SWT at various sites at hourly, daily, and monthly time steps. For more detail on traditional ANNs, with descriptions of ANN variants and back-propagation alternatives, we refer the reader to Appendix A.

2.3.3 Newer and recent ML algorithms

We define newer and recent ML algorithms as those introduced or reintroduced in the last decade for SWT modeling. These ML algorithms include deep (i.e., increased layers) ANNs such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), extreme learning machine (ELM), ML hybridizations, and subsets.

A "deep" neural network (DNN) has three or more hidden layers, MLPNNs being one such example. The purpose of added layers is to serve as optimizations for greater accuracy. Due to their complex nature, DNNs need extensive time spent solely on training the network on the input data (Abdi et al., 2021). Convolutional neural networks (CNNs) are FFNNs used to recognize objects and patterns in visual

data (LeCun et al., 1989, 2004). CNNs have convolutional layers, which hold one or more filters that calculate a local weighted sum as they analyze the input data. A CNN filter is a matrix (rows and columns) of randomized number values that convolves (i.e., moves) through the pixels of an image, taking the dot product of the matrix of values in the filter and the pixel values of the image. The dot product is used as input for the next convolutional layer. To ensure adequate performance, CNNs must be trained with examples of correct output in the form of labeled training data and should be calibrated (i.e., adjusting filters, implement loss functions) to optimize performance (Krizhevsky et al., 2012). For more detail on CNN, we refer the reader to LeCun et al. (2004), Krizhevsky et al. (2012), and Xu and Liang (2021). A disadvantage of CNNs is that they are not ideal for interpreting temporal or sequential information or data that require learning from past data to predict future output. For interpreting temporal information or sequential data, recurrent neural networks are preferred.

Unlike FFNNs, recurrent neural networks (RNNs) work in a chain-link nature that allows them to loop (i.e., keep) previously handled data for use in a present task to make better predictions (Hochreiter and Schmidhuber, 1997). The RNN architecture is better equipped (and preferred) to handle temporal (i.e., time series) or sequential (i.e., a video is a sequence of images) data due to their ability to learn from their past (Bengio et al., 1994). The Elman neural network (ELM-NN) is a type of RNN where the hidden layer (bi-directionally connected to the input layer and output layer) stores contextual information of the input that it sends back to the input layer with sequential time steps (Elman, 1990).

However, one of the issues that persists in RNNs is that there is a limit to how far back RNNs can access past data to make better predictions. This is described as the problem of long-term dependencies, also known as the vanishing gradient problem. The vanishing gradient problem is due to back-propagated gradients that can grow or shrink at each time step, increasing instability until the gradients "explode" or "vanish" (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997). Hochreiter and Schmidhuber (1997) introduced the long short-term memory (LSTM) model, a type of RNN explicitly designed to overcome the vanishing gradient problem. The LSTM architecture includes three gates (input, forget, and output gates) that control the flow of information in and out of the cell state, allowing the ANN to store and access data over longer time periods. In the last few decades, LSTMs have improved and variations introduced (Gers and Schmidhuber, 2000; Cho et al., 2014; Yao et al., 2015), and many have been cross-compared, with findings showing similar performance across LSTMs (Greff et al., 2016). In the last few years, LSTMs and their variations have been revisited and employed in hydrologic studies to examine possible relationships in time series data (Shi et al., 2015; Shen, 2018; Kratzert et al., 2018, 2019). For example, Sadler et al. (2022) used an LSTM model to multi-task, i.e., predict two related variables – streamflow and SWT. Their argument for forcing an LSTM to multi-task is that if two variables are driven by the same underlying physical processes, a multi-tasking LSTM could more holistically represent shared hydrologic processes and thus better predict the variable of interest. Their LSTM model consisted of added components: specifically, two parallel, connected output layers that represented streamflow output and SWT output (Sadler et al., 2022). Overall, using the multi-tasking LSTM improved accuracy for half the sites, but for those sites with marked improvement, more calibration was needed to reach improvement (Sadler et al., 2022).

Another type of NN is the graph neural network (GNN), which is used for representation learning (unsupervised learning of feature patterns) of graphed data (Bengio et al., 2013), where a "graph" denotes the links between a collection of nodes. At each graph node or link, information in the form of scalars or embeddings can be stored, making them very flexible data structures. Example of graphs that we interact with regularly are images, where each pixel is a node and is linked to adjacent pixels. A stream network is also an example of a graph, albeit a directed graph, which is a graph in which the links (also called "edges") have direction. Two examples of recent GNNs are recurrent graph convolution networks (RGCNs) and temporal convolution graph models (TCGMs). The RGCN utilizes LSTM network architecture (i.e., use of forget, input, output gates) for temporal recognition (Topp et al., 2023). In contrast to RGCN, TCGM uses 1D convolutions (i.e., input a three-dimensional object and output a three-dimensional object), pooling, and channel-wise normalization to capture low-, intermediate-, and high-level temporal information in a hierarchical manner (Lea et al., 2016). An example that utilizes this approach is Graph WaveNet (Wu et al., 2019), which has been used in spatial-temporal modeling of SWT (Topp et al., 2023). According to Topp et al. (2023), the temporal convolutional structure of Graph WaveNet is more stable in the gradient-based optimization process in contrast to the possible gradient explosion problem that the LSTM in the RGCN could experience.

While present studies continue to use ML models as standalones to evaluate SWT predictions, other studies have coupled modern ML with non-ML models to examine whether such combinations improve model performance (Graf et al., 2019; Qiu et al., 2020; Rehana and Rajesh, 2023). For example, Graf et al. (2019) coupled four discrete wavelet transform (WT) techniques with MLPNN to predict SWT for eight stations on the Warta River in Poland. For reference, WT is widely applied for the analysis and denoising of information (signals) and images both over time and on a domain scale (frequency). The unique characteristic of a wavelet neural network (WNN) is the use of the WT as the activation function in the hidden layer of the NN (Qiu et al., 2020). Zhu et al. (2019) coupled WT with MLPNN and ANFIS to evaluate daily SWT at two stations on the river Drava in Croatia

and separately compared the WT-ML coupling with MLR. The study found that the combination of WT and ML improved performance compared to the standalone models (Zhu et al., 2019d). A recent ML approach called differentiable modeling incorporates physics into ML modeling frameworks, where the basic model structure and parameters of a process-based model are inserted into an ANN to estimate parameters or replace existing process descriptions (Rahmani et al., 2023). Rahmani et al. (2023) examined model components that could improve an LSTM model's ability to better match model predictions to field observations. From their study, Rahmani et al. (2023) found that adding a separate shallow subsurface flow component to the LSTM model and a recency-weighted averaging of past air temperature for calculating source SWT resulted in improved predictions (Rahmani et al., 2023).

Attention-based transformers are a more novel type of deep learning that has led to advancements in natural language processing, in the form of ChatGPT, Microsoft's CoPilot, Google's Gemini, and others. Due to their exponential success in the last few years, attention-based transformer models have been used in geological science fields such as oceanography for sea surface temperature prediction (Shi et al., 2015), hydrology for streamflow and runoff prediction (Ghobadi and Kang, 2022; Wei, 2023), and remote sensing for streambed land use change classification (Bansal and Tripathi, 2024). As a relatively new AI tool, attention-based transformers have yet to be used for SWT (to our knowledge), but their applications in other geological science fields suggest it is only a matter of time before their use is observed in SWT modeling.

2.4 SWT predictions using ML

2.4.1 Identifying model complexity

The strong success of ML use in SWT modeling warrants a brief and broad overview on identifying model complexity to minimize overfitting and underfitting of models. When a model is too complex, i.e., has too many features or parameters relative to the number of observations, or is forced to overextend its capabilities, i.e., to make predictions with insufficient training data, the model runs the risk of overfitting (Srivastava et al., 2014). An overfitted model fits the training data "too well", capturing noise and details that provide high accuracy on a training dataset, only to perform poorly once the model encounters "unseen" data in testing and validation (Xu and Liang, 2021). Scenarios where overfitting may be temporarily acceptable are (1) model development is at preliminary stages, such as a "proof of life" concept; (2) when the objective is to identify heavily relied on features by the model, i.e., feature importance; or (3) in highly controlled modeling environments where the expected data will be consistently similar to the training dataset. The latter is more likely in industrial applications and unlikely in the changing nature of hydrology.

In contrast, underfitting occurs when a model is too simple to capture any patterns in the data, which can also lead to unsatisfactory performance in training, testing, and validation. Underfitting can occur with inadequate model features and poor model complexity or when regularization techniques (e.g., L1 or L2 regularization) are over-used, making the model too rigid and unable to respond to changes in the data. Given the propensity of ML models to effectively learn the training data, underfitting is less of an issue in ML, whereas overfitting can be widespread. In Fig. 1, we present an example workflow that researchers can use to transition away from overfitting and towards model generalizability. In the five-step outline (Fig. 1), we suggest the need for "temporal, unseen, ungaged region tests" (TUURTs) in SWT ML modeling. The idea behind TUURTs has been applied for decades in SWT process-based (Dugdale et al., 2017) and statistically based models (Benyahya et al., 2007; Gallice et al., 2015) to improve SWT model robustness. In TUURTs, testing for unseen cases means testing only within the developmental dataset, whereas testing for "ungaged" cases means testing for new sites that have no data and have not been previously seen by the model at all. Some statistically based models, such as DynWat (Wanders et al., 2019) and the Pacific Northwest (PNW) SWT model (Siegel et al., 2023), have tested for ungaged regions and unseen data. In the last few years, ML-SWT studies have begun applying TUURTs (Hani et al., 2023; Rahmani et al., 2020, 2021, 2023; Souaissi et al., 2023; Topp et al., 2023; Philippus et al., 2024a) but more ML-SWT studies need to apply these tests to improve user confidence in extrapolation capability. We further encourage researchers to shift towards more generalizable models, which are in theory more capable of performing well across diverse scenarios and datasets and stand to become increasingly important with the unpredictability of climate extremes.

2.4.2 Model inputs for ML-SWT

Using air temperature (AT) to better understand SWT has been considered since the 1960s, when Ward (1963) and Edinger et al. (1968) discussed the influence of air temperature on SWT. Since then, various input variables have been tested (see Table S1); however, the model inputs of AT and SWT continue to be the most used in ML modeling studies. For example, studies have used AT from time periods outside of the known SWT record to improve ML model performance (Sahoo et al., 2009; Piotrowski et al., 2015; Graf et al., 2019). In addition to AT and SWT, flow discharge has been used to attempt to constrain SWT (Foreman et al., 2001; Tao et al., 2020; St-Hilaire et al., 2011; Grbić et al., 2013; Piotrowski et al., 2015; Graf et al., 2019; Qiu et al., 2020). Other model inputs include precipitation (Cole et al., 2014; Jeong et al., 2016; Rozos, 2023), wind direction and speed (Hong and Bhamidimarri, 2012; Cole et al., 2014; Jeong

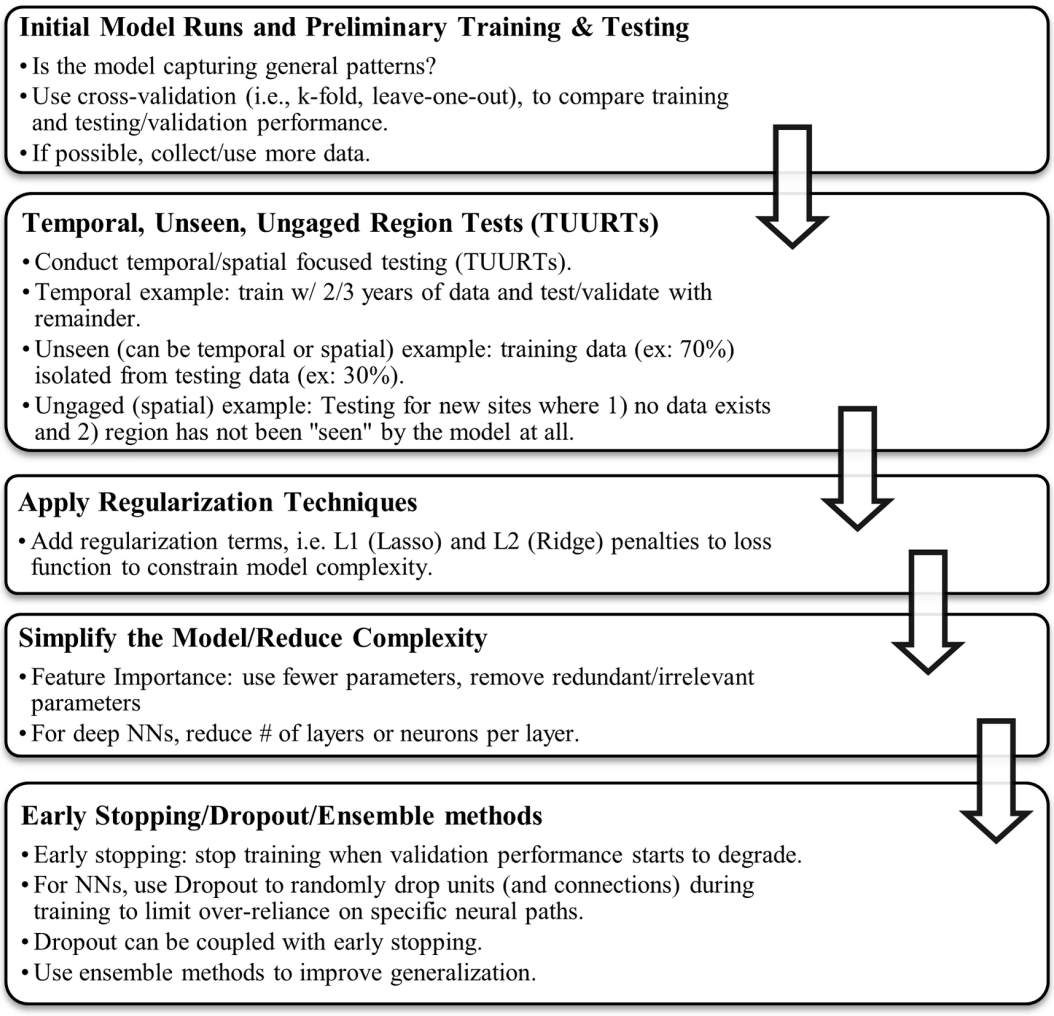

Figure 1. Diagram outlining steps that can be taken in the modeling process to mitigate overfitting.

et al., 2016; Kwak et al., 2016; Temizyurek and Dadaser-Celik, 2018; Abdi et al., 2021; Jiang et al., 2022), barometric pressure (Cole et al., 2014), landform attributes (Risley et al., 2003; DeWeber and Wagner, 2014; Topp et al., 2023; Souaissi et al., 2023), and many more (see Table S1).

In the last few years, including the day of year (DOY) as an input (Qiu et al., 2020; Heddam et al., 2022a; Drainas et al., 2023; Rahmani et al., 2023) and humidity (Cole et al., 2014; Hong and Bhamidimarri, 2012; Kwak et al., 2016; Temizyurek and Dadaser-Celik, 2018; Abdi et al., 2021) has also been shown to better capture the seasonal patterns of SWT (Qiu et al., 2020; Philippus et al., 2024a). With improved access to remote sensing data, there has also been a notable increase in satellite product inputs such as estimates of sky cover (Cole et al., 2014), solar radiation (Kwak et al., 2016; Topp et al., 2023; Majerska et al., 2024), sunshine per day (Drainas et al., 2023), and potential evapotranspiration or ET (Rozos, 2023; Topp et al., 2023). However, more research is needed to better understand the influence of newer model inputs on SWT (Zhu and Piotrowski, 2020).

Recently, SWT studies focused on the CONUS scale have chosen to use as many model inputs as available, with Wade et al. (2023) using a point-scale CONUS ML study using over 20 variables, while Rahmani et al. (2023) created an LSTM model and considered over 30 variables to simulate SWT. Despite the use of diverse data, the models in these studies performed only satisfactorily and were deemed not generalizable, leaving much room for improvement in CONUS-scale modeling of SWT. With the compilation of larger and larger datasets, feature importance in ML, which is the process of using techniques to assign a score to model input features based on how good the features are at predicting a target variable, can be an efficient way to improve data comprehension, model performance, and model interpretability, the latter of which can dually serve as a transparency marker of which features are driving predictions. Methods for measuring feature importance include using cor-

relation criteria (Pearson's r, Spearman's ρ), permutation feature importance (shuffling feature values, measuring decrease in model performance), and linear regression feature importance (larger absolute values indicate greater importance); if using CART/RF/gradient boosting, entropy impurity measurements can be insightful (Venkateswarlu and Anmala, 2023).

For example, one technique that can be used to improve ML model parameter selection is the Least Absolute Shrinkage and Selection Operator (LASSO), a regression technique used for feature selection (Tibshirani, 1996). Research utilizing ML models for SWT frequency analysis at ungaged basins used the LASSO method to select explanatory variables for two ML models (Souaissi et al., 2023). The LASSO method consists of a shrinkage process where the method penalizes coefficients of regression variables by minimizing them to zero (Tibshirani, 1996). The number of coefficients set to zero depends on the adjustment parameter, which controls the severity of the penalty. Thus, the method can perform both feature selection and parameter estimation, an advantage when examining large datasets (Xu and Liang, 2021).

2.4.3 Local: single rivers, site-specific ($\leq 100\,\mathrm{km}^2$)

SWT predictions using ML have extended from the local scale to nearly continental scales over the last 24 years. One of the first studies to use a neural network to estimate SWT using an MLPNN was done by Sivri et al. (2007), who predicted monthly SWT for Firtina Creek in Türkiye, a novel approach at the time. While the MLPNN model $R^2 \sim 0.78$ was not very good, the proof of concept was a success (Sivri et al., 2007). Chenard and Caissie (2008) used eight ANNs to calculate daily and max SWT for Catamaran Brook, a small drainage basin tributary to the Miramichi River in New Brunswick, Canada, for the years 1992 to 1999. Their ANN models performed best in late summer and autumn and performed comparatively to stochastic models for the same watershed (Chenard and Caissie, 2008). In 2009, Sahoo et al. (2009) compared an ANN, multiple regression analysis, and dynamic nonlinear chaotic algorithms (Islam and Sivakumar, 2002) to estimate SWT in the Lake Tahoe watershed area in along the California–Nevada border within the US. Their ANN models included available solar radiation and air temperature, with results showing a variation of the BPNN as having the best performance (Sahoo et al., 2009).

Hadzima-Nyarko et al. (2014) used a linear regression model, a stochastic model, and variations of two NNs – MLP (six variations) and RBF (two variations) – to compute and compare SWT predictions for four stations on the river Drava, along the Croatia–Hungary border in southern central Europe. While their ANN models performed better than the linear regression and stochastic models, a comparison of their NN models found that one of their six MLPNN variations barely outperformed the RBFNN, with a difference in RMSE of 0.0126 °C, within the margin of error. The authors stated that apart from the current mean AT, the daily mean AT of the prior 2 d and classification of the day of the year (DOY) were significant controls of the daily SWT (Hadzima-Nyarko et al., 2014). Rabi et al. (2015) conducted a study using the same gage stations on the river Drava using only AT as a predictor and restricted the use of NNs to only MLP, finding that the MLPNN outperformed the linear regression approaches (Rabi et al., 2015).

Cole et al. (2014) tested a suite of models including an FFNN to predict SWT downstream of two reservoirs in the upper Delaware River, in Delaware, US. During training, the FFNN was outperformed by an Auto Regressive Integrated Moving Average (ARIMA) model and performed similarly to the physically based Heat Flux Model (HFM) (Cole et al., 2014). During testing, the FFNN, ARIMA, and HFM models performed similarly, with HFM being slightly more accurate due to its advantage as a physically based model with data availability and calibration potential (Cole et al., 2014). The authors suggest that the under- or overpredictions of the models may have been from unaddressed groundwater inputs or unaccounted for nonlinear relationships (Cole et al., 2014). Hebert et al. (2014) focused on the Catamaran Brook area (like Chenard and Caissie, 2008) and included the Little Southwest Miramichi River in New Brunswick, Canada, to conduct ANN model predictions of hourly SWT. The study considered spring through autumn with hourly data from 1998 to 2007, finding that the ANN models performed similarly to or better than deterministic and stochastic models for both areas (Hebert et al., 2014).

Piotrowski et al. (2015) examined data from two streams, one mountainous and one lowland, in a moderately cold climate of eastern Poland to model SWT using MLPNN, PUNN, ANFIS, and WNN. The ANN models were independently calibrated to find the best fits, with results showing that MLPNN and PUNN slightly outperformed ANFIS and WNN (Piotrowski et al., 2015). The study also found current AT and information on the mean, maximum, and minimum AT from 1–2 d prior to be important for improving model accuracy (Piotrowski et al., 2015). Temizyurek and Dadaser-Celik (2018) used an ANN with observations of AT, relative humidity, prior month SWT, and wind speed to predict monthly SWT at four gages on the Kızılırmak River in Türkiye. Best results were obtained from using the sigmoidal (S-shape) activation function and the scaled conjugate gradient algorithm (Møller, 1993), though the average RMSE (~ 2.3 °C) for the NN used was higher (worse) than the average calculated from this literature review where RMSE ~ 1.4 °C.

Zhu et al. (2019a, c, d) conducted several studies that used NNs to examine SWT on the river Drava, Croatia (Zhu et al., 2019a, c, d). They also examined SWT of three rivers in Switzerland and three rivers in the US (Zhu et al., 2019a, b, c). Across the studies, the MLPNN models had better performance compared to ANFIS (Zhu and

Heddam, 2019), GPR (Zhu et al., 2019a), or MLR (Zhu et al., 2019b). Qiu et al. (2020) used variations of NNs (MLP/BPNN, RBFNN, WNN, GRNN, ELMNN) to examine SWT at two stations on the Yangtze River, China, finding that the MLP/BPNN outperformed all other models when the particle swarm algorithm (PSO) was used for optimization (Qiu et al., 2020). Stream discharge and DOY were also shown to improve model accuracy. Piotrowski et al. (2020) used various MLPNN shallow (one hidden layer) structures to test the use of an approach called dropout in SWT modeling using data from six stations in Poland, Switzerland, and the US. The dropout approach can be applied to deep ANNs due to its efficiency in preventing overfitting and low computation requirements (Piotrowski et al., 2020). The study found that use of dropout and drop-connect significantly improved performance of the worst training cases. For more information on the use of dropout with shallow ANNs, we refer the reader to Piotrowski et al. (2020).

Graf and Aghelpour (2021) compared stochastic and ANN (ANFIS, RBF, GMDH) SWT models for four gages on the Warta River in Poland, finding that all models performed similarly well (R^2>97.6 %). Results showed that the stochastic and ML models performed similarly, while the stochastic models had fewer prediction errors for extreme SWT (Graf and Aghelpour, 2021). Rajesh and Rehana (2021) used several ML models (ridge regression, K-nn, RF, SVR) to predict SWT at daily, monthly, and seasonal scales for a tropical river system in India. The authors found that the monthly SWT prediction performed better than the daily or seasonal (Rajesh and Rehana, 2021). Of the ML models, the SVR was the most robust, though a data assimilation algorithm notably improved predictions (Rajesh and Rehana, 2021). Jiang et al. (2022) examined SWT under the effects of the Jinsha River cascaded reservoirs using six ML models (i.e., adaptive boosting – AB, decision tree – DT, random forest – RF, support vector regression – SVR, gradient boosting – GB, and multilayer perceptron neural network – MLPNN). The study found that day of year (DOY) was most influential in each model for SWT prediction, followed by streamflow and AT (Jiang et al., 2022). With knowledge of the influential parameters, ML model variations were tested, finding that gradient boosting and random forest provided the most accurate estimation for the training dataset and the test dataset (Jiang et al., 2022). Abdi et al. (2021) used linear regression and a deep (multi-layered) neural network (DNN) to predict hourly SWT for the Los Angeles River, finding that the DNN outperformed the linear regressions. They suggested that using a variety of ML models to predict SWT could add robustness to a study but state that training ANNs is more time-consuming than training linear regression models for minimal improved accuracy (Abdi et al., 2021).

Khosravi et al. (2023) used an exploratory data analysis (EDA) technique, a type of feature engineering that prepares the dataset for best performance with an LSTM to identify SWT predictors (discharge, water level, AT, etc.) up to 1 week in advance for a monitoring station on the central Delaware River. The authors noted that though the LSTM performed satisfactorily, future studies should compare LSTMs with CNNs or other model types and that generalizability is limited to the specific location and dataset (Khosravi et al., 2023). Majerska et al. (2024) used GPR to simulate SWT for the years 2005–2022 for the Arctic catchment Fuglebekken in Svalbard, Norway. The unique opportunity to study SWT of an unglaciated High Arctic stream regime showed an alarming warming throughout the summer where SWT increased as much as 6 °C, highlighting a strong sensitivity of the Arctic system to ongoing climate change (Majerska et al., 2024).

2.4.4 Regional, continental scale (≥ 100 km^2)

DeWeber and Wagner (2014) conducted one of the first regional ANN ensemble studies, focusing on thousands of individual streams reaches across the eastern US. They used an ensemble of 100 ANNs to estimate daily SWT with varying predictors for the 1980–2009 period, finding that daily AT, prior 7 d mean AT, and catchment area were the most important predictors (DeWeber and Wagner, 2014). In Serbia, Voza and Vuković (2018) conducted cluster analysis, PCA, and discriminant analysis for the Morava River Basin using data from 14 river stations to identify monitoring periods for sampling. With discriminant parameters identified, an MLPNN was used to predict changes in the values of the discriminant factors (see Fig. 1 of Voza and Vuković, 2018) and identify controls on the monitoring periods, finding that seasonality and geophysical characteristics were most influential (Voza and Vuković, 2018).

Rahmani et al. (2020) used 4 years of SWT data for 118 sites across the CONUS to test three LSTM models that simulated SWT, finding that the LSTM trained with streamflow observations was the most accurate, which was unsurprising. Of interest to the reader would be the inner mechanisms of the LSTM, but the study did not explicitly state what physical laws were followed by the LSTM. Instead, the authors hypothesized that the LSTM could assume internal representations of physical quantities (i.e., water depth, snowmelt, net heat flux, baseflow temperature, SWT). The authors further stated that the LSTM was dependent on a good historical data record and would not generalize well to ungaged basins. A follow-up study by Rahmani et al. (2021) used 6 years of SWT data and relevant meteorological parameters for 455 sites across the CONUS (minus California and Florida) to test LSTM models for data-scarce, dammed, and semi-ungaged basins (discharge used as input). The follow-up study showed improved performance, but the models remained limited in capturing the influence of latent contributions such as baseflow and subsurface storage. Feigl et al. (2021) tested the performance of six ML models – stepwise linear regression, RF, XGBoost, FFNNs, and two RNNs (LSTM and GRU) – using

data from 10 gages in the Austria–Germany–Switzerland region to estimate daily SWT. From the comparison, FFNNs and XGBoost were the best-performing in 8 of 10 catchments (Feigl et al., 2021). For modeling SWT in large catchments ($>96\,000\,km^2 \sim$ Danube catchment size), the RNNs performed best due to their long-term dependencies (Feigl et al., 2021). Zanoni et al. (2022) used RF, DNN, and a linear regression to predict daily SWT in the Warta River basin and compared the results with those of stochastic models. Their results found that the DNN was the most effective in capturing nonlinear relationships between drivers (i.e., SWT) and water quality parameters (Zanoni et al., 2022). On parameter influence, the analysis also found that DOY was an adequate surrogate for AT input in modeling SWT, experiencing only a slight performance reduction.

Heddam et al. (2022b) used six ML models – K-nn, LSSVM, GRNN, CCNN, RVM, and LWPR – to evaluate SWT for several of Poland's larger rivers. For each ML, three variations were created: one calibrating with only AT as input, another calibrating with AT and DOY, and a third decomposing AT using the variational mode decomposition (VMD) (Heddam et al., 2022b). For more on VMD, we refer the reader to Heddam et al. (2022a). The study found that the VMD parameters improved RMSE and MAE performance metrics for some models, but neither GRNN nor K-nn showed improvement. Heddam et al. (2022a) examined how use of the Bat algorithm optimized the extreme learning machine (Bat-ELM) neural network and how that in turn affected modeling of SWT in the Orda River in Poland. Results from the Bat-ELM were compared with MLPNN, CART, and multiple linear regression (MLR), finding the Bat-ELM outperformed MLPNN, CART, and MLR (Heddam, Kim, et al., 2022). Focusing on a region of Germany, Drainas et al. (2023) trained and tested various ANNs with different inputs for 16 small ($\leq 1\,m^3\,s^{-1}$) headwater streams, finding that the best-performing (lowest RMSE) input combination was stream-specific, suggesting that the optimal input combination cannot be generalized across streams for the region (Drainas et al., 2023). The ANN prediction accuracy of SWT was negatively affected by river length, total catchment area, and stream water level (Drainas et al., 2023). Additionally, ANN accuracy suffered when dealing with open-canopy land use types such as grasslands but improved with semi-natural and forested land cover (Drainas et al., 2023). Recently, He et al. (2024) built an LSTM framework to model water dynamics in stream segments while attempting to capture spatial and temporal dependencies. First, they created a baseline LSTM+GNN, then improved it by using graph masking and adjusting the model based on constraints (He et al., 2024). For the Delaware River Basin, the Fair-Graph model performed slightly better than the baseline with an RMSE of 1.83 vs. 1.78, respectively. For the Houston River network, the Fair-Graph model also performed slightly better than the baseline (NSE of 0.721 vs. 0.580). While the relative performance compared to the baseline was not significantly better, we anticipate that graph masking (an algorithm that incorporates spatial awareness into ANN) will play an increasingly large role in hydrologic modeling (Shen, 2018; He et al., 2024).

2.5 Decision support and climate change scenarios

In 2003, the United States Geological Survey (USGS) used an FFNN to estimate hourly SWT for a summer season in western Oregon (Risley et al., 2003). Their work used the predicted SWT to better constrain future total maximum daily loads (TMDLs) for stream management. Jeong et al. (2016) used an ANN to evaluate SWT for the Soyang River, South Korea. The goal was to couple the ANN predictions with a cyber infrastructure prototype system to deliver automated, real-time predictions using weather forecast data (Jeong et al., 2016).

Liu et al. (2018) used a hydrological model called the Variable Infiltration Capacity (VIC) model to produce estimates of AT and river-section-based variables for the Eel River Basin, Oregon, US, to be used as input data for an ANN. The study considered the AT rise from the RCP8.5 scenario to estimate future (2093–2100) daily streamflow and SWT, finding that SWT was increasingly sensitive to the proportion of baseflow in the summer (Liu et al., 2018). Topp et al. (2023) used the Delaware River Basin in the eastern US to compare two DL models: a recurrent graph convolution network (RGCN) and a temporal convolution graph model (TCGM) called Graph WaveNet. The comparison included scenarios capturing climate shifts representative of long-term projections where warm conditions or drought persisted. Considered spatiotemporally aware, the two process-guided deep learning models performed well (test RMSE of 1.64 and 1.65 °C); however, Graph WaveNet significantly outperformed RGCN in four out of five experiments where test partitions represented diverse types of unobserved environmental conditions.

Further focusing on the Delaware River Basin, Zwart et al. (2023a) used data assimilation and an LSTM to generate 1 and 7 d forecasts of daily maximum SWT for the purpose of aiding reservoir managers in decisions about when to release water to cool streams. Following up on this study was Zwart et al. (2023b), who used an LSTM and an RGCN, both with and without data assimilation, to generate 7 d forecasts of daily maximum SWT for monitored and unmonitored locations in the Delaware River Basin, finding that the RGCN with data assimilation performed best for ungaged locations and at higher SWT, which is important for reservoir operators to be aware of while drafting release schedules.

Rehana and Rajesh (2023) used a standalone LSTM, a WT-LSTM, and a k-nearest neighbor (K-nn) bootstrap resampling algorithm with LSTM to assess climate change impacts on SWT using downscaled projections of AT with RCPs 4.5 and 8.5 for seven polluted river catchments in India. Comparing the coupled models and the physically based

Air2stream model, they found the K-nn coupled with LSTM to be the best-performing in terms of effectively predicting SWT at the monthly timescale. Considering the RCP scenarios, the predicted SWT increase for 2071–2100 for the rivers in India ranged from 3.0–4.7 °C.

3 Model evaluation metrics

The second part of this review compiles ML performance evaluation metrics as they pertain to SWT modeling and prediction and considers the commonly used metrics to suggest guidelines for easier comparison of ML performance across SWT studies. We considered journal articles from 2000–2024 that used ML to evaluate, predict, or forecast SWT and examine what model performance metrics were used. Performance metrics can be calculated during model calibration, testing, and (or) validation to compute a single value that denotes the agreeableness between simulated and observed data.

For this literature review, all journals examined used at least one metric to evaluate model performance, with two or more metrics used by >84 % of studies published in or after the year 2019. For review, the quantitative statistics were split into three categories: standard regression, dimensionless, and error index (Moriasi et al., 2007). Standard regression statistics (Pearson's r, R^2) are ideal for examining the strength of the linear relationship between model simulations or predictions and the observed or measured data. Dimensionless techniques (NSE, KGE) provide a relative assessment of model performance but due to their interpretational difficulty (Legates and McCabe, 1999) have been less commonly used. In contrast, error indices (RMSE, MAE) quantify the error in terms of the units of the data (i.e., °C) considered.

3.1 Model performance metrics: standard regression

The most basic statistics (slope, y-intercept mean, median, standard deviation) continue to be used in part due to their simplicity and ease of interpretability. These statistics are useful for preliminary examinations, where the assumption is that measured and simulated values are linearly related, and all the variance of error is contained within the predictions or simulations, whilst the observations are free of error. Unfortunately, observations are rarely error-free, and datasets are nonlinear, highlighting a need for using a diverse set of statistics (Helsel and Hirsch, 2002). One such set of statistics commonly used for standard regressions are called the correlation coefficients – Kendall's tau, Spearman's rho, and Pearson's r.

Pearson's r, also known as the correlation coefficient, is used to determine the strength and direction (i.e., positive, negative) of a simple linear relationship (Helsel and Hirsch, 2002). Values of r range from -1 to $+1$, where $r<0$ indicates a negative correlation and $r>0$ indicates a positive correlation (Legates and McCabe, 1999). The square of r is denoted as r^2, known as the square of the correlation coefficient, with values of r^2 ranging from 0 to 1. The r^2 metric is commonly used in simple linear regression to assess the goodness of fit by measuring the fraction of the variance in one variable (i.e., observations) that can be explained by the other variable (i.e., predictors). The metric r^2 tends to be confused with R^2, the latter of which is a statistical measure that represents the proportion of variance explained by the independent variable(s) in a multiple linear regression model (Helsel and Hirsch, 2002). Part of the confusion may be related to the fact that R^2 shares the same range of 0 to 1, with $R^2 = 1$ indicating that the model can explain all the variance, and vice versa. We note that while both r^2 and R^2 share similarities in that they measure the proportion of variance, R^2 is more commonly used for multiple linear regression context, while r^2 is best suited for simple linear regressions. To reduce confusion, we strongly suggest that r, r^2, and R^2 always be reported together (even if as a supplement to a manuscript) to characterize goodness of fit.

In contrast to the linear regression metrics, Spearman's rank correlation coefficient, rho (ρ), is a nonparametric rank-sum test useful for analyzing non-normally distributed data and nonlinear monotonic relationships (Helsel and Hirsch, 2002). The data are ranked on a range from -1 to $+1$, where $\rho = 0$ indicates no association and $\rho = -1$ or $+1$ suggests a perfect monotonic relationship. By ranking the data, Spearman's correlation coefficient quantifies monotonic relationships between two variables (converts nonlinear monotonic relationships to linear relationships), allowing ρ to be robust against outliers (Helsel and Hirsch, 2002).

3.2 Model performance metrics: error indices

The mean absolute error (MAE), mean square error (MSE), and root mean squared error (RMSE) are popular error indices used to assess model performance. The equations for MAE, MSE, and RMSE are as follows.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|O_i - P_i| \tag{1}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2 \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{N}} \tag{3}$$

For the equations, N is the number of samples, O_i is the observed SWT, and P_i is the predicted SWT at time i. The MAE computes the average magnitude of the errors in a set of predicted values to obtain the average absolute difference between the predicted P_i and the observed O_i. In contrast to MAE, the MSE squares the error terms, resulting in the squared average difference between the predicted and observed values. The resultant MSE is not in the same units

as the value of interest, making it difficult to interpret. As the square root of the MSE, RMSE provides an error index in the unit of the data (Legates and McCabe, 1999). However, both the MSE and RMSE are more sensitive to outliers and less robust than MAE.

$$\mathrm{PBIAS} = 100 \cdot \frac{\sum_{i=1}^{N}(P_i - O_i)}{\sum_{i=1}^{N} O_i} \tag{4}$$

Another error index used in SWT modeling is called the percent bias (PBIAS) index. PBIAS computes the average tendency of model predictions to be greater or smaller than the observations or measurements (Gupta et al., 1999). A PBIAS value of 0 is best, and low-magnitude values (closer to 0) denote stronger model accuracy. Positive PBIAS values suggest model underestimation, while negative PBIAS values suggest model overestimation (Moriasi et al., 2007).

3.3 Model performance metrics: dimensionless

The Nash–Sutcliffe efficiency (NSE, also called NSC, NS, or NASH) is a "goodness-of-fit" criterion that describes the predictive power of a model. Mathematically, the NSE is a normalized statistic that computes the relative magnitude of the variance of the residuals compared to the variance of the measured or observed data (Nash and Sutcliffe, 1970). Visually, the NSE shows how well the observed versus simulated data fit on a 1 : 1 line.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}, \tag{5}$$

where $\overline{O}$ is the average value of O_i. To compute the Kling–Gupta efficiency (KGE),

$$\mathrm{KGE} = 1 - \mathrm{ED},$$

$$\mathrm{ED} = \sqrt{(r-1)^2 + \left(\frac{\sigma_P}{\sigma_O} - 1\right)^2 + \left(\frac{\mu_P}{\mu_O} - 1\right)^2}, \tag{6}$$

where r is the linear correlation coefficient between predictions and observations. The purpose of the KGE metric is to reach a balance between optimal conditions of modeled and observed quantities being perfectly correlated (i.e., $r = 1$), with the same variance ($\sigma_p/\sigma_o = 1$) and minimizing model output bias ($\mu_p/\mu_o = 1$). The Kling–Gupta efficiency (KGE) is based on a decomposition of NSE into separate components (correlation, variability bias, and mean bias) and tries to improve on NSE weaknesses (Knoben et al., 2019). Like NSE, KGE = 1 is a perfect fit between model simulations or predictions and observations or measurements. However, NSE and KGE values cannot be directly compared because each metric is influenced by the coefficient of variation of the observed time series (Knoben et al., 2019).

The Willmott index of agreement, d, ranging from 0 to 1, is defined as a standardized measure of model prediction error, where a value of 1 is perfect agreement between measured and predicted values, and a value of 0 indicates no agreement.

$$d = 1 - \frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} \tag{7}$$

The Akaike information criterion (AIC) is a selection method used to compare several models to find the best approximating model for the dataset of interest (Akaike et al., 1973; Banks and Joyner, 2017; Portet, 2020). For details on the mathematical derivation and application of AIC, please see Banks and Joyner (2017), Portet (2020), and Piotrowski et al. (2021). The AIC equation version shown was developed for the least-squares approach (Anderson and Burnham, 2004):

$$\mathrm{AIC} = N \cdot \ln(\mathrm{MSE}) + 2 \cdot K, \tag{8}$$

where N is the number of samples, K is the number of model parameters + 1, and MSE is obtained by the model, for the respective dataset, per stream (Piotrowski et al., 2021). The Bayesian information criterion (BIC) was developed for studies where model errors are assumed to follow a Gaussian distribution (Faraway and Chatfield, 1998; Piotrowski et al., 2021). For other versions of BIC, please see Faraway and Chatfield (1998).

$$\mathrm{BIC} = N \cdot \ln(\mathrm{MSE}) + K\ln(N) \tag{9}$$

Unlike other performance metrics, the AIC and BIC are unique in their ability to penalize the number of parameters used by a model, thus favoring more parsimonious models. For both the AIC and BIC, lower values of the criterion point to a better model (Piotrowski et al., 2021).

3.4 Performance metrics for most-cited ML statistics

Reviewing ML studies focused on SWT modeling (Table S1, S2), the most-cited performance metrics were RMSE (45 citations), NSE (25), MAE (18), and R^2 (17). Having reviewed the literature and in agreement with previous published recommendations (Moriasi et al., 2007), we recommend that a combination of standard regression (i.e., r, r^2, R^2), dimensionless (i.e., NSE), and error index statistics (i.e., RMSE, MAE, PBIAS) be used for model evaluation and reported together in future publications. As part of our efforts to propose guidelines for easier comparison of ML performance across SWT studies, we identified the range in reported values for these four most-cited metrics and show the spread of values in the training and calibration as well as the testing and validation phases in box plot form.

We begin with the standard regression and dimensionless statistics, R^2 and NSE, both of which have an optimal value of 1. Figure 2 shows the median R^2 per ML model per model phase for the cited publications. For example, Foreman et al. (2001) used an ANN to model SWT in the Fraser

Watershed in British Columbia, Canada. Their model estimated 1995–1998 tributary and headwater temperatures and reported a median R^2 (Fig. 2) of 0.93 for the training and calibration phase. Over the review period, the R^2 range (2001–2024) was 0.65–1.00. We note that for process-based modeling, acceptable R^2 values start around $R^2 \sim 0.50$ (Moriasi et al., 2007). In stark contrast, ML models published between 2000–2024 exhibited significantly higher R^2 values, with a median of $R^2 \sim 0.93$ across 17 studies (Fig. 2).

Unlike the R^2 metric, NSE was not used as a metric in ML studies of SWT between 2000 and 2010 (Fig. 3). The first ML study to use NSE was St. Hilaire et al. (2011) to analyze SWT in Catamaran Brook, a small catchment in New Brunswick, Canada. Fig. 3 shows that the NSE range reported by studies using ML for SWT was between 0.25–1.00 over the reviewed period (2000–2024). Like R^2, NSE published values are high compared to traditional models (Moriasi et al., 2007, 2015), with a median NSE of 0.93 across 25 studies (Fig. 3). Overall, these complementary metrics should always be reported together as they provide a broader evaluation of model performance; i.e., NSE measures a model's predictive skill and error variance, while R^2 assesses how well the model explains the variability of the data.

Figure 4 shows the median RMSE (°C) and Fig. 5 shows the median MAE (°C) per ML model per model phase for each publication. RMSE (°C) and MAE (°C) are popular error indices used in model evaluation because the metrics show error in the units of the data of interest (i.e., °C), which helps analysis of the results. RMSE and MAE values equal to 0 are a perfect fit. Over the review period, median RMSE (Fig. 4) ranged from 0.0002–3.50 °C. The median RMSE was 1.35 °C across 45 studies (Fig. 4). Figure 5 shows that between 2000–2012, MAE was not used as a metric in ML studies of SWT. The first ML study to use MAE for SWT modeling was Grbić et al. (2013), where the Gaussian process regression (GPR) ML approach was compared with field observations of SWT from the river Drava in Croatia to assess the feasibility of model development in SWT prediction. In contrast to RMSE, the MAE range (Fig. 5) was 0.14–2.19 °C. The median MAE overall was 1.09 °C across 18 studies (Fig. 5).

3.5 Spatial scale

We examined the data for the possible influence of spatial scale on the most-cited performance metric, RMSE, by grouping publications into two spatial categories: local, which included studies that focused on point to plot, specific sites, and small watersheds less than $\sim 100\,\text{km}^2$ in area (about the size of a HUC-08), and regional, which included everything over $\sim 100\,\text{km}^2$ in area. For this analysis, all RMSE values reported by publications were compiled into a table (not shown) and classified as belonging to either the local/watershed or regional/CONUS scale. A comparison of the data found that the average RMSE was similar for the local (~ 1.52 °C) and regional (~ 1.55 °C) categories. The median local RMSE was slightly better than the regional RMSE (~ 0.04 °C) but arguably within a standard of error. The local/watershed category had a higher maximum and minimum RMSE than those reported for the regional category. Overall, neither category appeared significantly better or worse than the other.

Table 1. Average, median, maximum, and minimum RMSE (°C) for studies grouped by local/watershed and regional/CONUS spatial scales.

	Local/watershed ($<100\,\text{km}^2$ area)	Regional/CONUS ($>100\,\text{km}^2$ area)
Number of data points	900	1369
Average	1.52	1.55
Median	1.38	1.42
Maximum	5.170	4.387
Minimum	0.038	0.0002

3.6 Temporal scale

Across studies, there was large variability in the focus of temporal scales and use. For example, some studies used data collected at 5–15 min intervals to simulate SWT at daily or weekly intervals for an abbreviated period (Risley et al., 2003; Hong and Bhamidimarri, 2012). Other studies used data collected at hourly, daily, weekly, or monthly intervals (Foreman et al., 2001; Sivri et al., 2007; Temizyurek and Dadaser-Celik, 2018) for periods of record spanning weeks (Lu and Ma, 2020; Abdi et al., 2021) to several decades (Cole et al., 2014; Weierbach et al., 2022; Heddam et al., 2022a; Topp et al., 2023; Rehana and Rajesh, 2023) to simulate SWT. Concurrently, output for studies was then provided at resolutions ranging from hourly to monthly periods for the past, present, or future. Given the use of study-specific temporal outputs and the limited amount of reported peer-reviewed model performance data at the temporal scales used by researchers, it was difficult to conduct statistical comparisons for temporal scales, so they are not further discussed in this review. We strongly suggest to researchers that metrics be made available at the temporal scale of interest (and not just for the overall model) in appendices or supplementary information to encourage more comparison across studies.

4 Discussion

4.1 Model evaluation ratings

From our review of RMSE, R^2, NSE, and MAE, we compiled ratings for ML performance metrics that should be used to for cross-comparison across SWT studies. From Table S2, we note that there was not a consistent way of reporting train-

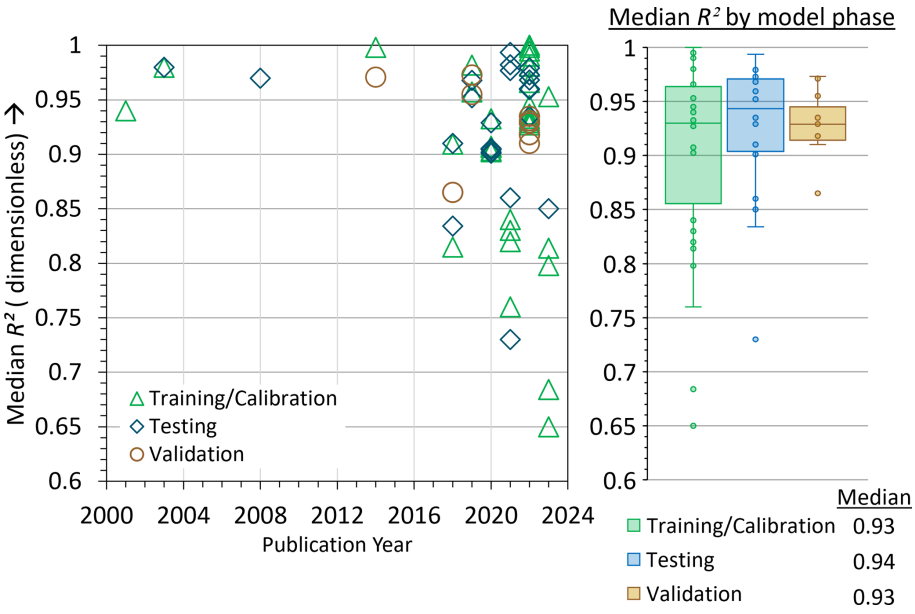

Figure 2. Median R^2 (dimensionless) values from published literature for training and calibration, testing, and validation phases of model evaluation.

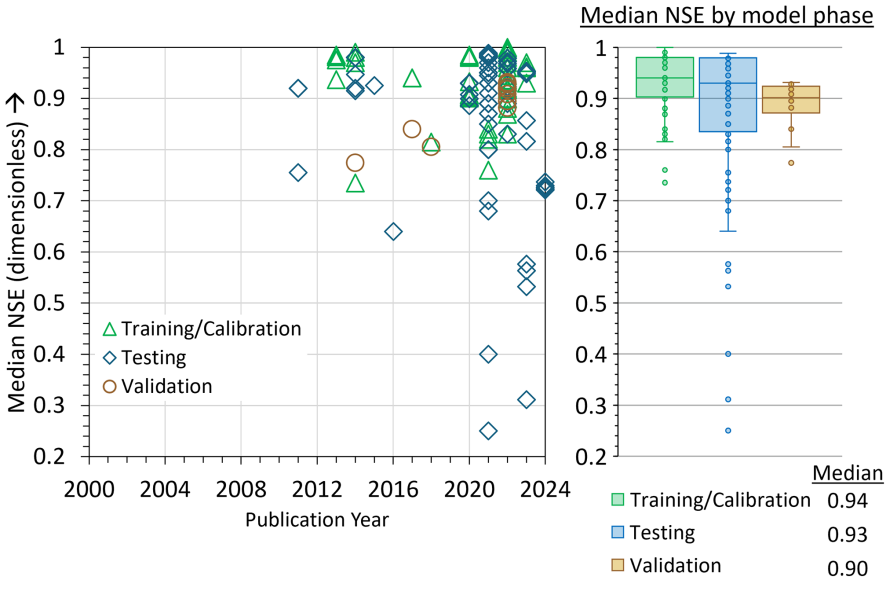

Figure 3. Median NSE (dimensionless) values from published literature for training and calibration, testing, and validation phases of model evaluation.

ing, validation, and testing percentages; for example, some studies only reported performance metrics for one modeling phase (i.e., training), while others used "testing" and "validation" interchangeably, which could affect interpretation of model performance (Laanaya et al., 2017; Voza and Vuković, 2018; Hani et al., 2023). Additionally, others stated information not by percentages but by years (i.e., training 2 years, testing 1 year, validation 1 year), which can make comparisons challenging. Despite all the different ways that researchers chose to compile performance metrics, most models had strong metrics, as shown by our calculated ratings for performance metrics shown in Table 2.

We posit that the definitions of satisfactory, good, and very good be updated to reflect the inherent capability of an ML algorithm to fit the input data more successfully than other model types, such as statistical models and process-based

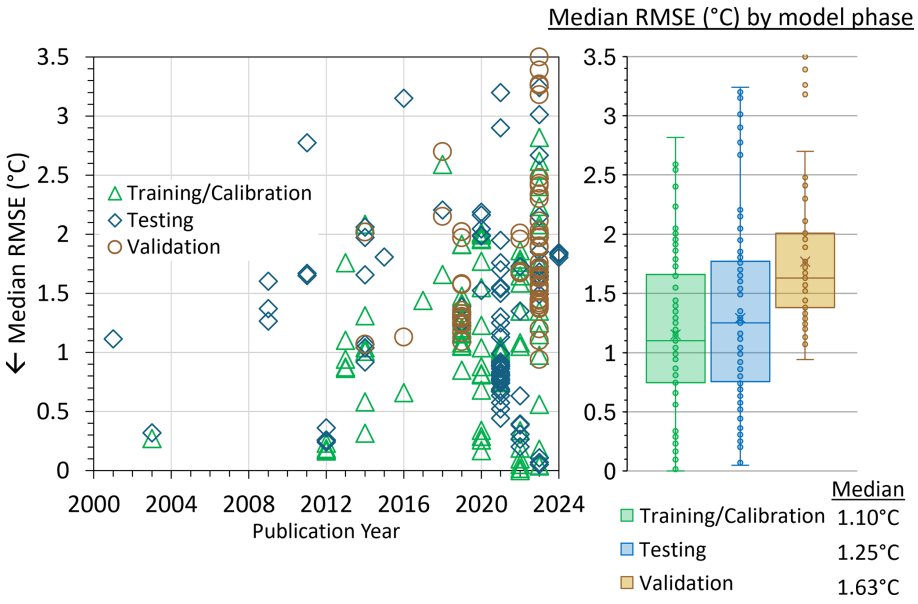

Figure 4. Median RMSE (°C) values from published literature for training and calibration, testing, and validation phases of model evaluation.

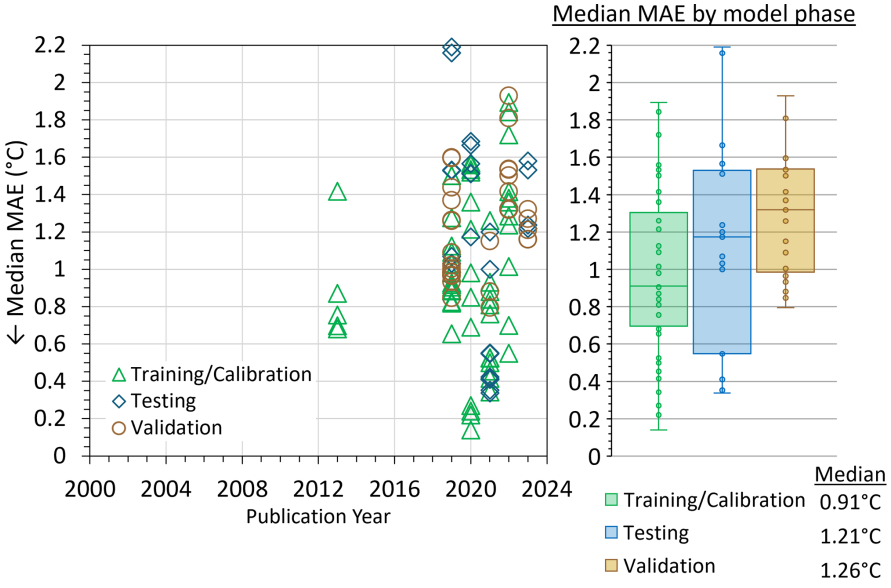

Figure 5. Median MAE (°C) values from published literature for training and calibration, testing, and validation phases of model evaluation.

models. For example, R^2 values from ML-SWT studies that may appear to be very good, such as $R^2 \sim 0.91$, should be considered satisfactory given the context of the performance metrics published in the ML-SWT literature. In Table 2, the very good and unsatisfactory ranges were calculated from the box plots by identifying the two-thirds distance from the upper (or lower) quartile to the respective extreme whisker. This calculation identifies the $\sim 8\,\%$ of the data that is relatively closest to the minimum or maximum values of the box plots, indicating a very good or unsatisfactory value. For Table 2, the separation between the satisfactory range and the good range was denoted as the halfway value between very good and unsatisfactory. The purpose of these guidelines is to serve as a reference for SWT studies looking to understand and consider ML performance relative to other SWT-ML studies.

4.2 ML data requirements vs. data availability

While, in recent years, access to hydrologic data has improved (CUAHSI Inc., 2024), data remain scarce for many hydrologic applications including SWT research, particu-

Table 2. Suggested ratings for performance metrics (median) using metrics published by ML studies examining SWT.

	R^2			NSE		
Rating	Training	Testing	Validation	Training	Testing	Validation
Very good (>)	0.99	0.99	0.96	0.99	0.98	0.93
Good (range)	0.89–0.99	0.92–0.99	0.94–0.96	0.92–0.99	0.84–0.98	0.88–0.93
Satisfactory (range)	0.79–0.89	0.86–0.92	0.91–0.94	0.85–0.92	0.70–0.84	0.83–0.88
Unsatisfactory (<)	0.79	0.86	0.91	0.85	0.70	0.83
	RMSE (°C)			MAE (°C)		
Rating	Training	Testing	Validation	Training	Testing	Validation
Very good (>)	0.25	0.26	1.15	0.33	0.42	0.86
Good (range)	1.34–0.25	1.51–0.26	1.80–1.15	1.01–0.33	1.19–0.42	1.32–0.86
Satisfactory (range)	2.43–1.34	2.77–1.51	2.45–1.80	1.70–1.01	1.97–1.19	1.79–1.32
Unsatisfactory (<)	2.43	2.77	2.45	1.70	1.97	1.79

larly because continual project management and funding to place and maintain stream temperature sensors can be expensive and/or time-consuming to undertake. As a result, in the 21st century, the scarcity of data remains a large impediment for the application of machine learning in SWT modeling. What is more, the question of data quantity (how many data do you have?) versus quality (how many diverse data are needed?) continues to hinder ML use in hydrologic applications. Xu and Liang (2021) make the excellent point that 1 year of streamflow data (can swap for stream temperature) at 15 min intervals equal about $\sim 35\,000$ points, which may seem extensive but is unlikely to be enough to properly train an ML model due to autocorrelation and limited exposure to diverse types of data that are naturally encountered with a longer time series (Xu and Liang, 2021). For example, machine learning models may only predict flood volumes they have previously seen (Kratzert et al., 2019). While data requirements for ML remain high, there are some strategies that researchers have used to alleviate this impact.

One strategy that hydrologists in other fields have used to tackle this problem is data augmentation, which can be applied spatially or temporally to create new training examples that the ML model can learn from. Spatial augmentation can be done by means of interpolation methods, i.e., kriging or distance weighting to create new data points or by generating synthetic data based on expected physical patterns to fill gaps in data coverage (Baydaroğlu and Demir, 2024). Temporal data augmentation can be done by shifting, scaling, or adding noise to existing time series to create new training examples (Skoulikaris et al., 2022). Alternatively, and not a new idea, is to use the statistical technique known as seasonal decomposition, which breaks down a time series into its main components, i.e., the trend, seasonal patterns, and residual components (Apaydin et al., 2021; He et al., 2024). These can then be recombined to generate new data and train the model for improved accuracy (Apaydin et al., 2021). In addition to data augmentations, data requirements can be alleviated by considering the help of unsupervised transfer learning, i.e., use pre-trained models on similar tasks to reduce the amount of data needed for training, or semi-supervised learning, such as few shot learning, i.e., combine a small percent of labeled data with a larger percent of unlabeled data to improve model performance (Yang et al., 2023). By implementing these strategies, researchers in other hydrologic fields have shown that models can be improved with fewer data, strategies that are likely transferable to SWT research.

4.3 ML use for knowledge discovery

It has been suggested that the increasingly prominent use of ML for hydrological predictions points to a paradigm shift, one where the adoption of ML in most if not all future physical hydrologic modeling appears certain (Xu and Liang, 2021; Varadharajan et al., 2022). As physical scientists try to stay afloat in a sea of ML algorithm options and processes, there is a critical need to examine how "newer" tools such as ML are improving our understanding of the natural world. Our review finds that ML studies examining SWT have been conducted from a computational perspective, one with a focus on comparing techniques and performance as opposed to explaining the nature of SWT dynamics or influencing processes.

While it is understandable that not every ML-SWT paper aims to explain physical processes, the SWT community should agree on a baseline of tests that all ML-SWT models undergo to assess model robustness and transferability. Specifically, we urge use of TUURTs (temporal, unseen, ungaged region tests) for future ML-SWT models as a helpful step towards better modeling practices, increased model transparency, and robustness (see Sect. 2.4.1, Fig. 1). From a computational perspective, the use of ML in SWT modeling has led to improvements in pattern identification (i.e., release of water from reservoirs; see Jiang et al., 2022) and examination of climate events (i.e., extreme droughts; see Qiu et al., 2020), with the aid of observations and remote sens-

ing data. The use of ML for estimating hydrologic variables (i.e., precipitation, snow water equivalent, and evapotranspiration) and approximating hydrologic processes (i.e., runoff generation) has also become increasingly common due to the ML's ability to use many inputs without the bounds of preexisting relationships (Xu and Liang, 2021). In addition, hybridizations that couple ML models (i.e., WT-LSTM or K-nn with LSTM) (Rehana and Rajesh, 2023) or couple ML with process-based models (i.e., SNTEMP-LSTM) (Rahmani et al., 2023) show potential for outperforming extensively calibrated hydrologic models, especially where physical constraints can be introduced (Rahmani et al., 2023).

Recent studies (Rahmani et al., 2023; Wade et al., 2023) have tried to infer drivers of SWT regimes by accounting for some level of physics. Compounding the challenge of applying physical laws without negatively affecting the performance of an ML model is the problem that the ML model itself is not immune to the difficulties met by statistical and process-based models such as data uncertainties, parameter uncertainties, and equifinality (Beven, 2020; Varadharajan et al., 2022). These uncertainties, coupled with the alarming trend of consistently high marks of the performance metrics discussed here, point to an imperative need to reevaluate how best to use ML in a manner that addresses knowledge gaps of physical systems instead of perfecting performance that is unlikely to be insightful for physical processes and trends. Our review of the literature and analysis of the performance data agree with the discussion by Beven (2020), who examined the future of hydrological sciences with ML and posed several important questions regarding better use of ML models for scientific inquiry.

4.4 Future directions of SWT modeling

The utility of ML in hydrologic modeling has advanced significantly, with interest seemingly growing exponentially (Nearing et al., 2021). With the novelty of ML, it is easy to over-value model performance and ignore the physics of the system, but with several decades of ML experience, we advocate for the necessity of purposefully using ML to address physically meaningful questions and not just creating ML for the sake of creating. Given this, Varadharajan et al. (2022) laid out an excellent discussion on opportunities for advancement of ML in water quality modeling; see Sect. 3 of Varadharajan et al. (2022). Here we highlight some of the questions from Varadharajan et al. (2022) that can be considered in the context of what objectives the SWT community should be using in the ML era, namely the following. (1) How do we use physical knowledge (regarding heat exchange equations, radiation influence) to improve models and process understanding? Rahmani et al. (2023) coupled NNs with the physical knowledge from SNTEMP, a one-dimensional stream temperature model that calculates the transfer of energy to or from a stream segment by either heat flux equations or advection, but found that even with SNTEMP, their flexible NNs exhibited substantial variance in prediction and needed to be constrained by further multi-dimensional assessments (Rahmani et al., 2023). In short, if our use of physics in machine learning makes our models worse, we should understand why.

A second question that needs addressing is (2) how do we deal with predictive uncertainty in ML used for SWT modeling? According to Moriasi et al. (2007), uncertainty analysis is the process of quantifying the level of confidence in any given model output based on five guidelines: (1) the quality and number of observations (data), (2) the lack of observations due to poor or limited field monitoring, (3) the lack of knowledge of physical processes or operational procedures, i.e., instrumentation, (4) the approximation of our mathematical equations, and (5) the robustness of model sensitivity analysis and calibration. For example, in rainfall-runoff modeling, researchers have proposed benchmarking to examine uncertainty predictions of ML rainfall-runoff modeling (Klotz et al., 2022). For stream temperature modeling, researchers have attempted to address the role of uncertainty in deep learning model (RGCN, LSTM) predictions using the Monte Carlo dropout (Zwart et al., 2023b) and a unimodal mixture density network approach (Zwart et al., 2023a).

Other questions that SWT-ML studies should consider are the following. (3) How do we make ML models generalize better, specifically with regards to ungaged basins? And (4) how can ML models be improved to predict extremes? As ML models advance to use satellite data, include more sensor networks, and/or couple with climate models, there is a logical next step toward creating generalizable models that can account for extremes. The challenge of prediction in ungaged basins in SWT modeling has been explored for at least a decade by process-based (Dugdale et al., 2017) and statistically based (Gallice et al., 2015, Isaak et al., 2017; Wanders et al., 2019; Siegel et al., 2023) models. Unfortunately, process-based models continue to be limited by data requirements and memory or processing and programming impediments (Dugdale et al., 2017; Ouellet et al., 2020), while statistically based models struggle to account for changing physical conditions (Benyahya et al., 2007; Arismendi et al., 2014; Lee et al., 2020). Physics-derived statistically based models have been applied in ungaged regions (Gallice et al., 2015), but models tend to be region-specific and not generalizable. We posit that a future direction of ML models is to expand on their ability to learn, identify, and mimic the complexity needed to improve SWT predictions for ungaged basins. To date, researchers have used ML to model SWT for partially ungaged (i.e., discharge used as input) regions across the CONUS (Rahmani et al., 2020, 2021), though limitations persist in hydrologically complex and critical regions in the west (CA) and southeast (FL). Recently, a satellite remote sensing paper used RF to model monthly stream temperature across the CONUS and tested for temporal (walk-forward validation), unseen, and "true" ungaged regions (Philippus et al., 2024a). Given community-wide mod-

eling interest expanding from SWT prediction to forecasting (Zhu and Piotrowski, 2020; Jiang et al., 2022; Zwart et al., 2023a), ML use could prove essential in capturing unknown, complex SWT patterns in space and time (Philippus et al., 2024b) and with shifting baselines. With regards to ML models such as LSTMs predicting extremes, a limitation that must be addressed is that they generally only make predictions within the bounds of their training data (Kratzert et al., 2019) though researchers are looking to improve on this by using ML hybridizations (Rozos, 2023). Overall, there is promising work in the community towards creating ML models for SWT that generalize better and/or are more robust for predictions of extremes.

Finally, (5) how can we build ML models such that they are seen as trustworthy and interpretable by the hydrologic community? To answer this question, we must address a technical barrier (black-box issues, data limitations, model uncertainty) and a social barrier (i.e., educated skepticism of ML due to novelty, little understanding of computer science basics and/or coding experience). If we are to incorporate ML into decision-making processes, it makes sense that ML must be transparent and understandable to more than just computer or data scientists (Varadharajan et al., 2022). For example, Topp et al. (2023) recently used explainable AI to elucidate how ML architectures affected the SWT model's spatial and temporal dependencies and how that in turn affected the model's accuracy. Addressing this technical barrier can also be done by improving access to data, which has seen remarkable progress thanks to web repositories such as CUAHSI's NSF-funded HydroShare (CUAHSI Inc., 2025) and GitHub (GitHub Inc., 2024). In the United States, data access to state and locally based data remains limited and should be addressed. In terms of the social barrier, education about ML and ML use is key.

Societal interest in ML has thankfully also led to a plethora of educational resources and ML walk-through videos and tutorials in Tensorflow (Abadi et al., 2016), PyTorch (Paszke et al., 2019), and Google Colab (Bisong, 2019). With the speed at which ML use is evolving, short communication pieces (Lapuschkin et al., 2019) and opinion pieces (Kratzert et al., 2024) with clear examples about an ML issue and practical solutions will also help make ML challenges more transparent and therefore accessible to the hydrologic community at large.

5 Conclusions

While initial examination of SWT began with statistical and process-based modeling many decades ago, there is now a strong interest within the hydrology community to use ML across the board to further our understanding of hydrologic causes and effects. Indeed, extensive progress has been made in using ML for SWT modeling solely in the last quarter century (2000–2024). As discussed in this review, applications of ML in SWT modeling have ranged from the local to the continental scale, as well as from the short-term period of hours to the longer-term period of decades.

In this review, we examined published literature that used ML for SWT modeling and provided a range of background information on the ML models used in these studies. Additionally, we compiled reported ML performance metrics and compare those most cited – RMSE, R^2, NSE, and MSE. We find that ML performance metrics surpass all our preconceived notions of what makes a very good vs. satisfactory model. We argue that as a scientific community, we need to redefine model success in the face of ML's consistently robust performance or, at the very least, hold ML to additional standards when comparing ML to physically based and statistically based models. To aid in redefining standards, we introduce updated designations (for ML studies only) of very good, good, and satisfactory performance metrics as derived from the literature. In addition to leveling the playing field when comparing ML results to process-based and statistically based models, we assert that raising the performance bar could also strengthen user confidence in ML models to the point that their consideration in decision-relevant predictions becomes more widely trusted and accepted.

Finally, our review finds that the increased accessibility to ML and its use in SWT modeling has yet to lead to better physical understanding of SWT causes and effects. Over the past 25 years, the focus on desired accuracy and performance metrics has overpowered much-needed trade-offs that earlier models of the 20th century considered, such as process complexity (scale, heterogeneity, generalizability), knowledge discovery, timeliness, and basic public understanding. Given our knowledge that most ML models consistently perform at a higher level, we believe it is time to take a step back and purposefully consider more thoughtful creation and purposefulness of ML models for the goal of decision-relevant predictions that include risk mitigation, water resource planning, and process understanding of stream water temperature influencers and effects.

Appendix A: Traditional artificial neural networks, detail, and descriptions

ANNs are composed of networks of interconnected neurons, also called nodes or units. The network architecture of a commonly used ANN, the feed-forward NN (FFNN), can be described as a three-layered (or more) network of connected neurons, organized from left to right, where the input layer is the first layer, the center layer (could be one or more) is "hidden", and the last layer is the output layer (Risley et al., 2003). Multi-layer perceptron NNs (MLPNNs) fall under the umbrella of FFNNs. In the FFNN architecture, the first (leftmost) layer creates input signals from a dataset. In the hidden layer, the neurons process the input signals using an activation function (i.e., step, sigmoid-shaped, hyperbolic

tangent, etc.) to calculate a hidden-layer output from the input, the hidden-layer weight, and the hidden-layer bias (Hinton, 1992). The hidden-layer weight is defined as the strength of the influence of neurons on each other and is modifiable (Hinton, 1992). For example, a connection between neurons A and B may be stronger (weight ~ 0.5) than a connection between neurons B and C (weight ~ 0.1). This weight can be adjusted, or "fine-tuned", to minimize errors. Depending on the output of the activation function, the output signals may be transmitted to other neurons in the network, eventually supplying output from the hidden layer to the final layer, which computes the final output using a summation function (Hinton, 1992).

The back-propagation (BP) learning algorithm (Hinton, 1992) is one of the more popular techniques that iteratively adjusts model weights and bias terms in a neural network. First, the FFNN is trained on a labeled and categorized dataset, called the "training" dataset. The BP algorithm then iteratively adjusts weights in the NN based on the calculated error between the predicted output and the actual output, allowing the NN to find underlying patterns or possible relationships in the data (Hinton, 1992). However, use of the BP learning algorithm for FFNNs can be time-consuming in terms of training and calibration (Huang et al., 2006). Huang et al. (2006) proposed an alternative learning algorithm called extreme learning machine (ELM) for shallow-layer BP FFNNs (also abbreviated to BPNNs). The ELM algorithm optimizes training by randomly choosing hidden nodes and analytically finding output weights (Huang et al., 2006). In a comparison study, ELM generally outperformed the BP algorithm in terms of learning and performance (Huang et al., 2006).

Another kind of ANN with a similar three-layer structure is the radial basis function NN (RBFNN). However, the RBFNN distinction is that only one hidden layer is used and that the width of connections and centers (distance between inputs and weights) must be calculated prior to adjusting weights (Musavi et al., 1992; Buhmann, 2000). We refer the reader to Musavi et al. (1992) and Buhmann (2000) for more detail on RBFNN. The Cascade Correlation Neural Network (CCNN), introduced by Fahlman and Lebiere (1990), proved to be much faster than back propagation (Fahlman and Lebiere, 1990). The CCNN was created with a cascade architecture, where hidden neurons are added to the network one at a time and remain unchanged; i.e., the input weights are frozen, allowing the neuron to become a feature detector in the network, capable of either producing outputs or creating other complex feature detectors (Fahlman and Lebiere, 1990). For more detail on CCNN, we refer the reader to Fahlman and Lebiere (1990).

General regression NN is a Bayesian type of FFNN based on kernel regression networks (Specht, 1991). Unlike MLPNN, GRNN does not need an iterative training procedure like back propagation. One of the advantages of GRNN with increasingly larger datasets is that it is consistent in forcing the estimation error to approach zero with only minor restrictions on the function (Specht, 1991). GRNN also differs from RBFNN in the method used to decide the weights of the hidden-layer nodes. GRNN does not train the weights as RBFNN does; instead, GRNN provides the target value (to the node weight) by considering the input training dataset and the related output (Specht, 1991). The product-unit NN (PUNN) uses product units (in contrast to the summation units used by MLPNN) to compute the product of its inputs, each raised to a variable power (Janson and Frenzel, 1993). While less used in SWT modeling, PUNNs have garnered interest due to their capacity for implementing higher-order functions (Martínez-Estudillo et al., 2006) and advantage of requiring fewer parameters for optimization when considering the same number of input nodes, hidden nodes, and output nodes (Piotrowski et al., 2015). For more on PUNN, we refer the reader to Janson and Frenzel (1993) and Martínez-Estudillo et al. (2006). A lesser known but used ANN is the Group Method of Data Handling (GMDH), created by Russian scientist Ivakhnenko in the late 1960s for the purpose of using inductive learning methods for modeling complex, nonlinear systems without the bias of the user (Ivakhnenko, 1970). Although not initially described as an ANN, GMDH is a polynomial NN. GMDH initiates only with input neurons; then during the training processes, neurons are "self-organized" to optimize the network with the help of "control data" to stop the training process when overfitting occurs (Ivakhnenko, 1970; Ivakhnenko and Ivakhnenko, 1995; Graf and Aghelpour, 2021). For more information on GMDH, we refer the reader to Ivakhnenko (1970) and Ivakhnenko and Ivakhnenko (1995).

Adaptive-network-based fuzzy inference systems (ANFISs) are types of NNs using fuzzy inference, initially proposed by Jang (1993). Fuzzy inference systems first interpret values in the input vector, and then (following a set of rules) the system assigns values to the output vector (Kalogirou, 2023). ANFIS uses a combination of fuzzy inference and adaptive network learning (a superset of all FFNNs) to make and improve upon its estimations (Jang, 1993). In SWT modeling, ANFIS has been included in comparisons with other ANNs for model performance evaluation (Piotrowski et al., 2015; Zhu et al., 2019; Zhu, Hadzima-Nyarko, Gao, Wang, et al., 2019; Graf and Aghelpour, 2021). A different type of fuzzy ANN is the dynamic neuro-fuzzy local modeling system (DNFLMS), which contrasts with ANFIS by its use of the one-pass clustering algorithm and sequential learning algorithm (Hong and Bhamidimarri, 2012). A comparison of ANFIS and DNFLMS showed that the latter requires less training in terms of fuzzy rules needed and fewer epochs, which can result in over 18.5 h saved in computing time (Hong and Bhamidimarri, 2012).

Code availability. Computer code was not used to conduct this review.

Data availability. All data were obtained from the cited publications. Data used for Figs. 1 to 4 and Tables 1 and 2 are available online at CUAHSI HydroShare: http://www.hydroshare.org/resource/68edf1673096480bacc6bd104215e9dc (Corona, and Hogue, 2025).

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/hess-29-2521-2025-supplement.

Author contributions. CRC's contributions included conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, and writing (original draft preparation; review and editing). TSH's contributions included conceptualization, funding acquisition, methodology, project administration, resources, supervision, and writing (review and editing).

Competing interests. The contact author has declared that neither of the authors has any competing interests.

Disclaimer. This manuscript and related items of information have not been formally disseminated by NOAA and do not represent any agency determination, view, or policy.

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors are thankful for the support from the NOAA Cooperative Institute for Research to Operations in Hydrology, CIROH, United States (NA22NWS4320003).

Financial support. Funding was awarded to CIROH through the NOAA Cooperative Agreement with The University of Alabama (NA22NWS4320003).

Review statement. This paper was edited by Christa Kelleher and reviewed by Jeremy Diaz and two anonymous referees.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: large-scale machine learning on heterogeneous distributed systems, https://tensorflow.org, last access: 1 November 2015.

Abdi, R., Rust, A., and Hogue, T. S.: Development of a multilayer deep neural network model for predicting hourly river water temperature from meteorological data, Front. Environ. Sci., 9, 738322, https://doi.org/10.3389/fenvs.2021.738322, 2021.

Acito, F.: Predictive analytics with KNIME: Analytics for citizen data scientists, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-45630-5, 2023.

Ahmadi-Nedushan, B., St-Hilaire, A., Ouarda, T. B. M. J., Bilodeau, L., Robichaud, É., Thiémonge, N., and Bobée, B.: Predicting river water temperatures using stochastic models: case study of the Moisie River (Québec, Canada), Hydrol. Process., 21, 21–34, https://doi.org/10.1002/hyp.6353, 2007.

Akaike, H., Petrov, B. N., and Csaki, F.: Information theory and an extension of the maximum likelihood principle, Second international symposium on information theory, Akademiai Kiado, Budapest, Hungary, 2–8 September 1971, 267–281, https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15, 1973.

Anderson, D. and Burnham, K.: Model selection and multimodel inference, Second, Springer-Verlag, NY, 63, 10, https://doi.org/10.1007/b97636, 2004.

Anmala, J. and Turuganti, V.: Comparison of the performance of decision tree (DT) algorithms and extreme learning machine (ELM) model in the prediction of water quality of the Upper Green River watershed, Water Environ. Res., 93, 2360–2373, https://doi.org/10.1002/wer.1642, 2021.

Apaydin, H., Taghi Sattari, M., Falsafian, K., and Prasad, R.: Artificial intelligence modelling integrated with Singular Spectral analysis and Seasonal-Trend decomposition using Loess approaches for streamflow predictions, J. Hydrol., 600, 126506, https://doi.org/10.1016/j.jhydrol.2021.126506, 2021.

Arismendi, I., Safeeq, M., Dunham, J. B., and Johnson, S. L.: Can air temperature be used to project influences of climate change on stream temperature?, Environ. Res. Lett., 9, 084015, https://doi.org/10.1088/1748-9326/9/8/084015, 2014.

Banks, H. T. and Joyner, M. L.: AIC under the framework of least squares estimation, Appl. Math. Lett., 74, 33–45, https://doi.org/10.1016/j.aml.2017.05.005, 2017.

Bansal, K. and Tripathi, A. K.: Dual level attention based lightweight vision transformer for streambed land use change classification using remote sensing, Comput. Geosci., 191, 105676, https://doi.org/10.1016/j.cageo.2024.105676, 2024.

Barbarossa, V., Bosmans, J., Wanders, N., King, H., Bierkens, M. F. P., Huijbregts, M. A. J., and Schipper, A. M.: Threats of global warming to the world's freshwater fishes, Nat. Commun., 12, 1701, https://doi.org/10.1038/s41467-021-21655-w, 2021.

Bartholow, J. M.: Stream temperature investigations: field and analytical methods, US Fish and Wildlife Service, Biological Report 89:17, Washington, D.C., 139 pp. 1989.

Baydaroğlu, Ö. and Demir, I.: Temporal and spatial satellite data augmentation for deep learning-based rainfall nowcasting, J. Hydroinform., 26, 589–607, https://doi.org/10.2166/hydro.2024.235, 2024.

Bengio, Y., Simard, P., and Frasconi, P.: Learning long-term dependencies with gradient descent is difficult, IEEE T. Neutral. Networ., 5, 157–166, https://doi.org/10.1109/72.279181, 1994.

Bengio, Y., Courville, A., and Vincent, P.: Representation learning: A review and new perspectives, IEEE T. Pattern Anal., 35, 1798–1828, https://doi.org/10.48550/arXiv.1206.5538, 2013.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. M. J., and Bobée, B.: A review of statistical water temperature models, Can. Water Resour. J., 32, 179–192, https://doi.org/10.4296/cwrj3203179, 2007.

Beven, K.: Deep learning, hydrological processes and the uniqueness of place, Hydrol. Process., 34, 3608–3613, https://doi.org/10.1002/hyp.13805, 2020.

Bisong, E.: Google Colaboratory, in: Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners, edited by: Bisong, E., Apress, Berkeley, CA, 59–64, https://doi.org/10.1007/978-1-4842-4470-8_7, 2019.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Brown, G. W.: Predicting temperatures of small streams, Water Resour. Res., 5, 68–75, https://doi.org/10.1029/WR005i001p00068, 1969.

Buhmann, M. D.: Radial basis functions, Acta Numer., 9, 1–38, https://doi.org/10.1017/S0962492900000015, 2000.

Cairns Jr., J., Heath, A. G., and Parker, B. C.: Temperature influence on chemical toxicity to aquatic organisms, Water Pollution Control Federation, 47, 267–280, 1975.

Caissie, D., El-Jabi, N., and St-Hilaire, A.: Stochastic modelling of water temperatures in a small stream using air to water relations, Can. J. Civil Eng., 25, 250–260, https://doi.org/10.1139/l97-091, 1998.

Chang, H. and Psaris, M.: Local landscape predictors of maximum stream temperature and thermal sensitivity in the Columbia River Basin, USA, Sci. Total Environ., 461–462, 587–600, https://doi.org/10.1016/j.scitotenv.2013.05.033, 2013.

Chen, Y. D., Carsel, R. F., McCutcheon, S. C., and Nutter, W. L.: Stream temperature simulation of forested riparian areas: I. Watershed-scale model development, J. Environ. Eng., 124, 304–315, https://doi.org/10.1061/(ASCE)0733-9372(1998)124:4(304), 1998a.

Chen, Y. D., McCutcheon, S. C., Norton, D. J., and Nutter, W. L.: Stream temperature simulation of forested riparian areas: II. Model application, J. Environ. Eng., 124, 316–328, https://doi.org/10.1061/(ASCE)0733-9372(1998)124:4(316), 1998b.

Chenard, J. and Caissie, D.: Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada, Hydrol. Process., 22, 3361–3372, https://doi.org/10.1002/hyp.6928, 2008.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP 2014, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 25–29 October 2014, Doha, Qatar, [cs, stat], 11, https://doi.org/10.48550/arXiv.1406.1078, 2014.

Cluis, D. A.: Relationship between stream water temperature and ambient air temperature, Hydrol. Res., 3, 65–71, https://doi.org/10.2166/nh.1972.0004, 1972.

Cole, J. C., Maloney, K. O., Schmid, M., and McKenna, J. E.: Developing and testing temperature models for regulated systems: A case study on the Upper Delaware River, J. Hydrol., 519, 588–598, https://doi.org/10.1016/j.jhydrol.2014.07.058, 2014.

Corona, C. R. and Hogue, T. S.: Stream Temperature Machine Learning Review Performance Metrics Data, https://www.hydroshare.org/resource/68edf1673096480bacc6bd104215e9dc/, last access: 27 January 2025.

Cortes, C. and Vapnik, V.: Support-vector networks, Mach. Learn., 20, 273–297, https://doi.org/10.1007/BF00994018, 1995.

Cover, T. and Hart, P.: Nearest neighbor pattern classification, IEEE Trans. Inform. Theory, 13, 21–27, https://doi.org/10.1109/TIT.1967.1053964, 1967.

CUAHSI Inc.: Consortium of Universities for the Advancement of Hydrologic Sciences Inc. (CUAHSI), HydroShare, https://www.hydroshare.org/, last access: 30 December 2024.

Dawdy, D. R. and Thompson, T. H.: Digital computer simulation in hydrology, Journal AWWA, 59, 685–688, https://doi.org/10.1002/j.1551-8833.1967.tb03398.x, 1967.

Detenbeck, N. E., Morrison, A. C., Abele, R. W., and Kopp, D. A.: Spatial statistical network models for stream and river temperature in New England, USA, Water Resour. Res., 52, 6018–6040, https://doi.org/10.1002/2015WR018349, 2016.

DeWeber, J. T. and Wagner, T.: A regional neural network ensemble for predicting mean daily river water temperature, J. Hydrol., 517, 187–200, https://doi.org/10.1016/j.jhydrol.2014.05.035, 2014.

Drainas, K., Kaule, L., Mohr, S., Uniyal, B., Wild, R., and Geist, J.: Predicting stream water temperature with artificial neural networks based on open-access data, Hydrol. Process., 37, e14991, https://doi.org/10.1002/hyp.14991, 2023.

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

Edinger, J. E., Duttweiler, D. W., and Geyer, J. C.: The response of water temperatures to meteorological conditions, Water Resour. Res., 4, 1137–1143, 1968.

Elman, J. L.: Finding structure in time, Cognitive Science, 14, 179–211, https://doi.org/10.1207/s15516709cog1402_1, 1990.

Fahlman, S. and Lebiere, C.: The cascade-correlation learning architecture, Adv. Neur. In., 2, 524–532, https://dl.acm.org/doi/10.5555/109230.107380, 1990.

Faraway, J. and Chatfield, C.: Time series forecasting with neural networks: A comparative study using the air line data, J. R. Stat. Sco. C-Appl., 47, 231–250, https://doi.org/10.1111/1467-9876.00109, 1998.

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

Fix, E. and Hodges, J. L.: Discriminatory analysis: Nonparametric discrimination: Small sample performance, https://www.jstor.org/stable/1403797 (last access: 27 January 2025), 1952.

Foreman, M. G. G., Lee, D. K., Morrison, J., Macdonald, S., Barnes, D., and Williams, I. V.: Simulations and retrospective analyses of fraser watershed flows and temperatures, Atmos.-Ocean, 39, 89–105, https://doi.org/10.1080/07055900.2001.9649668, 2001.

Friedberg, R. M.: A learning machine: Part I, IBM J. Res. & Dev., 2, 2–13, https://doi.org/10.1147/rd.21.0002, 1958.

Fuller, M. R., Detenbeck, N. E., Leinenbach, P., Labiosa, R., and Isaak, D.: Spatial and temporal variability in stream ther-

mal regime drivers for three river networks during the summer growing season, J. American Water Resour. Assoc., 60, 57–78, https://doi.org/10.1111/1752-1688.13158, 2023.

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

Gers, F. A. and Schmidhuber, J.: Recurrent nets that time and count, in: IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, Neural Computing: New Challenges and Perspectives for the New Millennium, 7 July 2000, Como, Italy, 189–194, https://doi.org/10.1109/IJCNN.2000.861302, 2000.

Ghobadi, F. and Kang, D.: Improving long-term streamflow prediction in a poorly gauged basin using geo-spatiotemporal mesoscale data and attention-based deep learning: A comparative study, J. Hydrol., 615, 128608, https://doi.org/10.1016/j.jhydrol.2022.128608, 2022.

GitHub Inc.: GitHub, https://www.github.com/github/, last access: 31 December 2024.

Google, Inc.: Gemini, a large language model, Alphabet, Inc., https://gemini.google.com, last access: 27 January 2025.

Graf, R. and Aghelpour, P.: Daily river water temperature prediction: A comparison between neural network and stochastic techniques, Atmosphere, 12, 1154, https://doi.org/10.3390/atmos12091154, 2021.

Graf, R., Zhu, S., and Sivakumar, B.: Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach, J. Hydrol., 578, 124115, https://doi.org/10.1016/j.jhydrol.2019.124115, 2019.

Grbić, R., Kurtagić, D., and Slišković, D.: Stream water temperature prediction based on Gaussian process regression, Expert Syst. Appl., 40, 7407–7414, https://doi.org/10.1016/j.eswa.2013.06.077, 2013.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J.: LSTM: A search space odyssey, IEEE T. Neur. Net. Lear., 28, 2222–2232, https://doi.org/10.48550/arXiv.1503.04069, 2016.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, J. Hydrol. Eng., 4, 135–143, https://doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135), 1999.

Hadzima-Nyarko, M., Rabi, A., and Šperac, M.: Implementation of artificial neural networks in modeling the water-air temperature relationship of the River Drava, Water Resour. Manage., 28, 1379–1394, https://doi.org/10.1007/s11269-014-0557-7, 2014.

Hani, I., St-Hilaire, A., and Ouarda, T. B. M. J.: Machine-learning modeling of hourly potential thermal refuge area: A case study from the Sainte-Marguerite River (Quebec, Canada), River Res. Appl., rra.4191, https://doi.org/10.1002/rra.4191, 2023.

Hastie, T. and Tibshirani, R.: Generalized additive models: Some applications, J. Am. Stat. Assoc., 82, 371–386, https://doi.org/10.1080/01621459.1987.10478440, 1987.

Hastie, T., Friedman, J., and Tibshirani, R.: The elements of statistical learning, Springer New York, New York, NY, https://doi.org/10.1007/978-0-387-21606-5, 2001.

He, E., Xie, Y., Sun, A., Zwart, J., Yang, J., Jin, Z., Wang, Y., Karimi, H., and Jia, X.: Fair graph learning using constraint-aware priority adjustment and graph masking in river networks, AAAI, 38, 22087–22095, https://doi.org/10.1609/aaai.v38i20.30212, 2024.

Hebert, C., Caissie, D., Satish, M. G., and El-Jabi, N.: Modeling of hourly river water temperatures using artificial neural networks, Water Qual. Res. J., 49, 144–162, https://doi.org/10.2166/wqrjc.2014.007, 2014.

Heddam, S., Kim, S., Danandeh Mehr, A., Zounemat-Kermani, M., Ptak, M., Elbeltagi, A., Malik, A., and Tikhamarine, Y.: Bat algorithm optimised extreme learning machine (Bat-ELM): A novel approach for daily river water temperature modelling, Geogr. J., 189, 78–89, https://doi.org/10.1111/geoj.12478, 2022a.

Heddam, S., Ptak, M., Sojka, M., Kim, S., Malik, A., Kisi, O., and Zounemat-Kermani, M.: Least square support vector machine-based variational mode decomposition: a new hybrid model for daily river water temperature modeling, Environ. Sci. Pollut. Res., 29, 71555–71582, https://doi.org/10.1007/s11356-022-20953-0, 2022b.

Helsel, D. R. and Hirsch, R. M.: Chapter A3: Statistical Methods in Water Resources, in: Techniques of Water Resources Investigations, Book 4, U.S. Geological Survey, 522, https://doi.org/10.3133/twri04A3, 2002.

Hinton, G. E.: How neural networks learn from experience, Sci. Am., 267, 144–151, 1992.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hong, Y.-S. T. and Bhamidimarri, R.: Dynamic neuro-fuzzy local modeling system with a nonlinear feature extraction for the on-line adaptive warning system of river temperature affected by waste cooling water discharge, Stoch. Environ. Res. Risk Assess., 26, 947–960, https://doi.org/10.1007/s00477-011-0543-z, 2012.

Hsu, K., Gupta, H. V., and Sorooshian, S.: Artificial neural network modeling of the rainfall-runoff process, Water Resour. Res., 31, 2517–2530, https://doi.org/10.1029/95WR01955, 1995.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K.: Extreme learning machine: Theory and applications, Neurocomputing, 70, 489–501, https://doi.org/10.1016/j.neucom.2005.12.126, 2006.

Irani, J., Pise, N., and Phatak, M.: Clustering techniques and the similarity measures used in clustering: A survey, IJCA, 134, 9–14, https://doi.org/10.5120/ijca2016907841, 2016.

Isaak, D. J., Wenger, S. J., Peterson, E. E., Ver Hoef, J. M., Nagel, D. E., Luce, C. H., Hostetler, S. W., Dunham, J. B., Roper, B. B., Wollrab, S. P., Chandler, G. L., Horan, D. L., and Parkes-Payne, S.: The NorWeST summer stream temperature model and scenarios for the Western U.S.: A crowd-sourced database and new geospatial tools foster a user community and predict broad climate warming of rivers and streams, Water Resour. Res., 53, 9181–9205, https://doi.org/10.1002/2017WR020969, 2017.

Islam, M. N. and Sivakumar, B.: Characterization and prediction of runoff dynamics: a nonlinear dynamical view, Adv. Water Resour., 25, 179–190, https://doi.org/10.1016/S0309-1708(01)00053-7, 2002.

Ivakhnenko, A. G.: Heuristic self-organization in problems of engineering cybernetics, Automatica, 6, 207–219, https://doi.org/10.1016/0005-1098(70)90092-0, 1970.

Ivakhnenko, A. G. and Ivakhnenko, G. A.: The review of problems solvable by algorithms of the group method of data handling (GMDH), Pattern recognition and image analysis, 5, 527–535, 1995.

Jaber, F. and Shukla, S.: MIKE SHE: Model use, calibration, and validation, T. ASABE, 55, 1479–1489, https://doi.org/10.13031/2013.42255, 2012.

Jang, J.-S. R.: ANFIS: adaptive-network-based fuzzy inference system, IEEE Trans. Syst., Man, Cybern., 23, 665–685, https://doi.org/10.1109/21.256541, 1993.

Janson, D. J. and Frenzel, J. F.: Training product unit neural networks with genetic algorithms, IEEE Expert, 8, 26–33, https://doi.org/10.1109/64.236478, 1993.

Jeong, K., Lee, J., Lee, K. Y., and Kim, B.: Artificial neural network-based real time water temperature prediction in the Soyang River, The Transactions of The Korean Institute of Electrical Engineers, 65, 2084–2093, https://doi.org/10.5370/KIEE.2016.65.12.2084, 2016.

Jiang, D., Xu, Y., Lu, Y., Gao, J., and Wang, K.: Forecasting water temperature in cascade reservoir operation-influenced river with machine learning models, Water, 14, 2146, https://doi.org/10.3390/w14142146, 2022.

Johnson, S. L. and Jones, J. A.: Stream temperature responses to forest harvest and debris flows in western Cascades, Oregon, Can. J. Fish. Aquat. Sci., 57, 10, https://doi.org/10.1139/f00-109, 2000.

Kalogirou, S. A.: Solar energy engineering: processes and systems, 3rd edn., Elsevier, 902 pp., https://doi.org/10.1016/B978-0-12-374501-9.X0001-5, 2023.

Karunanithi, N., Grenney, W. J., Whitley, D., and Bovee, K.: Neural networks for river flow prediction, J. Comput. Civil Eng., 8, 201–220, https://doi.org/10.1061/(ASCE)0887-3801(1994)8:2(201), 1994.

Khosravi, M., Duti, B. M., Yazdan, M. M. S., Ghoochani, S., Nazemi, N., and Shabanian, H.: Multivariate multi-step long short-term memory neural network for simultaneous stream-water variable prediction, Eng, 4, 1933–1950, https://doi.org/10.3390/eng4030109, 2023.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

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

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

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

Krishnaraj, A. and Deka, P. C.: Spatial and temporal variations in river water quality of the Middle Ganga Basin using unsupervised machine learning techniques, Environ. Monit. Assess., 192, 744, https://doi.org/10.1007/s10661-020-08624-4, 2020.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Adv. Neur. In., 25, 1–9, https://doi.org/10.1145/3065386, 2012.

Kwak, J., St-Hilaire, A., and Chebana, F.: A comparative study for water temperature modelling in a small basin, the Fourchue River, Quebec, Canada, Hydrol. Sci. J., 62, 64–75, https://doi.org/10.1080/02626667.2016.1174334, 2016.

Laanaya, F., St-Hilaire, A., and Gloaguen, E.: Water temperature modelling: comparison between the generalized additive model, logistic, residuals regression and linear regression models, Hydrolog. Sci. J., 62, 1078–1093, https://doi.org/10.1080/02626667.2016.1246799, 2017.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R.: Unmasking Clever Hans predictors and assessing what machines really learn, Nat. Commun., 10, 1096, https://doi.org/10.1038/s41467-019-08987-4, 2019.

Lea, C., Vidal, R., Reiter, A., and Hager, G. D.: Temporal convolutional networks: A unified approach to action segmentation, Proceedings of the 14th European Conference Computer Vision – ECCV 2016 Workshops: Part III, 11–16 October 2016, Amsterdam, Netherlands, 47–54, https://doi.org/10.1007/978-3-319-49409-8_7, 2016.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L.: Handwritten digit recognition with a back-propagation network, in: NIPS '89, Proceedings of the third International Conference on Neural Information Processing Systems, Denver, Colorado, USA, 1 January 1989, 396–404, https://dl.acm.org/doi/10.5555/2969830.2969879, 1989.

LeCun, Y., Huang, F. J., and Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting, in: CVPR 2004, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, 27 June–2 July 2004, Washington, D.C., U.S., II-97–104, https://doi.org/10.1109/CVPR.2004.1315150, 2004.

Lee, S.-Y., Fullerton, A. H., Sun, N., and Torgersen, C. E.: Projecting spatiotemporally explicit effects of climate change on stream temperature: A model comparison and implications for coldwater fishes, J. Hydrol., 588, 125066, https://doi.org/10.1016/j.jhydrol.2020.125066, 2020.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

Liu, D., Xu, Y., Guo, S., Xiong, L., Liu, P., and Zhao, Q.: Stream temperature response to climate change and water diversion activities, Stoch. Environ. Res. Risk Assess., 32, 1397–1413, https://doi.org/10.1007/s00477-017-1487-8, 2018.

Loh, W.-Y.: Classification and Regression Tree Methods. In Encyclopedia of Statistics in Quality and Reliability, edited by: Ruggeri, F., Kenett, R. S., and Faltin, F. W., 315–323, https://doi.org/10.1002/9780470061572.eqr492, 2008.

Loinaz, M. C., Davidsen, H. K., Butts, M., and Bauer-Gottwein, P.: Integrated flow and temperature modeling at the catchment scale, J. Hydrol., 495, 238–251, https://doi.org/10.1016/j.jhydrol.2013.04.039, 2013.

Lu, H. and Ma, X.: Hybrid decision tree-based machine learning models for short-term water quality prediction, Chemosphere, 249, 126169, https://doi.org/10.1016/j.chemosphere.2020.126169, 2020.

Maheu, A., Poff, N. L., and St-Hilaire, A.: A classification of stream water temperature regimes in the conterminous USA, River Res. Appl., 32, 896–906, https://doi.org/10.1002/rra.2906, 2016.

Majerska, M., Osuch, M., and Wawrzyniak, T.: Long-term patterns and changes of unglaciated High Arctic stream thermal regime, Sci. Total Environ., 923, 171298, https://doi.org/10.1016/j.scitotenv.2024.171298, 2024.

Martínez-Estudillo, A., Martínez-Estudillo, F., Hervás-Martínez, C., and García-Pedrajas, N.: Evolutionary product unit based neural networks for regression, Neural Networks, 19, 477–486, https://doi.org/10.1016/j.neunet.2005.11.001, 2006.

McCulloch, W. S. and Pitts, W.: A logical calculus of the ideas immanent in nervous activity, The bulletin of mathematical biophysics, 5, 115–133, https://doi.org/10.1007/BF02478259, 1943.

Microsoft, Inc.: Microsoft Copilot, a large language model, https://copilot.microsoft.com, last access: 27 January 2025.

Møller, M. F.: A scaled conjugate gradient algorithm for fast supervised learning, Neural networks, 6, 525–533, https://doi.org/10.1016/S0893-6080(05)80056-5, 1993.

Moore, A. W., Schneider, J., and Deng, K.: Efficient locally weighted polynomial regression predictions, in: ICML '97, Proceedings of the Fourteenth International Machine Learning Conference, 8–12 July 1997, San Francisco, California, USA, 9, 236–244, ISBN 978-1-55860-486-5, 1997.

Moriasi, D. N., Arnold, J. G., Liew, M. W. V., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, T. ASABE, 50, 885–900, https://doi.org/10.13031/2013.23153, 2007.

Moriasi, D. N., Gitau, M. W., Pai, N., and Daggupati, P.: Hydrologic and water quality models: Performance measures and evaluation criteria, T. ASABE, 58, 1763–1785, https://doi.org/10.13031/trans.58.10715, 2015.

Morse, W. L.: Stream temperature prediction model, Water Resour. Res., 6, 290–302, https://doi.org/10.1029/WR006i001p00290, 1970.

Morshed, J. and Kaluarachchi, J. J.: Application of artificial neural network and genetic algorithm in flow and transport simulations, Adv. Water Resour., 22, 145–158, https://doi.org/10.1016/S0309-1708(98)00002-5, 1998.

Musavi, M. T., Ahmed, W., Chan, K. H., Faris, K. B., and Hummels, D. M.: On the training of radial basis function classifiers, Neural networks, 5, 595–603, https://doi.org/10.1016/S0893-6080(05)80038-3, 1992.

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

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resour. Res., 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

OpenAI, Inc.: ChatGPT (27 Jan version), a large language model, https://chat.openai.com/chat, last access: 27 January 2025.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An imperative style, high-performance deep learning library, arXiv [preprint], https://doi.org/10.48550/arXiv.1912.01703, 3 December 2019.

Patra, R. W., Chapman, J. C., Lim, R. P., Gehrke, P. C., and Sunderam, R. M.: Interactions between water temperature and contaminant toxicity to freshwater fish, Environ. Toxicol. Chem., 34, 1809–1817, https://doi.org/10.1002/etc.2990, 2015.

Philippus, D., Sytsma, A., Rust, A., and Hogue, T. S.: A machine learning model for estimating the temperature of small rivers using satellite-based spatial data, Remote Sens. Environ., 311, 114271, https://doi.org/10.1016/j.rse.2024.114271, 2024a.

Philippus, D., Corona, C. R., and Hogue, T. S.: Improved annual temperature cycle function for stream seasonal thermal regimes, J. American Water Resour. Assoc., 60, 1080–1094, https://doi.org/10.1111/1752-1688.13228, 2024b.

Piccolroaz, S., Calamita, E., Majone, B., Gallice, A., Siviglia, A., and Toffolon, M.: Prediction of river water temperature: a comparison between a new family of hybrid models and statistical approaches, Hydrol. Process., 30, 3901–3917, https://doi.org/10.1002/hyp.10913, 2016.

Piotrowski, A. P., Napiorkowski, M. J., Napiorkowski, J. J., and Osuch, M.: Comparing various artificial neural network types for water temperature prediction in rivers, J. Hydrol., 529, 302–315, https://doi.org/10.1016/j.jhydrol.2015.07.044, 2015.

Piotrowski, A. P., Napiorkowski, J. J., and Piotrowska, A. E.: Impact of deep learning-based dropout on shallow neural networks applied to stream temperature modelling, Earth-Sci. Rev., 201, 103076, https://doi.org/10.1016/j.earscirev.2019.103076, 2020.

Piotrowski, A. P., Marzena, O., and Napiorkowski, J. J.: Influence of the choice of stream temperature model on the projections of water temperature in rivers, J. Hydrol., 601, 1–21, https://doi.org/10.1016/j.jhydrol.2021.126629, 2021.

Poff, N. L., Tokar, S., and Johnson, P.: Stream hydrological and ecological responses to climate change assessed with an artificial neural network, Limnol. Oceanogr., 41, 857–863, https://doi.org/10.4319/lo.1996.41.5.0857, 1996.

Poole, G. C. and Berman, C. H.: An ecological perspective on in-stream temperature: Natural heat dynamics and mechanisms of human-caused thermal degradation, Environ. Manage, 27, 787–802, https://doi.org/10.1007/s002670010188, 2001.

Portet, S.: A primer on model selection using the Akaike Information Criterion, Infectious Disease Modelling, 5, 111–128, https://doi.org/10.1016/j.idm.2019.12.010, 2020.

Qiu, R., Wang, Y., Wang, D., Qiu, W., Wu, J., and Tao, Y.: Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River, Sci. Total Environ., 737, 139729, https://doi.org/10.1016/j.scitotenv.2020.139729, 2020.

Rabi, A., Hadzima-Nyarko, M., and Šperac, M.: Modelling river temperature from air temperature: case of the River Drava (Croatia), Hydrolog. Sci. J., 60, 1490–1507, https://doi.org/10.1080/02626667.2014.914215, 2015.

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C.: Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data, Environ. Res. Lett., 16, 1–11, https://doi.org/10.1088/1748-9326/abd501, 2020.

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

Rahmani, F., Appling, A., Feng, D., Lawson, K., and Shen, C.: Identifying structural priors in a hybrid differentiable model for stream water temperature modeling, Water Resour. Res., 59, e2023WR034420, https://doi.org/10.1029/2023WR034420, 2023.

Rajesh, M. and Rehana, S.: Prediction of river water temperature using machine learning algorithms: a tropical river system of India, J. Hydroinform., 23, 605–626, 2021.

Rehana, S.: River water temperature modelling under climate change using support vector regression, in: Hydrology in a Changing World, edited by: Singh, S. K. and Dhanya, C. T., Springer International Publishing, Cham, 171–183, https://doi.org/10.1007/978-3-030-02197-9_8, 2019.

Rehana, S. and Rajesh, M.: Assessment of impacts of climate change on indian riverine thermal regimes using hybrid deep learning methods, Water Resour. Res., 59, e2021WR031347, https://doi.org/10.1029/2021WR031347, 2023.

Risley, J. C., Roehl, E. A., and Conrads, P. A.: Estimating water temperatures in small streams in western Oregon using neural network models, U.S. Geological Survey, https://doi.org/10.3133/wri024218, 2003.

Risley, J. C., Constantz, J., Essaid, H., and Rounds, S.: Effects of upstream dams versus groundwater pumping on stream temperature under varying climate conditions: Upstream Dam and Groundwater Pumping Impacts, Water Resour. Res., 46, 1–32, https://doi.org/10.1029/2009WR008587, 2010.

Rogers, J. B., Stein, E. D., Beck, M. W., and Ambrose, R. F.: The impact of climate change induced alterations of streamflow and stream temperature on the distribution of riparian species, PLoS ONE, 15, e0242682, https://doi.org/10.1371/journal.pone.0242682, 2020.

Rozos, E.: Assessing hydrological simulations with machine learning and statistical models, Hydrology, 10, 49, https://doi.org/10.3390/hydrology10020049, 2023.

Sadler, J. M., Appling, A. P., Read, J. S., Oliver, S. K., Jia, X., Zwart, J. A., and Kumar, V.: Multi-task deep learning of daily streamflow and water temperature, Water Resour. Res., 58, e2021WR030138, https://doi.org/10.1029/2021WR030138, 2022.

Sahoo, G. B., Schladow, S. G., and Reuter, J. E.: Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models, J. Hydrol., 378, 325–342, https://doi.org/10.1016/j.jhydrol.2009.09.037, 2009.

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

Shamseldin, A. Y.: Application of a neural network technique to rainfall-runoff modelling, J. Hydrol., 199, 272–294, https://doi.org/10.1016/S0022-1694(96)03330-6, 1997.

Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources scientists, Water Resour. Res., 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: NIPS '15, Proceedings of the 29th International Conference on Neural Information Processing Systems, 7–12 December 2015, Montreal, Canada, 802–810, https://dl.acm.org/doi/10.5555/2969239.2969329 (last access: 12 June 2025), 2015.

Siegel, J. E., Fullerton, A. H., FitzGerald, A. M., Holzer, D., and Jordan, C. E.: Daily stream temperature predictions for free-flowing streams in the Pacific Northwest, USA, PloS Water, 2, 1–27, https://doi.org/10.1371/journal.pwat.0000119, 2023.

Sinokrot, B. A. and Stefan, H. G.: Stream temperature dynamics: Measurements and modeling, Water Resour. Res., 29, 2299–2312, https://doi.org/10.1029/93WR00540, 1993.

Sivri, N., Kilic, N., and Ucan, O. N.: Estimation of stream temperature in Firtina Creek (Rize-Turkiye) using artificial neural network model, J. Environ. Biol., 28, 67–72, 2007.

Skoulikaris, C., Venetsanou, P., Lazoglou, G., Anagnostopoulou, C., and Voudouris, K.: Spatio-temporal interpolation and bias correction ordering analysis for hydrological simulations: An assessment on a mountainous river basin, Water, 14, 660, https://doi.org/10.3390/w14040660, 2022.

Smith, K. and Lavis, M. E.: Environmental influences on the temperature of a small upland stream, Oikos, 26, 228, https://doi.org/10.2307/3543713, 1975.

Solomatine, D. P., Maskey, M., and Shrestha, D. L.: Instance-based learning compared to other data-driven methods in hydrological forecasting, Hydrol. Process., 22, 275–287, https://doi.org/10.1002/hyp.6592, 2008.

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

Specht, D. F.: A general regression neural network, IEEE T. Neutral. Networ., 2, 568–576, https://doi.org/10.1109/72.97934, 1991.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: A simple way to prevent neural net- works from overfitting, J. Mach. Learn. Res., 15, 1929–1958, https://dl.acm.org/doi/10.5555/2627435.2670313 (last access: 30 December 2024), 2014.

St-Hilaire, A., Morin, G., El-Jabi, N., and Caissie, D.: Water temperature modelling in a small forested stream: implication of forest canopy and soil temperature, Can. J. Civil Eng., 27, 1095–1108, https://doi.org/10.1139/l00-021, 2000.

St-Hilaire, A., Ouarda, T. B. M. J., Bargaoui, Z., Daigle, A., and Bilodeau, L.: Daily river water temperature forecast model with a k-nearest neighbour approach, Hydrol. Process., 26, 1302–1310, https://doi.org/10.1002/hyp.8216, 2011.

Suykens, J. A. and Vandewalle, J.: Least squares support vector machine classifiers, Neural. Process. Lett., 9, 293–300, https://doi.org/10.1023/A:1018628609742, 1999.

Tao, Y., Wang, Y., Rhoads, B., Wang, D., Ni, L., and Wu, J.: Quantifying the impacts of the Three Gorges Reservoir on water temperature in the middle reach of the Yangtze River, J. Hydrol., 582, 124476, https://doi.org/10.1016/j.jhydrol.2019.124476, 2020.

Temizyurek, M. and Dadaser-Celik, F.: Modelling the effects of meteorological parameters on water temperature using artificial neural networks, Water Sci. Technol., 77, 1724–1733, https://doi.org/10.2166/wst.2018.058, 2018.

Theurer, F. D., Voos, K. A., and Miller, W. J.: Instream water temperature model, Western Energy and Land Use Team, Division of Biological Services, Research and Development, USDI, Fish W., Instream Flow Information Paper, 16, 352 pp., 1984.

Thirumalaiah, K. and Deo, M. C.: Real-Time flood forecasting using neural networks, Comput.-Aided Civ. Inf., 13, 101–111, https://doi.org/10.1111/0885-9507.00090, 1998.

Tibshirani, R.: Regression shrinkage and selection via the lasso, J. R. Stat. Soc., 58, 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x, 1996.

Tipping, M. E.: Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res., 1, 211–244, https://doi.org/10.1162/15324430152748236, 2001.

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

Topp, S. N., Barclay, J., Diaz, J., Sun, A. Y., Jia, X., Lu, D., Sadler, J. M., and Appling, A. P.: Stream temperature prediction in a shifting environment: Explaining the influence of deep learning architecture, Water Resour. Res., 59, e2022WR033880, https://doi.org/10.1029/2022WR033880, 2023.

Ulaski, M. E., Warkentin, L., Naman, S. M., and Moore, J. W.: Spatially variable effects of streamflow on water temperature and thermal sensitivity within a salmon-bearing watershed in interior British Columbia, Canada, River Res. Appl., 39, 2036–2047, https://doi.org/10.1002/rra.4200, 2023.

Varadharajan, C., Appling, A. P., Arora, B., Christianson, D. S., Hendrix, V. C., Kumar, V., Lima, A. R., Müller, J., Oliver, S., Ombadi, M., Perciano, T., Sadler, J. M., Weierbach, H., Willard, J. D., Xu, Z., and Zwart, J.: Can machine learning accelerate process understanding and decision-relevant predictions of river water quality?, Hydrol. Process., 36, e14565, https://doi.org/10.1002/hyp.14565, 2022.

Venkateswarlu, T. and Anmala, J.: Importance of land use factors in the prediction of water quality of the Upper Green River watershed, Kentucky, USA, using random forest, Environ. Dev. Sustain., 26, 23961–23984, https://doi.org/10.1007/s10668-023-03630-1, 2023.

Voza, D. and Vuković, M.: The assessment and prediction of temporal variations in surface water quality – a case study, Environ. Monit. Assess., 190, 434, https://doi.org/10.1007/s10661-018-6814-0, 2018.

Wade, J., Kelleher, C., and Hannah, D. M.: Mach. Learn. unravels controls on river water temperature regime dynamics, J. Hydrol., 623, 129821, https://doi.org/10.1016/j.jhydrol.2023.129821, 2023.

Walker, J. and Lawson, J.: Natural stream temperature variations in a catchment, Water Res., 11, 373–377, https://doi.org/10.1016/0043-1354(77)90025-2, 1977.

Wallace, J. B. and Webster, J. R.: The role of macroinvertebrates in stream ecosystem function, Annu. Rev. Entomol., 41, 115–139, https://doi.org/10.1146/annurev.en.41.010196.000555, 1996.

Wanders, N., Thober, S., Kumar, R., Pan, M., Sheffield, J., Samaniego, L., and Wood, E. F.: Development and Evaluation of a Pan-European Multimodel Seasonal Hydrological Forecasting System, J. Hydrometeorol., 20, 99–115, https://doi.org/10.1175/JHM-D-18-0040.1, 2019.

Ward, J. C.: Annual variation of stream water temperature, J. Sanit. Eng. Div. ASCE, 89, 1–16, https://doi.org/10.1061/JSEDAI.0000463, 1963.

Ward, J. V.: Riverine landscapes: Biodiversity patterns, disturbance regimes, and aquatic conservation, Biol. Conserv., 83, 269–278, https://doi.org/10.1016/S0006-3207(97)00083-9, 1998.

Webb, R. W., Fassnacht, S. R., and Gooseff, M. N.: Wetting and Drying Variability of the Shallow Subsurface Beneath a Snowpack in California's Southern Sierra Nevada, Vadose Zone J., 14, vzj2014.12.0182, https://doi.org/10.2136/vzj2014.12.0182, 2015.

Wei, X.: Evaluation of transformer model and self-attention mechanism in the Yangtze River basin runoff prediction, J. Hydrol.: Regional Studies, 47, 13, https://doi.org/10.1016/j.ejrh.2023.101438, 2023.

Weierbach, H., Lima, A. R., Willard, J. D., Hendrix, V. C., Christianson, D. S., Lubich, M., and Varadharajan, C.: Stream temperature predictions for river basin management in the Pacific Northwest and mid-Atlantic regions using machine learning, Water, 14, 1032, https://doi.org/10.3390/w14071032, 2022.

Wild, R., Nagel, C., and Geist, J.: Climate change effects on hatching success and embryonic development of fish: Assessing multiple stressor responses in a large-scale mesocosm study, Sci. Total Environ., 893, 164834, https://doi.org/10.1016/j.scitotenv.2023.164834, 2023.

Wu, Z., Pan, S., Long, G., Jiang, J., and Zhang, C.: Graph WaveNet for deep spatial-temporal graph modeling, in: IJCAI '19, Proceedings of the 28th International Joint Conference on Artificial Intelligence, 10–16 August 2019, Macao, China, 1907–1913, https://dl.acm.org/doi/10.5555/3367243.3367303 (last access: 30 December 2024), 2019.

Xu, T. and Liang, F.: Mach. Learn. for hydrologic sciences: An introductory overview, WIREs Water, 8, e1533, https://doi.org/10.1002/wat2.1533, 2021.

Yang, M., Yang, Q., Shao, J., Wang, G., and Zhang, W.: A new few-shot learning model for runoff prediction: Demonstration in two data scarce regions, Environ. Modell. Softw., 162, 105659, https://doi.org/10.1016/j.envsoft.2023.105659, 2023.

Yao, K., Cohn, T., Vylomova, K., Duh, K., and Dyer, C.: Depth-Gated LSTM, 2015 Jelinek Summer Workshop on Speech and Language Technology, arXiv [preprint], 5, https://doi.org/10.48550/arXiv.1508.03790, 2015.

Zanoni, M. G., Majone, B., and Bellin, A.: A catchment-scale model of river water quality by Machine Learning, Sci. Total Environ., 838, 156377, https://doi.org/10.1016/j.scitotenv.2022.156377, 2022.

Zhu, S. and Heddam, S.: Modelling of maximum daily water temperature for streams: Optimally pruned extreme learn-

ing machine (OPELM) versus radial basis function neural networks (RBFNN), Environ. Process., 6, 789–804, https://doi.org/10.1007/s40710-019-00385-8, 2019.

Zhu, S. and Piotrowski, A. P.: River/stream water temperature forecasting using artificial intelligence models: a systematic review, Acta Geophys., 68, 1433–1442, https://doi.org/10.1007/s11600-020-00480-7, 2020.

Zhu, S., Nyarko, E. K., Hadzima-Nyarko, M., Heddam, S., and Wu, S.: Assessing the performance of a suite of machine learning models for daily river water temperature prediction, PeerJ, 7, e7065, https://doi.org/10.7717/peerj.7065, 2019a.

Zhu, S., Heddam, S., Wu, S., Dai, J., and Jia, B.: Extreme learning machine-based prediction of daily water temperature for rivers, Environ. Earth Sci., 78, 202, https://doi.org/10.1007/s12665-019-8202-7, 2019b.

Zhu, S., Heddam, S., Nyarko, E. K., Hadzima-Nyarko, M., Piccolroaz, S., and Wu, S.: Modeling daily water temperature for rivers: comparison between adaptive neuro-fuzzy inference systems and artificial neural networks models, Environ. Sci. Pollut. Res., 26, 402–420, https://doi.org/10.1007/s11356-018-3650-2, 2019c.

Zhu, S., Hadzima-Nyarko, M., Gao, A., Wang, F., Wu, J., and Wu, S.: Two hybrid data-driven models for modeling water-air temperature relationship in rivers, Environ. Sci. Pollut. Res., 26, 12622–12630, https://doi.org/10.1007/s11356-019-04716-y, 2019d.

Zwart, J. A., Diaz, J., Hamshaw, S., Oliver, S., Ross, J. C., Sleckman, M., Appling, A. P., Corson-Dosch, H., Jia, X., Read, J., Sadler, J., Thompson, T., Watkins, D., and White, E.: Evaluating deep learning architecture and data assimilation for improving water temperature forecasts at unmonitored locations, Front. Water, 5, 1184992, https://doi.org/10.3389/frwa.2023.1184992, 2023a.

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