Supplemental Materials for:

Machine Learning in Stream/River Water Temperature Modelling: a review and metrics for evaluation

Claudia R. Corona[1], Terri S. Hogue[1,2]

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

Correspondence to: Claudia R. Corona (Claudia.corona@mines.edu)

For this section, we organize the studies discussed and relevant information in Tables S1 and S2. Table S1 includes the 54 publications from which performance metrics where obtained. Table S1 includes the following information (as provided by the study): period considered, region examined, temporal resolution of SWT, spatial scale of study, hydrometeorological parameters used for modeling, and evaluation metrics used to assess model performance. References are in chronological order by year, by year of publication acceptance. For studies that detailed more than 5-6 hydrometeorological parameters/forcings, the number of variables is stated on Table S1 but only a partial list is included. With regards to the spatial scale stated in Table S1, the first subsection discusses studies that focused on the local/watershed scale, and the second subsection focuses on regional/CONUS point-scale studies. Table S2 lists the data analysis techniques and/or ML algorithms used, as well as the training/validation/testing percentages/time periods as reported by the study.

Supplemental Materials, Table 1. Spatial/temporal data and parameters used by SWT studies, as reported. SWT = stream water temperature, ET = Evapotranspiration, PPT = Precipitation, AT = air temperature, DO = dissolved oxygen, DOY = day-of-year.

#	Reference	Time period Considered	Region	SWT Temporal Resolution	Spatial Scale	Hydrometeorological Parameters (partial list)	Evaluation Metrics Used
1	**Foreman et al., 2001**	1953–1999, May to October	Fraser, Thompson River, Canada	Hourly	Local/ Watershed	SWT	Slope, R^2, RMSE (called RMSD), standard error
2	**Risley et al., 2003**	1999, June to September	western Oregon (U.S)	Hourly	Regional	SWT, riparian shade, elevation, % forested	R^2, RMSE
3	**Sivri et al., 2007**	2001–2002	Province of Rize, NE Turkey	Daily	Local / Watershed	Daily SWT, monthly mean AT, PPT, water quality parameters (DO, ph, etc.)	R^2, Sum of Squares Error (SSE)
4	**Tao et al., 2008**	1986–2003	Yellow River, China	Daily	Local / Watershed	AT, water level, channel storage, discharge	R^2
5	**Chenard and Caissie, 2008**	10 years: 1992–2002	Miramichi River, Canada	Daily	Local / Watershed	Hourly water level, Daily AT, Hourly SWT	RMSE, PBIAS, R^2
6	**Sahoo et al., 2009**	1997–2003 (varied by station)	Lake Tahoe watershed	Daily	Regional	AT, short-wave radiation	Mean error (E), r, RMSE. MSE for training error.
7	**St-Hilaire et al., 2012**	5 years: 1993–1998,	Moisie River, Quebec, Canada	Daily	Local / Watershed	AT, stream discharge	RMSE, relative bias (RB), NSE (called NTD)
8	**Hong and Bhamidimarri, 2012**	Dec 2001– April 2002	Waikato River, New Zealand	10-min interval	Local / Watershed	AT, barometric pressure, wind direction/speed, solar radiation, SWT, humidity, water level	RMSE, MSE
9	**Grbic et al., 2013**	5 years: 1993–1998	Drava river, Croatia	Daily	Local / Watershed	SWT , Daily mean AT, Daily mean discharge	NSE (called NSC), RMSE, MAE
10	**Hebert et al., 2014**	9 years: 1998–2007, April– October	Miramichi River, Canada	Hourly	Local / Watershed	AT, SWT, daily mean discharge	RMSE, R^2, PBIAS
11	**DeWeber and Wagner, 2014**	19 years, 1980–2010: May– October	Eastern US	Daily	Regional	AT, landform attributes, and forested land cover	RMSE, NSE, RMSE/SD, PBIAS
12	**Cole et al., 2014**	99 years: 1913–2012 (varied by station)	Delaware river	Daily	Local / Watershed	AT, wind speed, solar radiation, barometric pressure, dew point temp., relative humidity, sky cover, daily PPT	RMSE, NSE, PBIAS, Willmott index of agreement (d)
13	**Hadzima-Nyarko et al. 2014**	19 years: 1991–2010	Drava river, Croatia.	Daily	Local / Watershed	Daily mean AT, SWT	RMSE, normalized RMSE %, r, NSE, NSE-A: adjusted coefficient of efficiency
14	**Piotrowski et al., 2015**	17 years: 11/01/1990– 10/31/2000	Biana, Spurasl, Tarnowska Rivers, Poland	Daily	Local / Watershed	AT, sun declination	MSE

15	**Rabi et al., 2015**	19 years: 1991–2010	Drava river, Croatia.	Daily	Local / Watershed	Daily mean AT data, Daily SWT data	RMSE, RMSE %, r, NSE, NSE-A
16	**Jeong et al., 2016**	18 years: 1991–2009	Soyang river, South Korea	Hourly	Local / Watershed	Hourly air temp., hourly SWT, hourly rainfall, wind speed, wind direction	RMSE, NSE (called NASH), r, d
17	**Kwak et al., 2016**	3 years: 2011–2014, Summer	Fourchue River, Quebec, Canada	Hourly data aggregated to Daily	Local / Watershed	SWT, AT, solar radiation, relative humidity, and wind speed	RMSE, PBIAS, NSE
18	**Laanaya et al., 2017**	10 years: 2005–2015, (varied by station)	Sainte-Marguerite River, Canada	Hourly data aggregated to Daily	Local / Watershed	SWT, AT and stream discharge	RMSE, PBIAS, NSE (called NSC)
19	**Liu et al., 2018**	7 years: 2007 2014	Eel River, Coastal CA	Daily	Regional	AT, SWT	NSE, R^2, RMSE (for ANN model)
20	**Voza and Vukovic, 2018**	10 years: 2005–2015	Morava River, Serbia	Monthly	Regional	SWT, water quality parameters (DO, pH, etc.)	R^2, Spearman's *rho*
21	**Temizyurek and Dadaser-Celik, 2018**	12 years: 1995–2007	Kizilirmak river, Turkey	Monthly	Local / Watershed	AT, wind speed, relative humidity, SWTs	R^2, RMSE
22	**Graf et al., 2019**	39 years: 1984–2013	Warta River, Poland	Daily	Regional	AT, SWT	R^2, RMSE, MAE
23	**Zhu, Heddam, Nyarko et al., 2019**	32 years: 1984–2016, (varies by station)	Drava River, Croatia; Mentue, Rhone, Dischmaback river, Switzerland	Daily	Local / Watershed	AT, discharge, and components of the Gregorian calendar	r d, RMSE, MAE
24	**Zhu, Heddam, Wu et al., 2019**	16 years: 2001–2017 (varies by station)	Cedar River (WA), Fanno Creek (OR), Irondequoit Creek (NY)	Daily	Local / Watershed	AT, discharge, and DOY	r, d, RMSE, MAE
25	**Zhu, Hadzima-Nyarko, Gao et al., 2019**	26 years: 1991–2017 (varies by station)	Drava river, Croatia.	Daily	Local / Watershed	Daily SWT, AT	r, d, RMSE, MAE
26	**Zhu, Nyarko, Hadzima-Nyarko et al., 2019**	26 years: 1991–2017 (varies by station)	Drava river, Croatia; 3 rivers in Switzerland; 3 rivers in U.S.	Daily	Local / Watershed	AT, discharge, DOY	r, d, RMSE, MAE
27	**Zhu and Heddam, 2019**	16 years: 2001–2017 (varies by station)	Cedar (WA), Fanno (OR), Irondequoit Creek (NY)	Daily	Local / Watershed	AT, discharge, DOY	r, d, RMSE, MAE
28	**Lu and Ma, 2020**	May–July 2019	Tualatin River, Oregon, U.S.	Hourly	Local / Watershed	AT, SWT, water quality parameters (DO, ph, etc.)	MAE, RMSE, MAPE, RMSPE, U1, U2
29	**Qiu et al., 2020**	28 years: 1977–1987, 1993–2011	Yangtze River, China	Daily	Regional	AT, discharge, DOY	MAE, RMSE, NSE, R^2

30	**Krishnaraj & Deka, 2020**	12 years: 2005–2017	Ganga River Basin, India	Monthly	Local / Watershed	20 parameters see Table 1, includes: SWT	Spearman's *rho*
31	**Rahmani et al., 2021**	4 years: 2010–2014	118 basins in CONUS w/o dams	Daily	Point-scale across CONUS	SWT, discharge	RMSE, median RMSE, PBIAS, NSE
32	**Feigl et al., 2021**	1977–2015 (varied by station)	Ten Austrian catchments	Daily	Regional	Daily mean/max/min AT, runoff, PPT, and global radiation.	MSE, MAE, RMSE
33	**Piotrowski et al., 2021**	46 years: 1971–2017	3 from NW U.S., 2 from Poland	Daily	Local / Watershed	AT and discharge	MSE, NSE (called NSC), AIC, BIC
34	**Abdi et al., 2021**	June 10–July 18, 2016	Los Angeles river, CA, U.S.	Hourly	Local / Watershed	AT, relative humidity, wind speed, barometric pressure	MAE, RMSE, R^2
35	**Graf and Aghelpour, 2021**	20 years: 11/1989–11/2009	Warta River, Poland	Daily	Regional	Time lag in SWT	R^2, RMSE, Normalized RMSE, MAE, NSE (called NS).
36	**Rajesh and Rehana, 2021**	1989-2014	Tunga-Bhadra river, Shimoga Station, India	Daily, also monthly, seasonal	Local/Watershed	AT (avg./max/min), SWT	R^2, MSE, RMSE, RSR, NSE, MAE, KGE
37	**Heddam, Kim et al., 2022**	53 years: 1961–2014	Orda river, several stations, Poland	Daily	Regional	AT, DOY	RMSE, MAE, NSE, *r*
38	**Jiang et al., 2022**	3 years: 2015–2018	Jinsha River Basin	Daily	Regional	Daily avg./max/min AT, dew temp., discharge, DOY, wind speed, and PPT.	RMSE, R^2, NSE
39	**Sadler et al., 2022**	34 years: 1980–2014	CONUS	Daily	Point-scale across CONUS	Daylight, PPT, short-wave radiation, snow-water equivalent, AT, vapor pressure	NSE, t-scores of the NSEs
40	**Weierbach et al., 2022**	40 years: 1980–2020	Mid-Atlantic, Pacific NW catchments	Monthly	Regional	AT, solar radiation, discharge, month of year, drainage area and dam info	RMSE, MAE, NSE, PBIAS
41	**Zanoni et al., 2022**	20 years: 1994–2014	Adige river, NE Italy	Daily	Regional	SWT, and other water quality parameters (As, Cl, DO, etc.)	PBIAS, NSE, RMSE, KGE
42	**Heddam, Ptak et al., 2022**	27 years: 1987 - 2014	Kaczawa, Orla, Nida, Krzna, Wieprz stations, Poland	Daily	Regional	SWT, mean AT, DOY	RMSE, MAE, NSE, R^2
43	**Hani et al., 2023**	2020, 2021: Summer months	Sainte-Marguerite River, Canada	Hourly	Local	Discharge, AT, SWT,	Mean, Median, SD, RMSE, r-RMSE, NSE, PBIAS, r-Bias
44	**Khosravi et al., 2023**	21 years: 2006–2022	Delaware River, New Jersey, U.S.	Varied due to different duration	Local / Watershed	SWT, discharge, water level, DO, turbidity, pH, specific conductance	RMSE, R^2, NSE (for loss function)
45	**Topp et al., 2023**	42 years: 1980–2022	Delaware River Basin (DRB), U.S.	Daily	Regional	AT, PPT, shortwave radiation, potential ET, stream width, slope, elevation	RMSE, NSE, bias

46	**Rehana and Rajesh, 2023**	37 years: 1980–2017 (varies by station)	India: 7 catchments	Monthly	Regional	Monthly min. and max AT, SWT	NSE, KGE, RSR, RMSE, MAE.
47	**Siegel et al., 2023**	33 years: 1990 - 2023	Pacific Northwest	Daily	Regional	SWT, Daily mean AT, daily modeled snow-pack data SWE, daily discharge	RMSE, MAE
48	**Wade et al., 2023**	4 years: 2016–2020	CONUS	Monthly	Point-scale across CONUS	23 variables, includes: AT, monthly PPT, discharge	RMSE normalized, R^2
49	**Souaissi et al., 2023**	15 years, 1960s–2023: May–October	24 river temp. stations, Switzerland	Monthly	Regional	14 variables include: catchment area, mean elevation, stream length, max annual AT, summer PPT	R^2, RMSE, RRMSE, BIAS, Pearson correlation coefficients: area, mean, etc.
50	**Drainas et al., 2023**	1–41 years, 1980–2021 (varied by station)	16 headwater streams, Bavaria, Germany	Daily	Regional	AT, DOY, discharge, water level and sunshine per day	RMSE, r, PBIAS
51	**Rozos, 2023**	1992–2014 (varied by station)	Arno River, Sieve River (Italy), Bakas River (Greece)	Daily, Hourly	Local / Watershed	Mean Areal Daily rainfall, ET, discharge, average annual rainfall/PPT	MSE, PBIAS
52	**Rahmani et al., 2023**	6 years: 2010–2016	CONUS, (GAGES-II)	Daily	CONUS	33 attributes including AT, SWT, DOY, etc.	PBIAS, NSE, KGE, RMSE (loss function, error minimization)
53	**He et al., 2024**	40 years: 1980–2020	Delaware River (NJ), Houston River (Texas)	Daily	Regional	Discharge, SWT, AT	RMSE, NSE
54	**Majerska et al., 2024**	10 years: 2012–2022, June to September	Fuglebekken stream, Svalbard, Norway	Daily	Local/ Watershed	SWT, water level, ground temperature, radiation	KGE

35

40

45

Supplemental Materials, Table S2. Data analysis techniques and/or ML algorithms used by studies. Reported training/calibration/validation included directly from studies.

#	Reference	Data Analysis Technique OR ML Algorithms Used	Training (Zhu et al. 2019 studies label it as "calibration"), validation, testing
1	**Foreman et al., 2001**	Linear regression and neural networks (NNs)	4 years, 1995–1998, training, 21 years, 1953–1994 validation
2	**Risley et al., 2003**	Feed-forward NN, three sets of NN models	33 % training, 66 % testing. Six of 148 streams randomly removed, used for validation.
3	**Sivri et al., 2007**	Multi-Layer Perceptron (MLPNN)	2001: training, 2002: testing
4	**Tao et al., 2008**	Back-propagation Neural network (BPNN)	14 years, 1986–1999, training, 4 years, 2000–2003, testing
5	**Chenard and Caissie, 2008**	MLPNNs models w/ variety of input parameters: 4 predict mean daily SWT and 4 predicted max daily SWT	Training: 5 years, Testing: 2 years, Validation: 4 years
6	**Sahoo et al., 2009**	NN, statistical model (multiple regression analysis (MRA), and chaotic non-linear dynamic algorithms (CNDA).	Training, validation: 3 years, 1999–2001 Testing: 1 year, 2002
7	**St-Hilaire et al., 2012**	K-nearest neighbor (K-nn)	"The relatively small sample size precluded us from performing a split sample validation".
8	**Hong and Bhamidimarri, 2012**	Two DNFLMS modes: (1) online one-pass clustering, extended Kalman filtering algorithm (mode 1); and (2) extended Kalman filtering with BP algorithm trained to MSE.	Training 2.5 months, testing 1.5 months
9	**Grbic et al., 2013**	Combination of Gaussian process regression (GPR) models	Training: 4 years, 1993–1996 Testing: 2 years, 1997–1998
10	**Hebert et al., 2014**	NN	Training: 4 years, 1998–2002 Validation: 5 years, 2003–2007
11	**DeWeber and Wagner, 2014**	Four variations of FFNN models	90 % training, 10 % validation, 2010 used as testing year
12	**Cole et al., 2014**	Generalized Least Squares w/ cosine (GLScos), AutoRegressive Integrated Moving Average (ARIMA), NN	Training: 4 years, 2008–2011, tested: 1 year, 2012
13	**Hadzima-Nyarko et al. 2014**	MLPNN and a radial basis function network (RBFNN)	80 % training, 20 % testing
14	**Piotrowski et al., 2015**	MLPNN, product-unit NN, wavelet NN, adaptive-network-based fuzzy inference systems (ANFIS) and K-nn	Two sites training/validation/testing: 41/24/35, 40/30/30
15	**Rabi et al., 2015**	Six different MLPNNs	80 % training, 20 % testing
16	**Jeong et al., 2016**	FFNN	Not Available
17	**Kwak et al., 2016**	(a) Deterministic model CEQUEAU; (b) ARMAX (AutoRegressive-Moving Average with eXogenous terms) stochastic model; and (c) NARX (Nonlinear AutoRegressive model w/ eXogenous input).	75 % of data for training/calibration and 25 % of data for validation
18	**Laanaya et al., 2017**	Generalized Additive Model (GAM), Logistic Model (LM), Residuals regression model (RRM), and Linear regression model (LRM)	Not explicitly stated
19	**Liu et al., 2018**	Integrated model of NN and VIC (not ML)	1.5–5 years (calibration), 0.75–2 years (validation)

20	**Voza and Vukovic, 2018**	MLPNN	Not explicitly stated
21	**Temizyurek and Dadaser-Celik, 2018**	NN	50 % training, 25 % testing, and 25 % used as holdout
22	**Graf et al., 2019**	Wavelet Transform (WT) and NN hybrid model, using 4 mother wavelets: Daubechies, Symlet, discrete Meyer, and Haar. Tested against MLPNN.	Stations used for: Training (4 of 9), testing (3 of 9), validation (2 of 9)
23	**Zhu, Heddam, Nyarko et al., 2019**	4 ML models: (MLPNN) and (ANFIS) – w/ fuzzy c-mean clustering (ANFIS_FC), ANFIS w/ grid partition (ANFIS_GP), and ANFIS w/ subtractive clustering method (ANFIS_SC),	Varied by station: Training data: 6–20 years, ~67–73 %, Validation data: 3–10 years, 27–33 %
24	**Zhu, Heddam, Wu et al., 2019**	Extreme learning machine (ELM), MLPNN and MLR models.	Training data: 8–11 years, ~70–75 %, Validation data: 3–4 years, ~20–25 %
25	**Zhu, Hadzima-Nyarko, Gao et al., 2019**	Hybrid models combining WT and NN (WTNN) and WT with ANFIS (WTANFIS), NN, ANFIS	Calibration/training: Botovo, 1991–2008, ~17 years. Donji 1993–2008, 15 years. Validation: 2009–2016, 7 years for Botovo, Donji
26	**Zhu, Nyarko, Hadzima-Nyarko et al., 2019**	(FFNN), (GPR), and (DT) models	Varied by station: training data: 6–20 years, ~67–73 %, validation: 3–10 years, 27–33 %
27	**Zhu and Heddam, 2019**	Optimally pruned extreme learning machine (OPELM) and RBFNN	Training data: 8–11 years, ~70–75 %, Validation data: 3–4 years, ~20–25 %
28	**Lu and Ma, 2020**	Extreme gradient boosting (XGBoost) and random forest (RF)	Training/testing: 9:1 ratio
29	**Qiu et al., 2020**	BP_PSO model variations, based on BPNN (back propagation NN) optimized by PSO (particle swarm optimization) algorithm.	Cuntan: 13/17 years = 76.5 % (training), 4/17 years = 24.5 % (testing), Datong: 7/9 years = 78 % (training), 2/9 years= 22 % testing
30	**Krishnaraj and Deka, 2020**	k-means cluster analysis (CA), principal component analysis (PCA) and correlation	Not explicitly stated
31	**Rahmani et al., 2021**	LSTM	4 years, 67 %, training 2 years, 33 % testing
32	**Feigl et al., 2021**	Step-wise linear regression, RF, XGBoost, FFNNs, 2 (RNNs): LSTM and recurrent gated unit (GRU)	60 % training, 20 % validation, 20 % testing
33	**Piotrowski et al., 2021**	MLPNNs, PUNN, extended logistic regression and air2stream (not ML)	66 % training, 33 % validation
34	**Abdi et al., 2021**	Four models: single-layer (SLR), multilayer linear regression (MLR) and SLR and MLR-NN	80% training, validation, and 20 % testing
35	**Graf and Aghelpour, 2021**	Autoregressive (AR), Moving Average (MA), Autoregressive Moving Average (ARMA). Autoregressive Integrated Moving Average (ARIMA), ANFIS, Radial Basis Function (RBF) and Group Method of Data Handling (GMDH).	75 % training, 25 % testing
36	**Rajesh and Rehana, 2021**	ML models - Ridge regression (RR), K-nn, Random forest, and SVR	Not explicitly stated, cross-validation mentioned.
37	**Heddam, Ptak et al., 2022**	K-nn, least square support vector machine (LSSVM), generalized regression (GRNN), cascade correlation artificial (CCNN), relevance vector machine (RVM), and locally weighted polynomials regression (LWPR).	70 % training, 30 % testing

38	**Jiang et al., 2022**	Decision trees (DT), RF, gradient boosting (GB), adaptive boosting (AB), SVR, MLPNN	60 % training, 40% testing
39	**Sadler et al., 2022**	LSTM	Years splits (Training/validation/testing): Exp. A, (25/4/4), Exp. B, (25/4/4), Exp. C, (2/4/4), Exp. D performance testing (31, 1, 1)
40	**Weierbach et al., 2022**	SVR and XGBoost	70 % training, 30 % testing
41	**Zanoni et al., 2022**	RF and FFNN, linear regression model	70 % training, 30 % testing
42	**Heddam, Kim et al., 2022**	Bat algorithm optimized extreme learning machines ELM (Bat-ELM), (MLPNN), (CART) and (MLR).	70 % training, 30 % validation
43	**Hani et al., 2023**	Multivariate adaptive splines regression (MARS), generalized additive model (GAM), support vector machine (SVM), and (RF)	Not explicitly stated
44	**Khosravi et al., 2023**	LSTM	Roughly 70 % training, 30 % validation, not explicitly stated
45	**Topp et al., 2023**	Process-guided deep learning (PGDL): Recurrent graph convolution network (RGCN), temporal convolution graph model (Graph WaveNet)	Two scenarios: Temp. shift 10/5/25 Drought 25/5/10
46	**Rehana and Rajesh, 2023**	LSTM, integrated w/ k-NN bootstrap resampling (k-NN + LSTM), WT-LSTM. Air2stream (not ML)	~77–89 % training, ~11–23 % testing (in years): 18/4; 10/3; 24/3; 8/2; 11/2; 24/3; 11/3.
47	**Siegel et al., 2023**	Generalized Additive Model (GAM)	leave-one-year-out cross validation; leave-one-region-out cross-validation
48	**Wade, Kelleher & Hannah, 2023**	RF	80 % training, 20 % testing
49	**Souaissi et al., 2023**	RF, XGBoost, w/ non-parametric MARS and GAM.	Not explicitly stated
50	**Drainas et al., 2023**	NN	90 % training, 10 % testing
51	**Rozos, 2023**	RNN w/ LSTM, K-nn w/ Bluecat	70 % training, 30 % testing
52	**Rahmani et al., 2023**	LSTM, SN-TEMP (process-based model)	67 % training (4 years), 33 % testing (2 years)
53	**He et al., 2024**	LSTM+GNN	67 % training, 33 % validation
54	**Majerska et al., 2024**	GPR	Five-fold cross-validation; unclear % training vs % testing/validation