water

MDPI

Article

Stream Temperature Predictions for River Basin Management in the Pacific Northwest and Mid-Atlantic Regions Using Machine Learning

Helen Weierbach [1], Aranildo R. Lima [2], Jared D. Willard [1,3], Valerie C. Hendrix [4], Danielle S. Christianson [4], Michaelle Lubich [1,5] and Charuleka Varadharajan [1,*]

1 Earth and Environmental Sciences Area, Lawrence Berkeley National Lab, Berkeley, CA 94720, USA; hweierbach@lbl.gov (H.W.); willa099@umn.edu (J.D.W.); mishalubich007@berkeley.edu (M.L.)
2 Aquatic Informatics, Vancouver, BC V6E 4M3, Canada; proj.brain@gmail.com or aranildo.lima@aquaticinformatics.com
3 Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA
4 Computing Sciences Area, Lawrence Berkeley National Lab, Berkeley, CA 94720, USA; vchendrix@lbl.gov (V.C.H.); dschristianson@lbl.gov (D.S.C.)
5 Berkeley Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA
* Correspondence: cvaradharajan@lbl.gov

Abstract: Stream temperature (T_s) is an important water quality parameter that affects ecosystem health and human water use for beneficial purposes. Accurate T_s predictions at different spatial and temporal scales can inform water management decisions that account for the effects of changing climate and extreme events. In particular, widespread predictions of T_s in unmonitored stream reaches can enable decision makers to be responsive to changes caused by unforeseen disturbances. In this study, we demonstrate the use of classical machine learning (ML) models, support vector regression and gradient boosted trees (XGBoost), for monthly T_s predictions in 78 pristine and human-impacted catchments of the Mid-Atlantic and Pacific Northwest hydrologic regions spanning different geologies, climate, and land use. The ML models were trained using long-term monitoring data from 1980–2020 for three scenarios: (1) temporal predictions at a single site, (2) temporal predictions for multiple sites within a region, and (3) spatiotemporal predictions in unmonitored basins (PUB). In the first two scenarios, the ML models predicted T_s with median root mean squared errors (RMSE) of 0.69–0.84 °C and 0.92–1.02 °C across different model types for the temporal predictions at single and multiple sites respectively. For the PUB scenario, we used a bootstrap aggregation approach using models trained with different subsets of data, for which an ensemble XGBoost implementation outperformed all other modeling configurations (median RMSE 0.62 °C).The ML models improved median monthly T_s estimates compared to baseline statistical multi-linear regression models by 15–48% depending on the site and scenario. Air temperature was found to be the primary driver of monthly T_s for all sites, with secondary influence of month of the year (seasonality) and solar radiation, while discharge was a significant predictor at only 10 sites. The predictive performance of the ML models was robust to configuration changes in model setup and inputs, but was influenced by the distance to the nearest dam with RMSE <1 °C at sites situated greater than 16 and 44 km from a dam for the temporal single site and regional scenarios, and over 1.4 km from a dam for the PUB scenario. Our results show that classical ML models with solely meteorological inputs can be used for spatial and temporal predictions of monthly T_s in pristine and managed basins with reasonable (<1 °C) accuracy for most locations.

Keywords: stream water temperature; machine learning; catchments; modeling; predictions in unmonitored basins (PUB)

Citation: Weierbach, H.; Lima, A.R.; Willard, J.D.; Hendrix, V.C.; Christianson, D.S.; Lubich, M.; Varadharajan, C. Stream Temperature Predictions for River Basin Management in the Pacific Northwest and Mid-Atlantic Regions Using Machine Learning. *Water* **2022**, *14*, 1032. https://doi.org/10.3390/w14071032

Academic Editor: Kaishan Song

Received: 18 February 2022
Accepted: 17 March 2022
Published: 24 March 2022

1. Introduction

The Earth's rivers and streams are under increasing stress due to climactic and urban changes including increased air temperatures, changing precipitation patterns, and more frequent disturbance events [1,2]. In particular, stream water temperature (T_s) is a master water quality variable that controls several physical, chemical, and biological processes, and has economic importance for industries such as thermoelectric power production [3,4] and fisheries [5]. Stream temperatures are projected to rise in the future due to expected increases in air temperatures(T_a) [1,6], which could have negative impacts on stream ecosystems such as degradation of fish habitats and increased likelihood of algal blooms [7–9]. Watershed management strategies and regulatory criteria often include maintenance of T_s as a key objective [10,11].

Accurate T_s predictions across local (reach to catchment) to regional (multi-catchment to basin) spatial scales and daily to decadal temporal scales can enable a scientific understanding of the processes affecting T_s and provide useful information for different types of watershed management decisions [12,13], particularly in the face of increasing extreme events and climate change [14–16]. While daily T_s forecasting is extremely useful for operational purposes, monthly and seasonal hydrological predictions are also important for long-term planning of catchment-scale water management solutions, and can be generated at higher spatial resolution with lower computational expense compared to predictions with higher temporal resolution [17]. Moreover, because the majority of the streams reaches are not monitored for T_s, models that can be used broadly for predictions in unmonitored basins (PUB) are especially useful [18]. Notably, models that predict T_s in both pristine and managed catchments (e.g., catchments containing dams, diversions) can provide actionable information for decision making.

Stream water temperatures can be affected by several factors that include climate (e.g., air temperature, solar radiation, and wind velocity), hydrological processes (e.g., advection from upstream reaches, snowmelt and runoff, groundwater exchange), land cover (e.g., riparian shading), and human activities (e.g., discharge from thermal power plants, dam releases, diversions) [19–22]. Thus modeling water temperatures, particularly for decision-relevant spatial and temporal scales using purely process-based models based on physical equations can be complex and scale-dependent, and more importantly lack adequate representation of human influences. For example, process models such as PRMS-SNTEMP [23], and MOSART-heat coupled to a water management module [20,24] incorporate thermodynamics and energy balance have made considerable advances in regional-scale T_s predictions, but do not fully account for the range of possible anthropogenic activities influencing T_s [24,25]. These models are also computationally expensive to run and require extensive data as inputs, which limits their applicability to regions that do not have the required datasets [26].

Several studies have alternatively used statistical approaches such as the popular multi-linear regression (MLR) method, nonlinear regression, kriging regression, logistic regression, and geographically weighted regression to predict T_s for both monitored and unmonitored basins (PUB) [26–30]. Hybrid statistical models have also been developed to improve upon existing statistical methods by solving analytical heat equations [31,32] and have used model predictions to understand the effects of watershed management and changing climate on stream temperature regimes [33–37]. Hybrid statistical estimates for temporal T_s predictions have fewer input requirements than physically based models, but are generally trained for specific catchments or regions [38] because they require river-specific information and calibration, and are thus more difficult to generalize for large spatial domains.

Machine learning (ML) methods, a subset of statistical methods that learn patterns from large datasets, are being increasingly used for hydrological predictions of streamflow, T_s and other water quality variables at a variety of temporal and spatial scales [18,26,39–42]. A few studies use classical ML regression approaches such as support vector regression [43], random forests [44], and gradient boosted trees [45] for monthly and daily T_s predictions.

These studies obtained relatively good performance for a variety of catchments (RMSE 0.5–1.5 °C) [42,46–49]. However, most prior implementations of classical ML methods train the models using historical T_s data from a given location for future forecasts at the same locations (hereafter referred to as the temporal single site predictions). A few studies have trained a classical ML model with data from multiple stations for catchment-scale T_s prediction [48]. To our knowledge no study has built a classical ML model for pristine and dam-impacted sites for regional, multi-year T_s predictions (hereafter referred to as temporal regional predictions).

More recently, deep learning (DL) models such as long short term memory networks (LSTMs) [50] and its variants have been used for daily T_s predictions at regional to continental scales using data from multiple sites to train a single model. For example, Jia et al. [51] developed a hybrid approach using graph-based LSTM pre-trained with a physics model and trained using 24 years of observations for daily T_s predictions across the Delaware River basin spanning pristine and human-impacted reaches, illustrating the value of combining process-based and ML models for data-sparse spatial predictions [51]. Yet, generalization of hybrid models to new regions can involve significant computation and data, because they have to incorporate outputs from a process-based model implemented for the region of interest. Rahmani et al. [40] used an LSTM model for daily T_s predictions in 118 pristine catchments across the continental US using data on meteorology, discharge (Q), and spatial catchment characteristics. A later study extended the LSTM model to make daily T_s predictions for >400 catchments in the U.S. with different levels of T_s data availability including human-impacted and unmonitored catchments [18]. When using T_s models for prediction, one important use case is prediction of T_s at ungaged or unmonitored basins (PUBs). When modeling PUB scenarios, models must be trained without monitoring data such as T_s and Q measurements often used to train statistical and process-based models; instead models must transfer information between catchments to new locations using widely available data. Statistical models with limited inputs have been employed for such predictions, but have limited prediction accuracy generally around 1.5 °C [32]. LSTM models trained with available catchment data have also been tested for PUB predictions [18,52], with Rahmani et al. [18] achieving notably high accuracy (RMSE 1.13 °C). However, no other studies have tested the use of classical ML for PUB predictions of T_s at large spatial scales, particularly without using catchment metadata that are not broadly available for all watersheds.

Here, we demonstrate the use of two classical ML models, an extreme gradient boosted tree algorithm XGBoost (XGB) [53] and support vector regression (SVR) [43], for monthly T_s predictions in 78 catchments within the Mid-Atlantic and Pacific Northwest hydrologic regions of the United States. We build models for temporal predictions using historical T_s records from single and multiple sites, and for spatiotemporal PUB predictions [18]. These sites are located in pristine and managed catchments with diverse climatology, geology, and land use. Our work adds unique contributions to existing T_s modeling research by (1) showing that classical ML models trained using data from multiple pristine and managed catchments can be used for regional T_s predictions with good performance (0.92–1.02 °C) when validated using long-term records, (2) demonstrating the first use of an ensemble XGBoost algorithm for spatiotemporal PUB, (3) exploring whether ML models that only used inputs that are easily available for any spatial location, namely meteorological data from high-resolution gridded products and simple catchment attributes (latitude, longitude, and elevation), can estimate T_s with reasonable accuracy, and (4) conducting a robust ML sensitivity analysis that examines the utility of different hyperparameter optimization approaches, and changes in model inputs and training.

2. Data and Methods

2.1. Data Sources and Software

We obtained T_s and Q data from the U.S. Geological Survey (USGS) National Water Information System (NWIS; [54]), and meteorological data from the Daymet data product

version 4 [55] in the Mid-Atlantic and Pacific Northwest Hydrological regions, identified by Hydrologic Unit Code (HUC 02, HUC 17, respectively). We used a custom data integration tool BASIN-3D (Broker for Assimilation, Synthesis, and Integration of eNvironmental Diverse, Distributed Datasets) to retrieve and harmonize data [56]. BASIN-3D retrieves USGS data and metadata (e.g., latitude, longitude, elevation) by querying the NWIS daily values service by the regional HUC code ('02' or '17' for this study), and meteorological data from a location's Daymet grid cell by querying the Daymet Single Pixel Extraction Tool [57]. The queries returned from NWIS and Daymet are harmonized into a uniform format and exported as an HDF5 file [58].

We also used the GAGES-II dataset [59], which contains extensive metadata on watershed and site characteristics compiled from multiple data sources for catchments with USGS stream gauges in the conterminous U.S. (CONUS) that have at least 20 years of Q records since 1950. Specifically, the GAGES-II dataset was used to obtain drainage area and information on major dams for available stations within our selected catchments (73/78 stations have metadata in the GAGES-II dataset).

We developed all the model and analysis code in Python . We implemented the ML models using primarily scikit-learn [60] for the MLR and SVR algorithms, package xgboost for XGB [53] and hyperopt [61] for tuning hyperparameters during training. We primarily used the package pandas [62,63] for data analysis. Version numbers of packages used are specified in the corresponding data and code release [58].

2.2. Sites/Station Selection

We used BASIN-3D to obtain daily T_s and Q for water temperature monitoring stations with >20 years of data records from 1 January 1980–1 December 2020. From the returned results, we selected 78 out of 93 stations, which had >100 months (8.3 years) of co-located T_s and Q observations to have sufficient training and testing data at the monthly scale. Of the 78 selected stations, 24 were in the HUC 02 Mid-Atlantic region and 54 in the HUC 17 Pacific Northwest region (Figure 1). These stations are used to monitor T_s in catchments with diverse attributes such as size, elevation, drainage area, and human disturbances (Table S1). For this paper, we refer to a catchment as the area monitored by a stream gauge station (identified using the 8-digit HUC code in the "STAID" column of GAGES-II or NWIS site number) with its corresponding boundary as delineated by the GAGES dataset. The minimum, median, and maximum drainage areas of these catchments are 5, 421, and 29,952 km^2 respectively, while catchment elevations vary from 0 to 1466 m, with a median elevation of 293 m. Average monthly T_s at the stations vary from 6.1 to 14.1 °C with a median of 10.8 °C. The stations also have varying levels of human activity; for example, the number of major dams in the catchments ranges from 0 to 123, with 46 out of 78 stations having at least one major dam in the station's corresponding watershed. Catchments with human activity are denoted as non-reference (NR) sites in GAGES-II (57 out of 78), while pristine catchments are classified as reference (R) sites (16 out of 78).

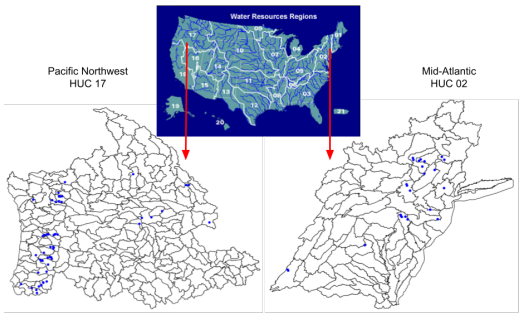

Figure 1. Map of study area showing the hydrologic regions within the continental United States (CONUS) and locations of water temperature monitoring stations within different watersheds of the mid-Atlantic (HUC 02) and Pacific Northwest (HUC 17) regions. Inset map of the water resources regions obtained from the USGS [64].

2.3. Model Setup

2.3.1. Model Description

For this study, we selected two classical ML models: support vector regression (SVR) and the extreme boosted regression tree algorithm XGBoost (XGB) which have been demonstrated to achieve high performance across a range of applications [65]. Deep learning models were not tested for this application, because they have been demonstrated to perform better for large datasets [66,67], and the size of the training dataset for monthly T_s was small (26,976 records across all stations). We compared the ML model performance to a baseline statistical multi-linear regression model (MLR), a common statistical technique that has been used frequently in previous T_s studies (see references in [32]). A MLR is a variation of linear regression for cases with more than one variable predictor which predicts a target variable ($\hat{\boldsymbol{Y}}$, in this case representing T_s) using a linear combination of variable predictors at each time step (bmX). The equation for MLR is given below, where $\hat{\boldsymbol{y}}$ is solved for by minimizing error in the model, epsilon ($\boldsymbol{\epsilon}$).

$$\hat{\boldsymbol{y}} = \boldsymbol{\beta X} + \boldsymbol{\epsilon} \tag{1}$$

Model training consists of finding coefficients for each predictor, which is performed by minimizing the loss function (error between observed and predicted target variable) over the record of training data. The coefficients obtained from model training are then used to make predictions at sample points in the test period. For more information on MLR, see Olive [68].

The SVR, on the other hand, is a nonlinear regression model based upon the support vector machine (SVM) classification algorithm developed by Cortes and Vapnik [69]. The SVM algorithm performs binary classification by generating an n-dimensional hyperplane with a decision boundary determined by training data points or "support vectors" to separate data into classes. To better separate high-dimensional, non-linear data with a hyperplane, SVM uses a kernel function (e.g., polynomial, Gaussian, radial basis function, etc.) to reduce the dimensions of data. The SVM algorithm then maximizes the boundary between support vectors with a penalty for misclassification to obtain an optimal hyperplane equation for the separation of classes. The SVR is the regression counterpart of SVM [43,70], and uses the same approach but instead maximizes the number of points within the boundary to find a hyperplane equation of best fit. This hyperplane equation is used to obtain a continuous value for new testing data points (rather than a discrete class as for SVM). The equation for SVR using a kernel function ϕ is given below:

$$\hat{\boldsymbol{y}} = \sum_{i=1}^{N} w_i \phi(\boldsymbol{X_i}, \boldsymbol{X}) + b \tag{2}$$

All SVR models here used a Gaussian kernel function, with hyperparameters of ϵ, C, and γ tuned using hyperparameter optimization (HPO).

The XGB is an ensemble tree-based algorithm developed by Chen and Guestrin [53] which makes use of gradient boosted trees as first introduced by Friedman et al. [45]. The XGB models train an ensemble of regression trees iteratively through a process called "boosting". Each regression tree is built to predict a given sample's value by mapping a sample's input data to a leaf with a continuous score. Each successive tree is built to predict the previous tree's errors or residuals, rather than to predict a specific target variable, and is then combined into the model ensemble to improve upon the model's prediction. The equation for forming an XGB model is given below [42], where the dependent variable $\hat{\boldsymbol{Y}}$ is solved for using a series of M successive boosted trees.

$$\hat{\boldsymbol{y}} = 0.5 + \sum_{m=1}^{M} \eta f_m(\boldsymbol{X}) \tag{3}$$

When adding new models, XGB uses gradient descent to minimize loss. The XGB is a popular choice for classical ML algorithms, and can also make predictions using sparse data.

2.3.2. Modeling Scenarios

We tested the use of classical ML models for two objectives: temporal predictions for unseen time periods and spatiotemporal predictions for unseen geographic locations. To address these two objectives, we trained three model scenarios (Figure 2). The first two scenarios involve temporal predictions: (1) single station models (hereafter referred to as 'single station' or SS) trained and tested on an individual station's data, which is the most commonly used scenario in prior ML studies of T_s, and (2) multi-station models for the individual and combined HUC 02 and HUC 17 regions (hereafter referred to as 'regional'). The third scenario is the PUB predictions, where we used an bootstrap-aggregation ensemble approach (hereafter referred to as 'ensemble PUB' or 'PUB'), trained on a group of stations and tested on withheld stations for each HUC region. For each spatial and temporal configuration, we evaluated model performance using different metrics (Section 2.4.1), and tested model sensitivity to different architectures, hyperparameter optimization (HPO) and input features using different modeling configurations (Table 1). More information on model training for each configuration is provided in Section 2.3.5.

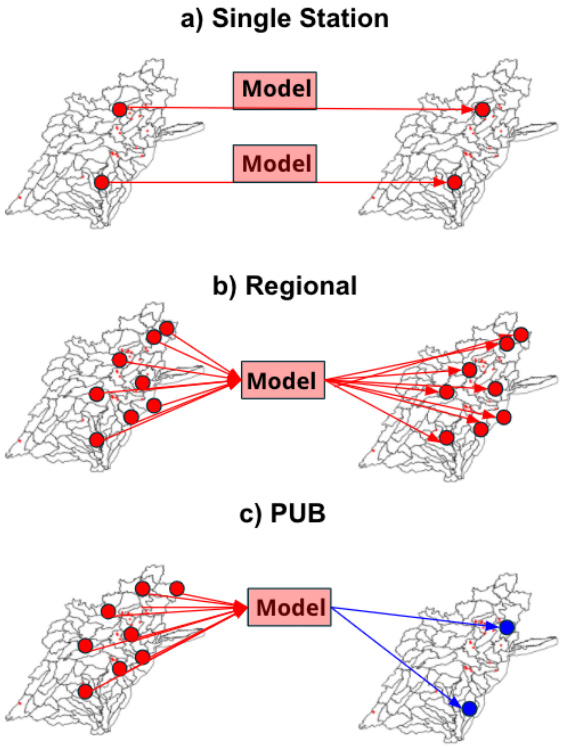

Figure 2. Graphical illustration of the three regional-scale modeling scenarios used in this study: (**a**) temporal single station predictions (SS), (**b**) temporal regional predictions (regional), and (**c**) spatiotemporal PUB predictions (PUB). Sites used for training data are shown on the left and for testing data are shown on the right. Sites with training data seen by the model are colored in red, and sites with no data seen by the model prior to testing are in blue.

Table 1. Details for all modeling configurations tested across the three scenarios used in this study. The 'Configuration Name' represents each configuration's training subset (HUC codes), training attributes (Att or noAtt), HPO (HPO or noHPO). 'MR' represents multi-region training while SR represents single region training. The optimal configuration(s) for each scenario are highlighted in bold and italics.

Configuration Name	Training Subset [1]	Training Attributes	HPO
	Temporal Single Station Scenario		
SS_HPO	***HUC 02 + 17***	***NA***	***HyperoptTPE***
SS_noHPO	HUC 02 + 17	NA	None
	Temporal Regional Scenario		
02_17_MR_noAtt_noHPO	HUC 02 + 17	lat/lon/elev	None
02_17_MR_noAtt_HPO	HUC 02 + 17	None	HyperoptTPE
02_17_MR_Att_HPO	HUC 02 + 17	None	HyperoptTPE
02_17_MR_Att_noHPO	HUC 02 + 17	lat/lon/elev	None
02_17_MR_Att_Drain_noHPO	HUC 02 + 17 GAGES	lat/lon/elev/drain	None
02_17_MR_Att_Drain_HPO	HUC 02 + 17 GAGES	lat/lon/elev/drain	HyperoptTPE
02_17_SR_noAtt_noHPO	HUC 02 + 17	lat/lon/elev	None
02_17_SR_noAtt_HPO	HUC 02 + 17	None	HyperoptTPE
02_17_SR_Att_HPO	HUC 02 + 17	None	HyperoptTPE
02_17_SR_Att_noHPO	***HUC 02 + 17***	***lat/lon/elev***	***None***
02_17_SR_Att_Drain_noHPO	HUC 02 + 17 GAGES	lat/lon/elev/drain	None
	PUB Scenario		
PUB_02_17_SR	***HUC 02 + HUC 17***	***lat/lon/elev***	***None***

[1] The column 'Training Subset' shows the set of stations that were used across either one or both regions.

2.3.3. Input Feature Selection and Preprocessing

We selected relevant input features from the list of all daily Daymet variables (minimum and maximum T_a averaged to mean T_a, precipitation (*prcp*), vapor pressure (*vp*), shortwave solar radiation (*srad*), snow water equivalent (*swe*), day length(*dayl*), Q, and month of the year (*moy*) (encoded as numeric values 1:12) through literature review and exploratory data analysis (EDA). The final inputs selected were mean T_a (hereafter T_a), *srad*, Q, *prcp*, and *moy*. The first lagged T_s showed significant (p-value > 0.7) autocorrelations for 22 stations, but was not included as an input feature as observed T_s is not available for PUB scenarios.

For the SS scenario, we tested the sensitivity of the *moy* feature engineering using a fuzzy month approach [42] and found no considerable differences in RMSE between using the two methods. We additionally tested the importance of drainage area as an input feature in modeling configurations, as it is indicative of catchment size and can impact the energy balance in the watershed (e.g., determining the amount of snowmelt runoff into streams).

CVA summary of our preprocessing workflow is presented in Figure 3. We preprocessed time series input features by first resampling each variable to the monthly frequency, using a monthly mean for all variables except for *prcp*, where we used a cumulative sum. To minimize edge effects, where models may be unable to capture a signal surrounded by gaps in data, we dropped periods where there were less than 6 months of consecutive T_s measurements. For Q measurements, we additionally scaled the variable to a log scale using the transformation in Rahmani et al. [40] to obtain a more Gaussian distribution. Finally, we standardized each station's data using a standard scalar, which normalizes data to a $[0, 1]$ range before training.

2.3.4. Model Training

Both temporal prediction scenarios (SS and regional) were trained with a 70–30 train-test split where the first 70% of the data was used for training (Figure 3). We trained all 78 SS models using only the corresponding meteorological variables, Q,

and *moy* for the station. For regional models, we additionally ran models with and without static station attributes (latitude, longitude, elevation, and drainage area of the corresponding catchment), and tested different training configurations (Table 1, Section 2.3.5). Regional scenarios with hyperparameter optimization (02_17_*_HPO) were run on the National Energy Research Scientific Computing (NERSC) Center Perlmutter Graphical Processing Unit (GPU) node.

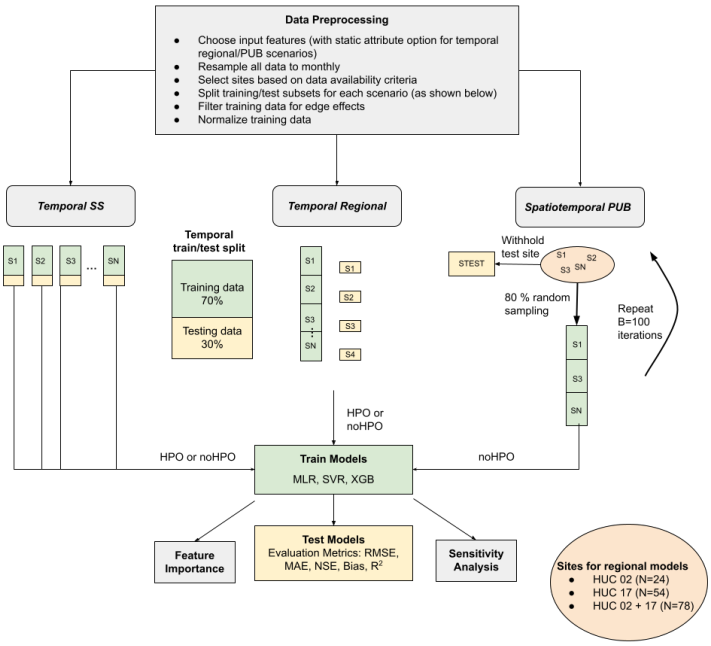

Figure 3. Summary of modelling methodology and workflow. The testing/training data split are shown for each of the three modelling scenarios. For the PUB scenario, an example of one iteration of the algorithm (presented in Algorithm 1) is depicted. S1…SN represent stations in the training subset.

Algorithm 1 The ensemble algorithm used to train the models for the spatiotemporal PUB scenario.

```
B = 100
frac_test = 0.8
for test_station in all_stations do
    train_set = all_stations \ test_station
    n_samples = ceil(len(train_set) * frac_train)
    for i in range(B) do
        train_stations = random.sample(train_set, n_samples)
        models = train_models(train_stations)
        predictions[i]= model.test(test_station)
    end for
    ensemble_mean[test_station] = predictions.mean()
end for
```

The algorithm used for training the PUB scenario follows an ensemble approach with bootstrap aggregation (Algorithm 1). For each PUB configuration, we iteratively withheld a test station (test_station) from the subset of stations chosen for the modeling configuration (all_stations). We then sampled 80% (frac_test) of the remaining stations for training without replacement (n_samples), and ran the model to obtain an ensemble of test predictions for the test_station (Figure 3). This process is repeated to obtain a 100 member ensemble (B = 100) of predictions for each test_station. This process of resampling and training models with varying training datasets is known as bootstrap aggregation or "bagging" [71]. The individual ensemble predictions were averaged to obtain an ensemble

mean, which is then used to quantify average prediction accuracy. The distribution of these predictions also gives a measure of model sensitivity to input features.

2.3.5. Model Configurations

For each of the modelling scenarios (Figure 2), we tested model sensitivity to several different training configurations (Table 1). The SS and regional model scenarios tested configurations where training was performed with HPO (indicated as '_HPO' in the configuration name) and without HPO ('_noHPO'). More information on hyperparameter optimization is located in SI Section A.2. Hyperparameters for SVR and XGB were tuned using the sequential model-based optimization Tree of Parzen Estimators [72] algorithm with the hyperopt package [73] (HyperoptTPE), a method of sequential model-based optimization. For the SS scenario, the Hyperopt TPE algorithm was compared with random search and grid search and selected because of its computational efficiency and improved model accuracy. All HPO results are calculated using 4-fold cross-validation on the training datasets with 50 iterations of the hyperopt TPE algorithm [72].

Figure 3 summarizes training information and variation in configurations for the three modelling scenarios.

Each configuration was trained with meteorology (mean air temperature ($\overline{T_a}$), solar radiation (srad)), river discharge (Q), and month of the year (moy). The regional model configurations additionally tested the importance of training with input features of widely available static station attributes (lat, lon, and elev indicated by '_Att') to provide some information on climactic influences and geographic proximity into the model. We also ran configurations including drainage area of the corresponding catchments ('_Att_Drain') for stations that had the metadata available in GAGES-II. To test the generalizability of the models across hydrologic regions, we also trained the regional models using data from stations within an individual region (single region or 'SR', i.e., separate models for stations within HUC 02 and HUC 17) and stations from both regions (multi-region or 'MR', i.e., a single model for stations within HUC 02 and HUC 17). The results from the regional configurations were used to determine PUB configurations. For example, the PUB scenario was run with static attributes ('_Att') and without HPO ('_noHPO') because the regional models did not benefit from HPO (see Section 3.2.3), and for computational efficiency.

2.4. Model Evaluation

2.4.1. Evaluation Metrics

For all three model scenarios, 5 error metrics were calculated: Root Mean Squared Error (RMSE, Equation (4)), Mean Absolute Error (MAE), Nash-Sutcliffe Error (NSE), and bias as defined by [74]; Equations (S1)–(S3)). R^2 values are not reported as it is comparable to NSE [32]. The metrics were computed by comparing model predictions with observations for the test period.

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

where $\hat{y}_i$ represents the i_{th} modelled T_s prediction, y_i represents the i_{th} T_s observation, and n is the number of observations in the test period.

To test the improvement in model performance under different model configurations, we additionally used RMSE skill score to compare runs with and without hyperparameter optimization and compare ML models in reference to MLR. The equation used to represent RMSE -based skill score is:

$$S_{score} = 1 - \frac{RMSE_{pred}}{RMSE_{ref}} \tag{5}$$

where $RMSE_{ref}$ is the reference RMSE (e.g., baseline MLR) and $RMSE_{pred}$ is the prediction RMSE (e.g., SVR ML model).

Skill scores are interpreted as a fractional improvement over reference model conditions. Skill score values are positive if a prediction (e.g., with HPO) has lower errors

compared its reference prediction (e.g., no HPO), negative if vice versa, and zero if the errors are equal. Lastly, for each model configuration, we also quantified performance at extremes in T_s ($T_s < 0.05$ and $T_s > 0.95$ quantiles) for a given station's temperature record.

2.4.2. Feature Importance

Tree-based algorithms such as XGB make predictions based on a series of splitting samples at 'nodes' of a tree, each based on a single input feature. Analysis of these splits provides additional information about how a trained model uses specific input features. For XGB, we calculate a fractional input feature importance (FI) score for each model configuration using the feature importance gain metric [53], a measure of the increase of accuracy a given feature brings to the branch it is on.

3. Results

The results from each of the model runs across different scenarios, configurations and model types (MLR, SVR, XGB) are summarized in Table 2. In this section and in the discussion (Section 4), we present model performance for each of the three modeling scenarios based on the optimal configuration that typically had lowest median RMSE. For the regional and PUB scenarios, all error metrics presented are averages across 78 stations (or 73 stations for the '_Drain' configurations) regardless of whether the models were trained for the single region (SR) or multi-region (MR) configurations. For each of the scenarios, we also identify the features that were most important for the model predictions and present model sensitivities to the different configurations.

3.1. Temporal Single Station (SS) Scenario

3.1.1. Model Performance

The optimal performance for the SS scenario was achieved using the SVR model with hyperparameter optimization (SS_02_17_HPO configuration, median RMSE = 0.69 °C). For the 78 SS_HPO models, both SVR and XGB consistently outperformed the standard regression benchmark (MLR) model, with median skill scores of 0.25 and 0.15 for SVR and XGB, respectively (Figure 4a). The ML models also outperformed MLR based on MAE and NSE (Figure S1). All models, however, have some bias in predictions, ranging from −0.02 to 0.02 °C.

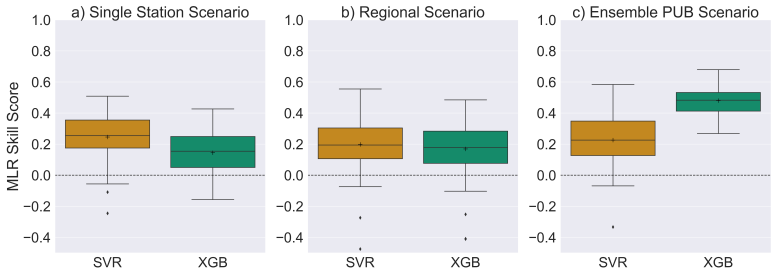

Figure 4. Skill Score of ML Models (SVR, XGB) in reference to baseline MLR Model for (**a**) SS, (**b**) Regional, and (**c**) PUB scenarios.

For the SS_HPO configuration, the majority of the stations (37, 60, and 54 out of 78 stations for MLR, SVR, and XGBoost, respectively) had RMSE <1 °C. A few stations (11, 6, and 4 for MLR, SVR, and XGB respectively) had RMSE > 1.5 °C, and all but one of these stations had at least one major dam in the catchment. The one exception was a station for which we did not have information on major dams as it was not included in GAGES-II.

The SVR and XGB also outperformed MLR at time periods with extremes in T_s. Median RMSE at extremes for the SS_HPO configuration was 1.28, 0.72, and 0.90 °C for MLR, SVR, and XGB respectively, showing a slight decline in accuracy in comparison to the full test period (Table 2).

Table 2. Root mean squared error (RMSE) metrics for each model configuration presented in Table 1. The metrics include Q1 (25th percentile), mean (μ), median or the second quartile Q2 (50th percentile), third quartile Q3 (75th percentile), and median performance at extremes (Ext). The optimal configuration is presented in bold and italicized font.

Configuration Name	MLR				SVR				XGB			
	Q1	μ/Q2	Q3	Ext [1]	Q1	μ/Q2	Q3	Ext [1]	Q1	μ/Q2	Q3	Ext [1]
Temporal Predictions—SS												
SS_02_17_HPO	0.72	1.07/1.03	1.23	1.28	***0.53***	***0.81/0.69***	***0.93***	***0.72***	0.62	0.89/0.84	1.04	0.72
SS_02_17_noHPO	0.72	1.07/1.03	1.23	1.28	0.59	0.89/0.80	1.06	1.09	0.68	0.94/0.92	1.06	1.09
Temporal Predictions—Regional												
02_17_MR_noAtt_noHPO	0.93	1.35/1.32	1.67	1.58	0.79	1.15/1.11	1.35	1.10	0.76	1.10/1.13	1.37	1.10
02_17_MR_noAtt_HPO	0.93	1.35/1.32	1.67	1.58	0.80	1.18/1.18	1.42	1.21	0.90	1.23/1.20	1.44	1.21
02_17_MR_Att_HPO	0.93	1.35/1.32	1.67	1.58	0.78	1.12/0.96	1.25	1.10	0.89	1.24/1.25	1.44	1.10
02_17_MR_Att_noHPO	0.93	1.35/1.32	1.67	1.58	0.73	1.08/0.94	1.27	1.06	0.77	1.05/1.03	1.30	1.06
02_17_MR_Att_Drain_noHPO [2]	0.91	1.31/1.25	1.62	1.57	0.70	0.99/0.86	1.19	1.02	0.73	1.02/1.02	1.22	1.02
02_17_MR_Att_Drain_HPO [2]	0.91	1.31/1.25	1.62	1.57	0.81	1.16/1.08	1.26	1.34	0.91	1.23/1.22	1.41	1.34
02_17_SR_noAtt_noHPO	0.97	1.30/1.24	1.44	1.56	0.75	1.11/0.99	1.30	1.19	0.80	1.09/1.08	1.28	1.19
02_17_SR_noAtt_HPO	0.97	1.30/1.24	1.44	1.56	0.75	1.14/1.01	1.32	1.14	0.94	1.21/1.17	1.42	1.14
02_17_SR_Att_HPO	0.97	1.30/1.24	1.44	1.56	0.74	1.09/0.93	1.25	1.05	0.90	1.17/1.08	1.39	1.05
02_17_SR_Att_noHPO	0.97	1.30/1.24	1.44	1.56	***0.69***	***1.03/0.92***	***1.21***	***1.14***	0.77	1.03/1.02	1.25	1.14
02_17_SR_Att_Drain_noHPO	0.96	1.26/1.21	1.41	1.55	0.65	0.95/0.89	1.14	1.05	0.76	1.01/0.98	1.21	1.05
Spatial Predictions—PUB												
PUB_02_17_SR	0.92	1.30/1.20	1.47	1.56	0.70	0.99/0.89	1.15	1.08	***0.50***	***0.64/0.61***	***0.72***	***1.08***

[1] Extremes defined as T_s < 0.05 percentile or >0.95 percentile. [2] These configurations were calculated with 73 stations in GAGES-II dataset that had drainage area estimates.

3.1.2. Feature Importance

Feature importance (FI) scores for the SS_HPO configuration showed that the XGB model predicts primarily based upon $\overline{T_a}$ and to a lesser extent *moy* ($\overline{T_a}$; mean FI = 0.82, *moy*; mean FI = 0.1), with all other predictors having mean FI < 0.05 (Figure 5a). For 7 stations with $\overline{T_a}$ FI < 0.5, *moy* was used as a primary predictor of T_s (mean FI score of 0.55). The models generally performed poorly for these stations (RMSE > 1 °C). For temporal predictions, 10 catchments (6 of which had major dams) used Q as an important predictor (FI > 0.05) but all other stations had no significant FI for Q (<0.05). Thus, Q was a slightly more important predictor for T_s in a few stations located in a dammed catchment.

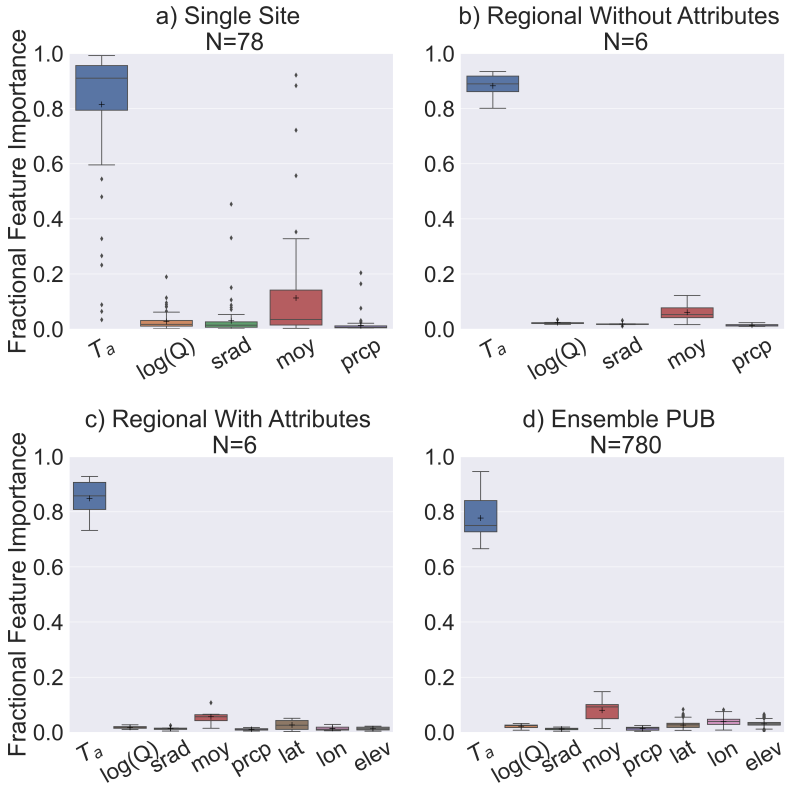

Figure 5. XGB FI scores for (**a**) SS scenario (SS_02_17_HPO configuration) (**b**) regional Models without attributes (**c**) regional Models with attributes and (**d**) all PUB scenario configurations.

3.1.3. Model Sensitivity

The SS_HPO configuration outperformed the SS_02_17_noHPO configuration with decreases in RMSE of 0.11 and 0.08 °C for the SVR and XGB models respectively. This indicates a marginal improvement in performance due to HPO, corresponding to 6% and 5% improvement in prediction accuracy on average for SVR and XGB respectively (Figure S2). Notably, there were a few stations (13 and 23 stations for SVR, and XGB respectively) that did not improve from HPO as indicated by a skill score < 0. The chosen optimal hyperparameters from the hyperopt TPE algorithm varied substantially between stations; for example, the optimal value of the hyperparameter 'C' which controls the penalty for misclassified points varies from 2 to 980, spanned nearly the full predefined search space (see supplemental dataset for details).

3.2. Temporal Regional Scenarios

3.2.1. Model Performance

The different regional runs had similar model performance with small variations in RMSE depending on training configurations (Table 2, Section 3.2.3). We chose the optimal regional model configuration as the one trained with single regions, static attributes and no

hyperparameter optimization ('02_17_SR_Att_noHPO') as it had one of the lowest median RMSEs. For this configuration, SVR and XGB again outperformed MLR for MAE and NSE error metrics. Although configurations with drainage area ('_Drain') had slightly lower median RMSEs, they were not chosen as the optimal configuration for consistency because drainage area values were not available for all 78 stations. Additional results for the SR configurations are presented in Table S2.

All models again showed some bias in predictions with SVR and XGB having a slightly positive mean bias (Figure S3). Median RMSE for the optimal regional configuration (02_17_SR_Att_noHPO) increased by 0.23 °C for SVR and 0.18 °C for XGB compared to the optimal SS configuration. The SVR slightly outperformed XGB with median RMSE of 0.92 °C, whereas XGB achieved median RMSE of 1.02 °C (Table 2). Median RMSE at extremes was higher than overall median RMSE with scores of 1.56 °C for MLR and 1.14 °C for both SVR and XGB.

3.2.2. Feature Importance

The XGB FI scores for regional models were similar to the SS configuration. At the regional scale, $\overline{T_a}$ emerged as the primary predictor based on the average XGB FI calculated for all 6 regional configurations without and including static attributes (Figure 5b,c respectively). For the optimal regional configuration (02_17_SR_Att_noHPO),$\overline{T_a}$ was the main predictor with median FI of 0.92. For the optimal configuration *moy* also was a notable predictor (average FI = 0.07). However, all other features did not contribute substantially to the predictions with FI < 0.05. Although the static attributes improved overall model performance, they did not contribute substantially to the predictions with FI < 0.05 (Figure 5d). Median FI for $\overline{T_a}$ was slightly higher for runs with no static attributes (median FI score 0.89) than for runs with static attributes (median FI score 0.86). For the 73 stations for which drainage area was available, its inclusion in the inputs lowered median RMSE by 0.03–0.09 °C depending on the model (Table 2), but the drainage area FI was still <0.05.

3.2.3. Model Sensitivity

Although there were slight differences in RMSE across different model runs, the performance was robust to changes in training configuration including HPO and addition of static attributes (Table 2). The static attributes did not improve model performance for MLR, but on average slightly increased accuracy for SVR and XGB (median skill score ranging 0–0.03 for the optimal configurations across different models). For runs with regions trained separately (02_17_SR_Att_noHPO), 51 stations for SVR and 66 stations for XGB had an increase in performance (positive skill score) using static attributes. For multi-region training (02_17_MR_Att_noHPO) 43 stations for SVR and 33 stations for XGB had increased performance (positive skill score). Thus, on average the accuracy of predicting T_a for each of the stations improved by using attributes for single region runs than for multi-region runs.

The performance of models trained with individual and multiple regions was similar with median RMSE with the SR configuration achieving slightly higher accuracy by 0.075 °C for MLR, and 0.015 °C for both SVR and XGB compared to the MR configuration (Figure S4). The performance varied between different regions for SR scenarios with the accuracy in HUC 02 being slightly higher than in HUC 17. However, this could be biased due to different sample sizes in the data since there were 24 stations from HUC and 54 from HUC 17 (Table S1). At the regional scale HPO showed no substantial improvement in prediction accuracy (median skill score SVR = 0.0, XGB = −0.1, Figure S2).

3.3. PUB Scenario

3.3.1. Model Performance

For the ensemble models used in the PUB scenario, the SVR and XGB similarly outperformed MLR with median skill scores of 0.23 and 0.48, respectively (Figure 4c). However, the XGB achieved substantially higher accuracy compared to SVR (median RMSE of 0.60 and 0.63 °C for HUC 02 and HUC 17, respectively, Table 2). Model performance

showed reduction in bias compared to the temporal scenarios (Figure S5) and was robust to differences in the length of the testing period (Figure S6).

3.3.2. Feature Importance

Although there was considerable variation in FI scores between ensemble runs, for all runs at all stations, $\overline{T_a}$ was consistently the primary predictor (median FI score = 0.8), with slight influence from *moy* (median FI score = 0.1). The static attributes again had low FI < 0.05 (median FI scores of 0.03, 0.04, and 0.03 for lat, lon, and elevation respectively) (Figure 5d).

3.3.3. Model Sensitivity

There was significant variation in RMSE between ensemble members that were tested with one station's data, but trained with different datasets. This indicates that the different models used are individually sensitive to the choice of training data, but the ensemble approach substantially reduces the variability of model performance. Both XGB and SVR showed a relatively high range of differences in RMSE between ensemble members; SVR had a median range in RMSE of 0.72 °C while XGB was more sensitive to the choice of training datasets with median range in RMSE of 1.12 °C (Figure 6). However, the MLR had little variation in RMSE between ensemble members (Figure 6), with median range in RMSE of only 0.10 °C.

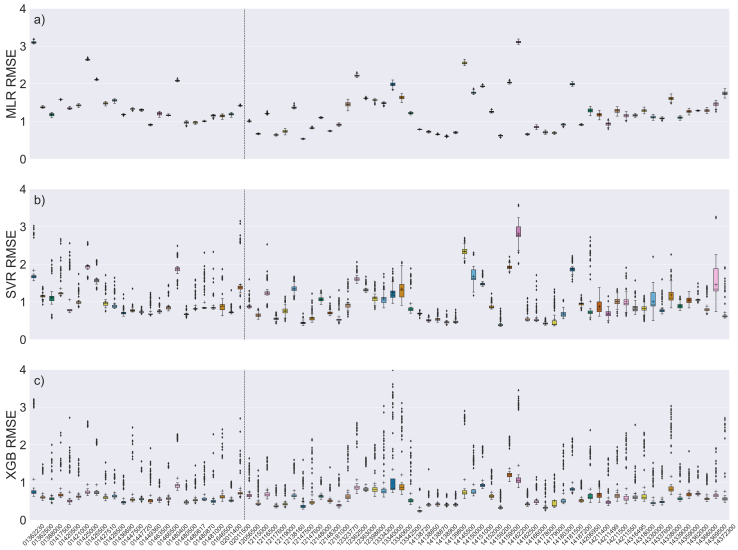

Figure 6. RMSE distributions PUB scenarios of (**a**) MLR, (**b**) XGB and (**c**) SVR models. HUC 02 stations are located to the left of the dashed vertical line, and HUC 17 stations are located to the right.

4. Discussion

4.1. Comparison of Machine Learning and Statistical Model Performance

Our results show that the classical ML models used in this study are 15–48% more accurate than the statistical MLR model for monthly T_s predictions depending on the site and scenario (Table 2). In general, for the temporal SS and regional scenarios, the SVR models achieved the highest accuracy for sites across different catchments. For the spatiotemporal PUB predictions, the ensemble XGB models achieved substantially higher accuracy, being the only models to accurately predict T_s (<1° RMSE) across stations with varying catchment sizes and number of dams (see examples in Figures 7 and S7). These results suggest that the boosted tree models benefited from bootstrap aggregation where ensembles of models were trained with different datasets (chosen with random sampling in this study). Such ensembles have previously shown improvement in predictions both in methodological theory [75] and in hydrological applications [52,76–78] such as operational models predicting Q [79,80]. Several of these studies build ensembles by changing random seeds or

initial conditions in contrast to our use of bootstrap aggregation to train ensemble members with different datasets. Classical ML models such as boosted trees are particularly useful for such ensemble approaches as they are relatively simple and can run many ensemble members efficiently.

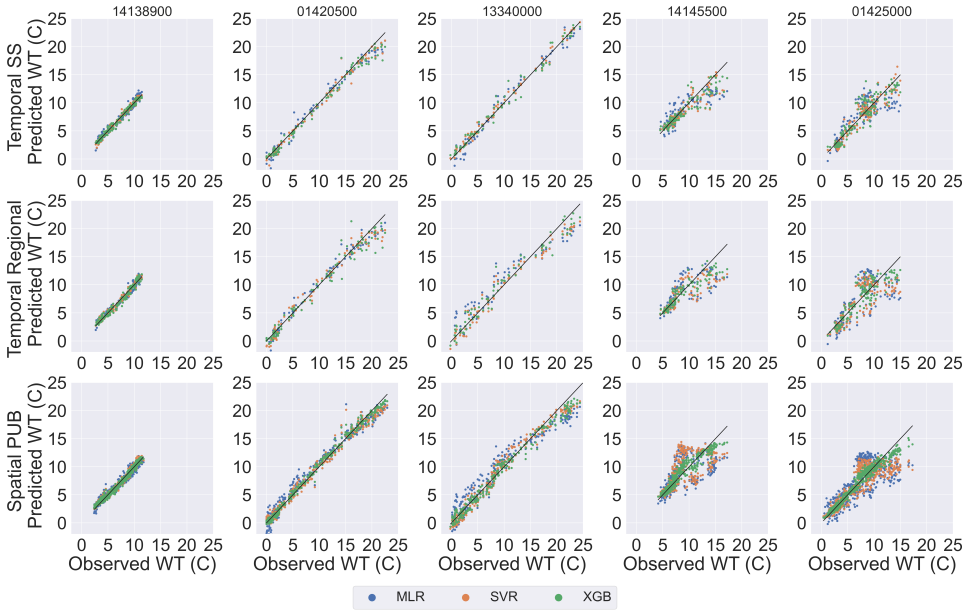

Figure 7. Observed vs. predicted T_s for 5 stations in 3 pristine catchments and two dammed catchments. The pristine catchments include: the North Fork of Bull Run River near Multnomah Falls (14138900)—small (Drainage area: 21.7 km^2), Beaver Kill at Cooks Falls NY (01420500)—mid size (627 km^2), and Clearwater River at Orofino ID (13340000)—large size (14269 km^2). The two dammed catchments include: the Middle Fork of Willamette River (14145500)—large size (1017 km^2) and the West Branch of the Delaware River (01425000)—large size (1181 km^2).

While the performance of our monthly models cannot be directly benchmarked against accuracy metrics of daily T_s predictions from prior studies, we provide a brief comparison with other ML models that had similar training scenarios. Our models implemented for the temporal SS scenario achieve comparable accuracy to prior studies that correspondingly trained ML models for single sites using meteorology (and in some cases discharge) data, which represent the vast majority of T_s predictions in published literature. For example, Feigl et al. [42] achieved a median RMSE of 0.55 °C for 10 catchments in Austria that included stations with human impacts and dams when comparing six different ML model architectures. In this study, we did not use DL models for monthly T_s predictions, but note that the performance of our temporal regional models are somewhat comparable to predictions using LSTMs that achieved a median RMSE of 0.69 °C for 118 undammed, pristine catchments across the CONUS [40]. Furthermore, our classical ML models had greater accuracy than the LSTM models used for CONUS-scale PUB predictions in Rahmani et al. [18] when comparing predictions for five stations in the Northwest US, which had a monthly aggregated mean RMSE of 1.07 °C in Rahmani et al. [18]. Three out of the five stations used for this comparison met our data requirements (Section 2.2), for which our XGB ensemble PUB achieved a mean RMSE of 0.55 °C. Although this comparison with [18] was done on a small number of stations, it indicates that classical ML models can achieve an aggregate monthly accuracy that is comparable to a DL approach, but with substantially fewer parameters to adjust. Our models also out-performed other statistical methods for the same stations as presented in Gallice et al. [32], where the RMSE was 1.3 °C.

The low computational cost of the ML models used here also allowed us to explore a full range of uncertainties associated with input features and training through different modeling configurations. These model configuration tests showed that our results were

robust to HPO, single vs. multi-region training, and to using additional static attributes. Contrary to studies where DL models improved performance by training with more data across different regions, the classical statistical and ML models used here did not benefit substantially from training with many stations that had different attributes. Furthermore, while HPO can considerably improve model performance for DL models, we did not find any meaningful improvement in performance using HPO for the classical ML models (Figure S2). This is important to note given the substantial computational costs associated with HPO for training large datasets. In contrast, we find that model performance was considerably affected by the choice of training datasets, which varied in the PUB scenario (Figure 6). Both temporal and spatiotemporal predictions were also affected by periods with discontinuous T_s data; these "edge effects" (or difficulty predicting a signal surrounding gaps in data) suggest that more precise strategies for gap filling and quality assurance/quality control could also improve model performance.

4.2. Factors Influencing Monthly Stream Temperature

We tested the influence of several factors known to influence T_s including climate, streamflow, and dams. Overall, our results suggest that meteorological variables encode most of the necessary information for monthly T_s predictions, based on the FI score of $\overline{T_a}$ in both the temporal and spatiotemporal models. Surprisingly, we found that Q was not a significant predictor for a majority of stations. This is counter to the findings of Rahmani et al. [40], where Q was found to be an important input for daily predictions of T_s. This difference may be because the effect of discharge is more apparent at daily time scales, where flow from upstream reaches can significantly impact diurnal T_s variability. However, at coarse monthly scales, our models indicate that the effects of T_a outweigh all other influences. Similarly, the FI scores of static attribute inputs, including drainage area were small, even though their inclusion marginally improved the performance of the temporal regional models (Section 3.2.3) The low FI scores of the static attributes can be partially attributed to information implicitly being encoded in other input features; for example T_a is generally lower in mountainous areas with high elevation.

The spatial distribution and bias of station locations may have also impacted model performance and FI scores, as the sites chosen for this study were not evenly distributed across the HUC02 and HUC 17 regions (Figure 1). In some cases multiple sites were located within the same catchment, while other catchments did not have T_s monitoring data that met the requirements of this study (Section 2.2). Thus, the catchment characteristics of the stations chosen for this study may be more similar than expected from a random sample, potentially impacting model performance and primary predictors. The spatial distributions of RMSE in the HUC 02 and HUC 17 regions for the PUB scenario XGB model are presented in Figure S8.

Our results also showed that the SVR and XGB models can accurately predict monthly T_s for most catchments with major dams across the temporal and spatiotemporal prediction scenarios. However, we found that model performance was impacted by the distance between a station and the nearest major dam. Notably, the models could only achieve RMSE < 1 °C accuracy when the sites were located greater than 16.4, 44.2, and 1.4 km of straight line distance from a major dam for the SS, regional, and PUB scenarios respectively (Figure 8). While the straight line distance retrieved from the GAGES-II dataset does not account for distances along the river, or indicate whether the dam is located upstream or downstream of the gage, this finding suggests that meteorological predictors can adequately capture monthly T_s dynamics for stations which are not proximal to a major dam.

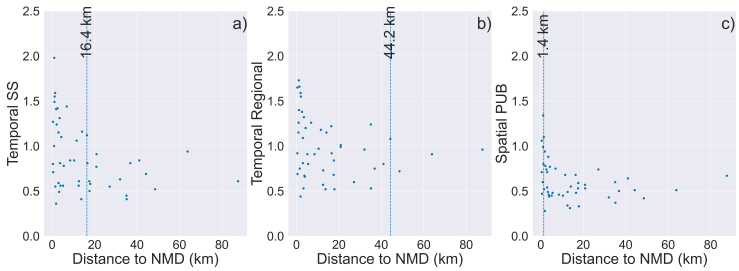

Figure 8. Distance to nearest major dam vs. best RMSE for each optimal model configuration of the three modeling scenarios: (**a**) Temporal SS, (**b**) Temporal Regional, and (**c**) PUB. For each scenario, the distance after which RMSE < 1 °C is achieved is marked with a vertical line.

4.3. Local and Regional Predictions

We implemented three modeling scenarios for T_s predictions from local to regional spatial scales: (a) temporal (local) predictions at a single monitored site, (b) temporal regional predictions across monitored sites, and (c) spatiotemporal PUB predictions for unmonitored sites. The classical ML models used in this study, SVR and XGB, achieve good performance (typically <1 °C RMSE) for all these scenarios. For example, Figure 7 presents predictions across the three modelling scenarios for five catchments of varying drainage area and number of dams, demonstrating general trends in prediction accuracy between scenarios.

The exact scenario that decision makers and other stakeholders can use for local to regional predictions of monthly T_s for will depend on the water management objectives and constraints. For example, if reliable long-term records for a specific site are available, the temporal SS scenario may be most useful for short-term seasonal forecasting of T_s and planning corresponding water management actions. This is because the prediction accuracy decreased slightly for all models when applying the temporal regional scenario in comparison to the temporal SS scenario. Thus, the same models when trained with multiple stations of data, were able to generalize some but not all station-specific dynamics given the simplistic attributes chosen in this study (latitude, longitude, elevation, and drainage area when available). On the other hand, if there is a need to predict T_s at locations where there is no historical data for river basin management purposes, the PUB scenario leveraging historical data records from the same region to transfer to unmonitored locations will be most useful, especially given the accuracy achieved using the ensemble modeling approach.

Overall, we find that classical ML models are relatively easy to implement, and are able to predict monthly T_s with reasonable accuracy (<1 °C) for both local and regional predictions. Given the results of the sensitivity tests to different input features, we find that our models can be used with low data requirements (i.e., widely available gridded meteorological data for inputs and prior T_s data for testing), which will considerably increase the number of sites and regions for which suitable data are available. Prior regional models in hydrology have used data-intensive approaches needing both Q and substantial amounts of watershed metadata (for example, 33 CAMELS attributes used by Kratzert et al. [81] and 55 GAGES attributes used by Rahmani et al. [40]). While adding simplistic static attributes and Q minimally improve our model skill, we find that the classical ML models can learn sufficient catchment-specific information to make reasonably accurate predictions (<1 °C) without data-intensive attribute metadata that are unavailable for the majority of locations, and for ungaged basins where no Q records are available. In particular, we note the ability of the models to predict T_s in dammed catchments, which is particularly useful because previous ML modeling studies (with the exception of Rahmani et al. [18]) have largely focused on PUB predictions in regions with little to no anthropogenic impacts. While the results presented in this study are from the mid-Atlantic and Pacific Northwest hydrologic regions of the US, the models and methodology for predicting monthly T_s are broadly applicable to other regions with different hydrological,

geological, meteorological, and land use characteristics, including locations that have human impacts.

5. Conclusions

Here we have shown how classical ML models can be used for monthly T_s temporal and spatial predictions using three prediction scenarios: temporal SS and regional scenarios, and a spatial ensemble PUB scenario. In total, we tested 17 different configuration scenarios to understand the sensitivity of models to changes in training datasets, hyperparameter optimization and input features. Across all scenarios, we have shown that ML models (XGB, SVR) outperform traditional statistical approaches (MLR). Temporal predictions achieve accuracy comparable to other studies with SS median RMSE 0.69–0.84 °C, and regional median RMSE and 0.92–1.02 °C. Predictions for temporal models were also shown to be robust to changes in model architecture and training.

Spatial predictions in the ensemble PUB scenario showed substantial improvement in model performance particularly for XGB models, where model errors are reduced to median RMSE of 0.63 °C. The ensemble PUB scenario with XGB models are able to predict not only general T_s dynamics, but also dynamics impacted by dams, in large basins, and under extremes in T_s where other models are generally less precise. Few studies have used ensembles of models in a hydrological context [82,83], and to our knowledge, this is the first study to use ensembles to make PUB predictions.

The simplistic inputs of this model (meteorology, Q, and simple static attributes) allow for broad application for water management where extensive metadata and measurements are not available. These models are not only accurate, but also advantageous for predicting T_s in catchments with minimal measurements. While daily models are better for operational forecasting, monthly models can play an important role in near-term seasonal forecasting to plan for and understand future impacts on stream temperatures due to a changing climate and extreme events.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/w14071032/s1, Table S1: Station Metadata, Equations S1–S2: MAE, NSE, and bias equations, Table S2: Individual Region RMSE Metrics, Figure S1: SS Error Metric Boxplots, Figure S2: HPO Skill Scores, Figure S3: Regional Error Metric Boxplots, Figure S4: Single vs. Multiple Region Training RMSE Boxplots, Figure S5: PUB Error Metric Boxplots, Figure S6: Testing Period Boxplots, Figure S7: Station Time Series, Figure S8: PUB Spatial RMSE Metrics.

Author Contributions: Conceptualization, H.W., J.D.W., C.V. and A.R.L.; methodology, C.V., A.R.L. and H.W.; software, H.W., M.L., V.C.H., D.S.C. and C.V.; validation, H.W.; formal analysis, H.W., A.R.L. and C.V.; investigation, H.W., C.V. and A.R.L.; resources, C.V.; data curation, C.V., D.S.C. and V.C.H.; writing—original draft preparation, H.W. and C.V.; writing—review and editing, H.W., J.D.W., C.V. and A.R.L.; visualization, H.W. and J.D.W.; supervision, C.V.; project administration, C.V.; funding acquisition, C.V. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work supported by the Early Career Research Program funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under the Berkeley Lab Contract Number DE-AC02-05CH11231.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Data synthesized using the BASIN-3D integration tool and code used to run the models in this study are published as a dataset in the US Department of Energy's ESS-DIVE data repository [58] under a CC-BY 4.0 license.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Van Vliet, M.; Ludwig, F.; Zwolsman, J.; Weedon, G.; Kabat, P. Global river temperatures and sensitivity to atmospheric warming and changes in river flow. *Water Resour. Res.* **2011**, *47* . [CrossRef]
2. Abbott, B.W.; Bishop, K.; Zarnetske, J.P.; Hannah, D.M.; Frei, R.J.; Minaudo, C.; Chapin, F.S.; Krause, S.; Conner, L.; Ellison, D.; et al. A water cycle for the Anthropocene. *Hydrol. Process.* **2019**, *33*, 3046–3052. [CrossRef]
3. Förster, H.; Lilliestam, J. Modeling thermoelectric power generation in view of climate change. *Reg. Environ. Chang.* **2010**, *10*, 327–338.
4. Van Vliet, M.T.; Wiberg, D.; Leduc, S.; Riahi, K. Power-generation system vulnerability and adaptation to changes in climate and water resources. *Nat. Clim. Chang.* **2016**, *6*, 375–380.
5. Lawrence, D.J.; Stewart-Koster, B.; Olden, J.D.; Ruesch, A.S.; Torgersen, C.E.; Lawler, J.J.; Butcher, D.P.; Crown, J.K. The interactive effects of climate change, riparian management, and a nonnative predator on stream-rearing salmon. *Ecol. Appl.* **2014**, *24*, 895–912. [PubMed]
6. van Vliet, M.T.; Franssen, W.H.; Yearsley, J.R.; Ludwig, F.; Haddeland, I.; Lettenmaier, D.P.; Kabat, P. Global river discharge and water temperature under climate change. *Glob. Environ. Chang.* **2013**, *23*, 450–464.
7. Heck, M.P.; Schultz, L.D.; Hockman-Wert, D.; Dinger, E.C.; Dunham, J.B. *Monitoring Stream Temperatures—Guide for Non-Specialists*; Technical Report; US Geological Survey: Washington, DC, USA, 2018.
8. Benyahya, L.; Caissie, D.; St-Hilaire, A.; Ouarda, T.B.; Bobée, B. A review of statistical water temperature models. *Can. Water Resour. J.* **2007**, *32*, 179–192.
9. Caissie, D. The thermal regime of rivers: A review. *Freshw. Biol.* **2006**, *51*, 1389–1406.
10. Gitau, M.W.; Chen, J.; Ma, Z. Water quality indices as tools for decision making and management. *Water Resour. Manag.* **2016**, *30*, 2591–2610.
11. Liu, L.; Hejazi, M.; Li, H.; Forman, B.; Zhang, X. Vulnerability of US thermoelectric power generation to climate change when incorporating state-level environmental regulations—Nature Energy. *Nat. Energy* **2017**, *2*, 1–5. [CrossRef]
12. Huang, B.; Langpap, C.; Adams, R.M. The value of in-stream water temperature forecasts for fisheries management. *Contemp. Econ. Policy* **2012**, *30*, 247–261.
13. Mijares, V.; Gitau, M.; Johnson, D.R. A method for assessing and predicting water quality status for improved decision-making and management. *Water Resour. Manag.* **2019**, *33*, 509–522.
14. Mantua, N.; Tohver, I.; Hamlet, A. Climate change impacts on streamflow extremes and summertime stream temperature and their possible consequences for freshwater salmon habitat in Washington State. *Clim. Chang.* **2010**, *102*, 187–223.
15. Wilby, R.; Orr, H.; Watts, G.; Battarbee, R.; Berry, P.; Chadd, R.; Dugdale, S.; Dunbar, M.; Elliott, J.; Extence, C.; et al. Evidence needed to manage freshwater ecosystems in a changing climate: turning adaptation principles into practice. *Sci. Total Environ.* **2010**, *408*, 4150–4164. [PubMed]
16. Lovett, G.M.; Burns, D.A.; Driscoll, C.T.; Jenkins, J.C.; Mitchell, M.J.; Rustad, L.; Shanley, J.B.; Likens, G.E.; Haeuber, R. Who needs environmental monitoring? *Front. Ecol. Environ.* **2007**, *5*, 253–260.
17. Neumann, J.L.; Arnal, L.; Emerton, R.E.; Griffith, H.; Hyslop, S.; Theofanidi, S.; Cloke, H.L. Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity. *Geosci. Commun.* **2018**, *1*, 35–57.
18. Rahmani, F.; Shen, C.; Oliver, S.; Lawson, K.; Appling, A. Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. *Hydrol. Process.* **2021**, *35*, e14400.
19. Kędra, M.; Wiejaczka, Ł. Climatic and dam-induced impacts on river water temperature: Assessment and management implications. *Sci. Total Environ.* **2018**, *626*, 1474–1483.
20. Zhang, X.; Li, H.Y.; Leung, L.R.; Liu, L.; Hejazi, M.I.; Forman, B.A.; Yigzaw, W. River Regulation Alleviates the Impacts of Climate Change on U.S. Thermoelectricity Production. *J. Geophys. Res. Atmos.* **2020**, *125*, e2019JD031618. [CrossRef]
21. Kelleher, C.; Wagener, T.; Gooseff, M.; McGlynn, B.; McGuire, K.; Marshall, L. Investigating controls on the thermal sensitivity of Pennsylvania streams. *Hydrol. Process.* **2012**, *26*, 771–785.
22. Borman, M.; Larson, L. A case study of river temperature response to agricultural land use and environmental thermal patterns. *J. Soil Water Conserv.* **2003**, *58*, 8–12.
23. Sanders, M.J.; Markstrom, S.L.; Regan, R.S.; Atkinson, R.D. *Documentation of a Daily Mean Stream Temperature Module—An Enhancement to the Precipitation-Runoff Modeling System*; Technical Report; US Geological Survey: Washington, DC, USA, 2017.
24. Li, H.Y.; Leung, L.R.; Tesfa, T.; Voisin, N.; Hejazi, M.; Liu, L.; Liu, Y.; Rice, J.; Wu, H.; Yang, X. Modeling stream temperature in the Anthropocene: An earth system modeling approach. *J. Adv. Model. Earth Syst.* **2015**, *7*, 1661–1679.
25. Van Vliet, M.; Yearsley, J.; Franssen, W.; Ludwig, F.; Haddeland, I.; Lettenmaier, D.; Kabat, P. Coupled daily streamflow and water temperature modelling in large river basins. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 4303–4321.
26. Zhu, S.; Piotrowski, A.P. River/stream water temperature forecasting using artificial intelligence models: A systematic review. *Acta Geophys.* **2020**, *68*, 1433–1442. [CrossRef]
27. Wehrly, K.E.; Brenden, T.O.; Wang, L. A comparison of statistical approaches for predicting stream temperatures across heterogeneous landscapes 1. *JAWRA J. Am. Water Resour. Assoc.* **2009**, *45*, 986–997.
28. Chang, H.; Psaris, M. Local landscape predictors of maximum stream temperature and thermal sensitivity in the Columbia River Basin, USA. *Sci. Total Environ.* **2013**, *461*, 587–600.

29. Daigle, A.; St-Hilaire, A.; Peters, D.; Baird, D. Multivariate modelling of water temperature in the Okanagan watershed. *Can. Water Resour. J.* **2010**, *35*, 237–258.
30. Sohrabi, M.M.; Benjankar, R.; Tonina, D.; Wenger, S.J.; Isaak, D.J. Estimation of daily stream water temperatures with a Bayesian regression approach. *Hydrol. Process.* **2017**, *31*, 1719–1733.
31. Toffolon, M.; Piccolroaz, S. A hybrid model for river water temperature as a function of air temperature and discharge. *Environ. Res. Lett.* **2015**, *10*, 114011.
32. Gallice, A.; Schaefli, B.; Lehning, M.; Parlange, M.B.; 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.* **2015**, *19*, 3727–3753. [CrossRef]
33. Hill, R.A.; Hawkins, C.P.; Carlisle, D.M. Predicting thermal reference conditions for USA streams and rivers. *Freshw. Sci.* **2013**, *32*, 39–55.
34. Isaak, D.J.; Luce, C.H.; Rieman, B.E.; Nagel, D.E.; Peterson, E.E.; Horan, D.L.; Parkes, S.; Chandler, G.L. Effects of climate change and wildfire on stream temperatures and salmonid thermal habitat in a mountain river network. *Ecol. Appl.* **2010**, *20*, 1350–1371. [PubMed]
35. Arismendi, I.; Safeeq, M.; Dunham, J.B.; Johnson, S.L. Can air temperature be used to project influences of climate change on stream temperature? *Environ. Res. Lett.* **2014**, *9*, 084015.
36. Hrachowitz, M.; Soulsby, C.; Imholt, C.; Malcolm, I.; Tetzlaff, D. Thermal regimes in a large upland salmon river: A simple model to identify the influence of landscape controls and climate change on maximum temperatures. *Hydrol. Process.* **2010**, *24*, 3374–3391.
37. 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.; et al. The NorWeST summer stream temperature model and scenarios for the western US: A crowd-sourced database and new geospatial tools foster a user community and predict broad climate warming of rivers and streams. *Water Resour. Res.* **2017**, *53*, 9181–9205.
38. Piotrowski, A.P.; Napiorkowski, J.J. Simple modifications of the nonlinear regression stream temperature model for daily data. *J. Hydrol.* **2019**, *572*, 308–328.
39. Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. [CrossRef]
40. Rahmani, F.; Lawson, K.; Ouyang, W.; Appling, A.; Oliver, S.; Shen, C. Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environ. Res. Lett.* **2021**, *16*, 024025.
41. Zhi, W.; Feng, D.; Tsai, W.P.; Sterle, G.; Harpold, A.; Shen, C.; Li, L. From hydrometeorology to river water quality: Can a deep learning model predict dissolved oxygen at the continental scale? *Environ. Sci. Technol.* **2021**, *55*, 2357–2368.
42. Feigl, M.; Lebiedzinski, K.; Herrnegger, M.; Schulz, K. Machine-learning methods for stream water temperature prediction. *Hydrol. Earth Syst. Sci.* **2021**, *25*, 2951–2977.
43. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **1997**, *9*, 155–161.
44. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
45. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **2000**, *28*, 337–407.
46. Rajesh, M.; Rehana, S. Prediction of river water temperature using machine learning algorithms: A tropical river system of India. *J. Hydroinform.* **2021**, *23*, 605–626.
47. Lu, H.; Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **2020**, *249*, 126169. [PubMed]
48. Turschwell, M.P.; Peterson, E.E.; Balcombe, S.R.; Sheldon, F. To aggregate or not? Capturing the spatio-temporal complexity of the thermal regime. *Ecol. Indic.* **2016**, *67*, 39–48.
49. Rehana, S. River water temperature modelling under climate change using support vector regression. In *Hydrology in a Changing World*; Springer: Cham, Switzerland 2019; pp. 171–183.
50. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
51. Jia, X.; Lin, B.; Zwart, J.; Sadler, J.; Appling, A.; Oliver, S.; Read, J. Graph-based Reinforcement Learning for Active Learning in Real Time: An Application in Modeling River Networks. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM) SIAM, Virtual Event, 29 April–1 May 2021; pp. 621–629.
52. Kratzert, F.; Klotz, D.; Herrnegger, M.; Sampson, A.K.; Hochreiter, S.; Nearing, G.S. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.* **2019**, *55*, 11344–11354.
53. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
54. USGS National Water Information System. Available online: https://waterdata.usgs.gov/nwis (accessed on 16 July 2022).
55. Thornton, P.E.; Shrestha, R.; Thornton, M.; Kao, S.C.; Wei, Y.; Wilson, B.E. Gridded daily weather data for North America with comprehensive uncertainty quantification. *Sci. Data* **2021**, *8*, 1–17.
56. Varadharajan, C.; Hendrix, V.C.; Christianson, D.S.; Burrus, M.; Wong, C.; Hubbard, S.S.; Agarwal, D.A. BASIN-3D: A brokering framework to integrate diverse environmental data. *Comput. Geosci.* **2022** , *159*, 105024.

57. Daymet Pixel Extraction Tool. Available online: https://daymet.ornl.gov/single-pixel/api (accessed on 16 July 2022).
58. Weierbach, H.; Lima, A.; Willard, J.; Hendrix, V.; Christianson, D.; Lubich, M.; Varadharajan, C. Dataset for "Stream Temperature Predictions for River Basin Management in the Pacific Northwest and Mid-Atlantic Regions Using Machine Learning". *ESS-DIVE Repos.* **2022**. [CrossRef]
59. Falcone, J.A. *GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow*; Technical Report; US Geological Survey: Washington, DC, USA, 2011.
60. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
61. Bergstra, J.; Yamins, D.; Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Proceedings of the International Conference on Machine Learning. PMLR, Atlanta, GA, USA, 16–21 June 2013; pp. 115–123.
62. Pandas Development Team, T. Pandas-dev/Pandas: Pandas. 2020. [CrossRef]
63. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 56–61. [CrossRef]
64. USGS Site Information Figure. Available online: https://help.waterdata.usgs.gov/tutorials/site-information/what-is-my-watershed-address-and-how-will-it-help-me-find-usgs-data (accessed on 7 March 2022).
65. Nielsen, D. Tree Boosting with Xgboost—Why Does Xgboost Win "Every" Machine Learning Competition? Master's Thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2016.
66. Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G.; Jun, H.; Kianinejad, H.; Patwary, M.; Ali, M.; Yang, Y.; Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv* **2017**, arXiv:1712.00409.
67. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 843–852.
68. Olive, D.J. Multiple linear regression. In *Linear Regression*; Springer: Cham, Switzerland, 2017; pp. 17–83.
69. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
70. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.
71. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.
72. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 2546–2554.
73. Bergstra, J.; Yamins, D.; Cox, D.D. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In Proceedings of the 12th Python in Science Conference, Austin, TX, USA, 24–29 June 2013; Volume 13, p. 20.
74. Jackson, E.K.; Roberts, W.; Nelsen, B.; Williams, G.P.; Nelson, E.J.; Ames, D.P. Introductory overview: Error metrics for hydrologic modelling—A review of common practices and an open source library to facilitate use and adoption. *Environ. Model. Softw.* **2019**, *119*, 32–48. [CrossRef]
75. Hsieh, W.W. *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*; Cambridge University Press: Cambridge, UK, 2009.
76. Zounemat-Kermani, M.; Batelaan, O.; Fadaee, M.; Hinkelmann, R. Ensemble machine learning paradigms in hydrology: A review. *J. Hydrol.* **2021**, *598*, 126266.
77. Dion, P.; Martel, J.L.; Arsenault, R. Hydrological ensemble forecasting using a multi-model framework. *J. Hydrol.* **2021**, *600*, 126537.
78. Jiang, S.; Ren, L.; Yang, X.; Ma, M.; Liu, Y. Multi-model ensemble hydrologic prediction and uncertainties analysis. *Proc. Int. Assoc. Hydrol. Sci.* **2014**, *364*, 249–254.
79. Fleming, S.W.; Bourdin, D.R.; Campbell, D.; Stull, R.B.; Gardner, T. Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river. *JAWRA J. Am. Water Resour. Assoc.* **2015**, *51*, 502–512.
80. Fleming, S.W.; Garen, D.C.; Goodbody, A.G.; McCarthy, C.S.; Landers, L.C. Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence. *J. Hydrol.* **2021**, *602*, 126782.
81. Kratzert, F.; Klotz, D.; Shalev, G.; Klambauer, G.; Hochreiter, S.; Nearing, G. Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine-Learning Applied to Large-Sample Datasets. *arXiv* **2019**, arXiv:1907.08456.
82. DeWeber, J.T.; Wagner, T. A regional neural network ensemble for predicting mean daily river water temperature. *J. Hydrol.* **2014**, *517*, 187–200.
83. Zhu, S.; Nyarko, E.K.; Hadzima-Nyarko, M. Modelling daily water temperature from air temperature for the Missouri River. *PeerJ* **2018**, *6*, e4894.