Check for updates

LIMNOLOGY AND OCEANOGRAPHY
Letters

ASLO
Open Access

Limnology and Oceanography Letters 2022
© 2022 The Authors. *Limnology and Oceanography Letters* published by Wiley Periodicals LLC on behalf of Association for the Sciences of Limnology and Oceanography.
doi: 10.1002/lol2.10249

DATA ARTICLE

Daily surface temperatures for 185,549 lakes in the conterminous United States estimated using deep learning (1980–2020)

Jared D. Willard [1,2]* *Jordan S. Read* [2] *Simon Topp* [2] *Gretchen J. A. Hansen* [3] *Vipin Kumar* [1]

[1]University of Minnesota Department of Computer Science, Minneapolis, Minnesota; [2]U.S. Geological Survey, Reston, Virginia; [3]University of Minnesota Department of Fisheries, Wildlife, and Conservation Biology, St. Paul, Minnesota

Scientific Significance Statement

Measured or estimated water temperatures are necessary to understand basic aquatic functions and to assess habitat suitability for numerous species. However, the vast majority of lakes in the United States do not have observed temperatures on most days. A nationally consistent dataset of long-term daily retrospective surface water temperatures is needed to support broad-scale limnological synthesis and to quantify local to regional ecosystem change. Deep learning can provide more accurate temperature estimates at greater scale than existing process-based and statistical models; this dataset provides deep learning-estimated daily surface temperature from 1980 to 2020 and spans over 185,549 lakes in the conterminous United States.

Abstract

The dataset described here includes estimates of historical (1980–2020) daily surface water temperature, lake metadata, and daily weather conditions for lakes bigger than 4 ha in the conterminous United States (n = 185,549), and also in situ temperature observations for a subset of lakes (n = 12,227). Estimates were generated using a long short-term memory deep learning model and compared to existing process-based and linear regression models. Model training was optimized for prediction on unmonitored lakes through cross-validation that held out lakes to assess generalizability and estimate error. On the held-out lakes with in situ observations, median lake-specific error was 1.24°C, and the overall root mean squared error was 1.61°C. This dataset increases the number of lakes with daily temperature predictions when compared to existing datasets, as well as

*Correspondence: willa099@umn.edu

Associate editor: Kendra Spence Spence Cheruvelil

Author Contribution Statement: JDW, GJAH, and JR conceived the presented idea. JDW, JR, and ST wrote the manuscript with support from GJAH and VK. JDW preprocessed all modeling data, developed the entity-aware long short-term memory, and implemented all data-driven models. JR compiled all data for the data release (meteorology, lake metadata, observed temperatures, and ERA5 data) and created figures 2, 4, and 5. JDW and JR performed the technical validation. JR performed quality assurance and quality control of observed temperatures. All authors discussed the results and contributed to the final manuscript.

Data Availability Statement: The data are available in the "Daily surface temperature predictions for 185,549 U.S. lakes with associated observations and meteorological conditions (1980-2020)" repository at https://doi.org/10.5066/P9CEMS0M (Willard et al. 2022) and URL of the Metadata with permanent identifier at https://www.sciencebase.gov/catalog/item/60341c3ed34eb12031172aa6.

Temporal range: 01 January 1980 to 31 December 2020.

Frequency or sampling interval: Daily estimated surface water temperature; daily or less frequently measured surface water temperature.

Spatial scale: Conterminous U.S.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

substantially improves predictive accuracy compared to a prior empirical model and a debiased process-based approach (2.01°C and 1.79°C median error, respectively).

Measured or estimated water temperatures are necessary to understand basic aquatic functions (such as microbial decomposition rates and gas exchange; Raymond et al. 2013) and to assess habitat suitability for numerous species (Fang et al. 2004). Diversity in lake temperatures results from unique combinations of weather, climate, and lake-specific properties that modulate responses to meteorological inputs (Livingstone 2008; Rose et al. 2016). Observing lake water temperatures at a temporal resolution sufficient to resolve short-term dynamics (such as temperature drops resulting from cold fronts) and of temporal duration sufficient to measure long-term trends is challenging and often prohibitively expensive, especially when attempting to capture diverse thermal regimes across many lakes. Despite these challenges, water temperature is the most common variable in the United States' Water Quality Portal (WQP; Read et al. 2017), and numerous satellite data products include a measure of surface water temperature (Schaeffer et al. 2018; Vanhellemont 2020). This seemingly high abundance of temperature measurements has been aided by the low cost and simplicity of thermistor sensors for in situ measurements as well as advances in atmospheric correction and emissivity algorithms in remote sensing. However, of the over 270,000 U.S. lakes in the National Hydrography Dataset PlusV2, fewer than 5% have in situ temperature observations and only 62% are resolvable by satellite (Schaeffer et al. 2018). These numbers are significantly lower when accounting for the millions of smaller waterbodies in the United States not included in NHDPlusV2, ultimately meaning that temperature of the vast majority of U.S. lakes is unobserved on most days.

New environmental modeling methods that are equipped to leverage existing data are improving prediction accuracy and being used to create useful data products. Machine learning algorithms are increasingly viable prediction methods for water resources applications due to surging availability of observational data and computational power (Sun and Scanlon 2019). In particular, deep learning algorithms composed of large, multilayer artificial neural networks (ANNs) can extract hierarchical features from raw data and have increased accuracy without the need for feature construction by experts (Shen 2018; Sit et al. 2020). The entity-aware long short-term memory (EA-LSTM) network is one deep learning architecture specifically developed for environmental time series prediction using a mix of static and dynamic input drivers (Kratzert et al. 2019). These modeling and data advances provide powerful tools to support the need to create broad-coverage foundational datasets (such as water temperature). We have used the EA-LSTM approach to reconstruct the daily historical surface temperature record for 185,549 lakes in the conterminous United States from 1980 to 2020. Here, we describe the dataset, methods used to create it, provide an overview of the evaluation of the predictions, and compare this data resource to other existing methods or datasets.

Data description

This dataset, summarized in Fig. 1, includes predicted daily surface water temperatures for 185,549 lakes and reservoirs (hereafter referred to simply as lakes) in the conterminous United States (the lower 48 states and the District of

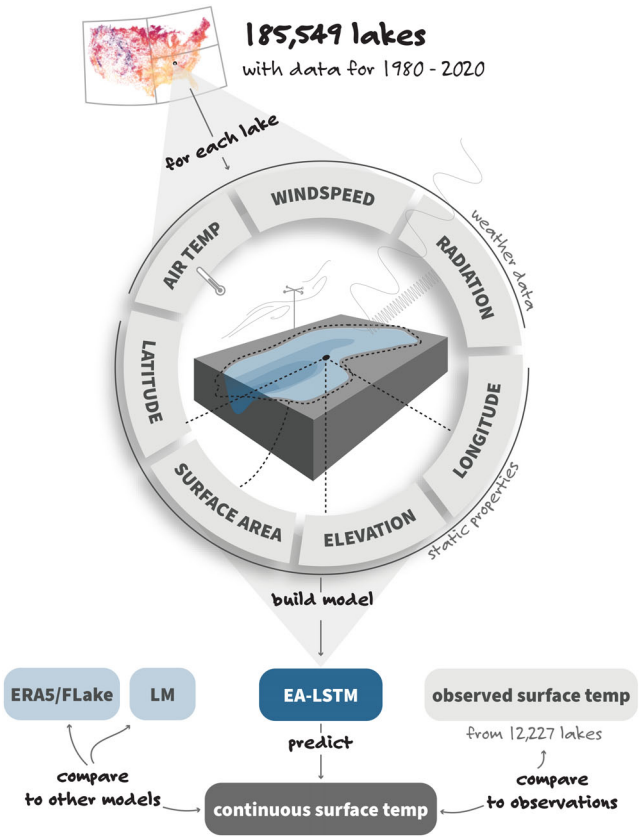

Fig. 1. Overview of the data and modeling flow used to create the continuous surface temperature predictions. The EA-LSTM neural network, a deep learning approach designed for time series and other sequential data, is built using the seven input drivers shown below as well as observed surface temperatures. EA-LSTM outputs are compared against the ERA5 reanalysis-simulated epilimnetic lake temperature outputs and a LM described in Bachmann et al. (2019). Each data component shown is available as part of the data release (https://doi.org/10.5066/P9CEMS0M). Inset map displays summer predictions for a single date and the spatial division used to break up the largest files (prediction and weather data) into three NetCDF files.

Columbia) from 1980 to 2020. Lake surface temperatures were predicted using an advanced deep learning model that is described in the Methods section; this model was compared to two published models for predicting lake surface temperatures: the ERA5 climate reanalysis aggregation of the process-based Fresh-water Lake (FLake) model (Mironov et al. 2010; Sabater 2019) and the empirical linear regression model (LM) developed by Bachmann et al. (2019). The dataset also includes data used to develop and evaluate the deep learning model, including observed water temperatures, historical downscaled weather conditions, lake-specific properties, and model evaluation metrics. Temperature observations, weather data, and lake properties were compiled from publicly available data portals and existing data publications. All data are referenced to the national hydrography dataset (NHD; Moore et al. 2019) high-resolution waterbodies using the NHD's PermID field (this dataset prefixes the value of this field with "nhdhr_"), with the exception of gridded weather data. Weather data are referenced by the longitude and latitude index of the source dataset grid cells because more than one lake can be contained within a single grid cell.

Lake surface temperature predictions are accessible from three NetCDF (Rew and Davis 1990) files covering sections of the conterminous United States as broken up by longitude and latitude boxes. Each file contains data for all lakes with surface area larger than 4 ha within each file's spatial boundary. These data include dimensions for time and NHD lake identifier and variables for surface temperature in degrees Celsius, the elevation of the lake, and the latitude and longitude of the lake centroid. Meteorological data used to drive daily temperature models are included in three additional NetCDF files that share the same spatial extents of the temperature prediction files. Meteorological data include downward longwave radiation flux, downward shortwave radiation flux, air temperature 2 m above the surface, and zonal and meridional wind speeds at 10 m above the surface. All lake surface temperature observations are included in a single comma-delimited file, with a column for lake identifier, time, observed water temperature in degrees Celsius, and estimated temperatures from each of the three temperature models. All lake-specific static values that were used to quantify lake properties were inputs to the predictive model, describe model error, or are used to connect to the appropriate NetCDF file names or indices, and are included in a single metadata file. Model accuracy was calculated using a cross validation technique (*see* the Methods section for additional details), and the root mean square error (RMSE;°C) of predicted vs. observed temperatures for lakes in each validation fold is included in the metadata file. Additionally, as mentioned above, the matchups for daily predicted and observed temperatures for each fold are available in a data file and can be used for analyzing additional dimensions of model performance not presented in this paper (e.g., estimated accuracy of predictions in a certain time of year for a particular subset of lakes).

All data files are available for download directly from https://doi.org/10.5066/P9CEMS0M using the web interface, or programmatically with the sbtools R package (Winslow et al. 2016). Example workflows for extracting surface temperatures for a single lake or all lakes for a single date are shared in the data release code repository (*see* "readme.md").

Methods

Our objective was to produce the most accurate and comprehensive predictions of daily surface water temperatures for lakes in the conterminous United States and to expose all underlying data that were used to build, drive, and evaluate these predictions to enable future expansion and comparison. Based on prior information from existing datasets and modeling efforts (Sharma et al. 2008; Read et al. 2017; Soranno et al. 2017; Bachmann et al. 2019), we excluded predictors that may be useful in temperature models but were be available broadly due to data limitations (e.g., lake depth and water clarity). We also treat our predictions as daily mean water temperatures even though observed values may be at specific times throughout the day for simplicity. Here, we describe the methods used to select models and assemble the various data included in this dataset. The code to reproduce these results is available at (https://doi.org/10.5281/zenodo.6210917).

Model descriptions

We compared three different approaches to broad-scale lake surface temperature modeling, the EA-LSTM neural network (Kratzert et al. 2019), the process-based FLake model (Spacio et al. 2008; Dutra et al. 2010; Mironov et al. 2010) used in the European Centre global reanalysis ERA5 at 0.1° latitude and longitude grid resolution (Muñoz-Sabater et al. 2021) at 17:00 UTC (Coordinated Universal Time; approximately noon local time for much of the U.S. domain), and the LM for summer temperature prediction described in Bachmann et al. (2019). This choice of methods represents state-of-the-art deep learning in the EA-LSTM, the only process-based simulation model with comprehensive global coverage via FLake and ERA5, and a simpler data-driven model in the LM. EA-LSTM is an adaptation of the standard deep learning LSTM architecture (Hochreiter and Schmidhuber 1997) for time series modeling that includes additional architectural distinction between static (e.g., lake surface area) and dynamic (e.g., air temperature) input features. Many temporal processes in environmental and engineering systems that involve complex temporal dependencies cannot be captured by a simple feed-forward ANN. LSTM models have been shown to outperform ANN models for lake temperature prediction in Jia et al. (2018), and Daw and Karpatne (2019) showed ANNs to have superior performance compared to support vector regression and boosted regression trees. However, providing time awareness to simpler machine

learning models via additional inputs (such as lagged meteorological conditions and day-of-year time vectors) can substantially boost performance (Kreakie et al. 2021), but these inputs must be selected a priori or learned from independent data to avoid overfitting to the training data. The EA-LSTM in particular has previously been applied in continental-scale rainfall-runoff modeling where it substantially outperformed all calibrated process-based hydrological models, and also showed learned similarities between different catchments that matched prior expert hydrological understanding (Kratzert et al. 2019). ERA5 makes use of the one-dimensional FLake model, a two-layer parametric representation of the dynamic water temperature profile and the integral energy budgets of these layers (for further FLake details, *see* Mironov et al. 2010). The FLake model is forced at the surface by reanalysis-derived data of wind, temperature, precipitation, humidity, and shortwave and longwave radiation. ERA5 cells for some near-coastal lakes did not have temperature estimates, and were not included in model evaluation.

Input: Meteorological conditions and lake-specific properties

Both EA-LSTM and LM models predicted surface water temperature from lake-specific properties that were static over time (log-transformed surface area, latitude, longitude, and elevation) and daily meteorological drivers that changed over time for each lake (air temperature, longwave radiation, shortwave radiation, and components of wind speed). EA-LSTM used all of these inputs, while LM only used air temperature (8 d lag-averaged), latitude, longitude, elevation, and month of the year. The choice of these dynamic features comes from the well-established understanding of connections between meteorological conditions and water temperature change (Edinger et al. 1968; Piccolroaz et al. 2018; Schmid and Read 2021). Latitude, longitude, and elevation features allow the model to learn spatial coherence in temperature, and surface area has a known role mediating lake responses to meteorological drivers (Woolway et al. 2016). Because neural networks benefit from input normalization (Sola and Sevilla 1997), an additional *z*-score normalized version of the inputs was created for the EA-LSTM based on the mean and standard deviation for each input calculated across the 12,227 observed lakes in the dataset.

The NHD (Moore et al. 2019) high-resolution polygons (based on 1 : 24,000 scale data) were downloaded as geodatabase files for each of 48 states in the conterminous United States, as well as the District of Columbia. Lakes and reservoirs were extracted using the "NHDWaterbody" layer from the geodatabase and filtered to values in the "FType" attribute that corresponded to 390, 436, and 361 (lake/pond, reservoir, and playa, respectively). The Great Lakes, several improperly labeled coastal lagoons, and lakes less than 4 ha (based on the value in the "AreaSqKm" NHD attribute) were removed from the dataset, and the remaining 185,549 lakes defined the complete lake coverage used in this data release.

Hourly meteorological data for the five variables described above were downloaded from the North American Land Data Assimilation System (NLDAS; Mitchell et al. 2004); we used a NASA earthdata login to access NetCDF files through https://hydro1.gesdisc.eosdis.nasa.gov/dods/NLDAS_FORA0125_H.002, and daily datasets were created by applying a U.S. central time zone offset for the entire spatial range and calculating the daily mean for each variable. The 0.125° NLDAS latitude and longitude grid was then used to assign NLDAS grid IDs to each lake's centroid using the "st_centroid" and "st_intersects" functions from the "sf" R package (Pebesma 2018). All grid cells that did not contain a lake were excluded and the remaining daily dataset was transformed from a spatial grid (latitude, longitude, and time) into a flatter and smaller discrete sampling geometry NetCDF format (Blodgett and Winslow 2019) indexed to grid ID and time.

Approximate lake surface area and elevations were calculated based on the vector polygon data ("st_area"; Pebesma 2018) and lake centroid, respectively. Lake surface elevation was estimated for each lake using the "get_aws_points" function from the "elevatr" package (Hollister 2021) at the zoom level of nine, providing a centroid-based value in meters for each lake from an elevation raster from the Shuttle Radar Topography Mission data (Farr et al. 2007).

In situ lake temperature data

Lake temperature data were compiled from two main sources: digitized or spreadsheet-based historical records shared directly with researchers (Read et al. 2021) and through programmatic access to discrete monitoring data in the joint Environmental Protection Agency and U.S. Geological Survey WQP (Read et al. 2017). High-frequency buoy data and remote sensing data were not used in this dataset due to extreme differences in temporal coverage that would favor a small number of lakes (as in the case of buoy data) and the large drop in measurement accuracy in satellite-based estimates of surface water temperatures when compared to in situ observations (e.g., mean absolute error (MAE) ranging from 1.34°C to 4.89°C depending on distance to lake shore with the Landsat analysis ready surface temperature product; Schaeffer et al. 2018). Prior compiled data from Read et al. (2021) included temperatures from lakes in U.S. midwestern states and were combined with updated national pulls of water temperature data from the WQP from 1980 to 2020. Unique WQP lake monitoring sites with temperature data were captured by breaking the spatial extent of the conterminous United States into 2.5° by 2.5° latitude/longitude cells and calling "whatWQPdata" function from the "dataRetrieval" R package (Hirsch and De Cicco 2015) for "Lake, Reservoir, Impoundment" siteTypes and "Temperature," "Temperature, sample," "Temperature, water," and "Temperature, water, deg F" characteristicNames on each cell's bounding box. Monitoring sites were then ranked according to expected

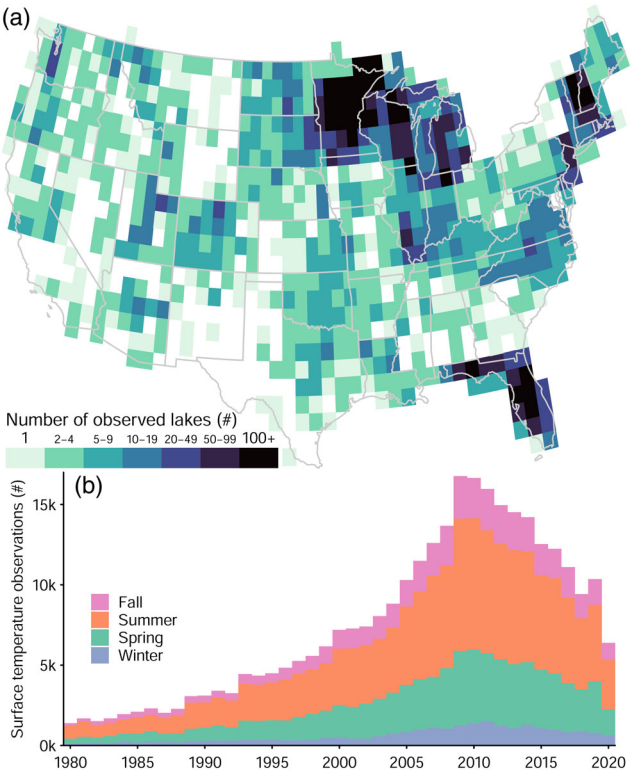

Fig. 2. Geographic and temporal coverage of in situ surface temperature data. Panel (**a**) shows geographic coverage of the 12,227 observed lakes across single degree latitude and longitude cells in the conterminous United States. Panel (**b**) shows observations by season and by year between 1980 and 2020.

number of observations (the "resultCount" value from the "whatWQPdata" result) and broken up into site groups containing no more than 500,000 total results or less than 200 unique sites, and each site group was queried for all available temperature data using the same characteristicNames as listed above. Resulting data were converted into standard depth as measured in meters and temperature as measured in degrees Celsius and then all observations deeper than 1 m were removed and basic quality control measures were applied (*see* the Technical validation methods section). Monitoring site locations, which are defined by a single spatial location, were joined to lakes by using point-in-polygon analysis and sites falling outside of the 185,549 lakes in this data release were excluded. The above process resulted in 306,553 in situ temperature observations from 12,227 lakes for model development. Geographic coverage density of the observed lakes is shown in Fig. 2a, and the temporal coverage is shown in Fig. 2b.

Oversampling

Only 955 temperature observations (0.3% of total) were greater or equal to 33°C. To compensate for a lack of very high temperatures leading to an observation distribution imbalance, we used a simple random oversampling method. Oversampling duplicates samples from a minority class, or in this case a minority temperature range, to address data imbalances for statistical or machine learning models (Japkowicz and Stephen 2002; Estabrooks et al. 2004). First, we defined a histogram with forty 1°C bins covering the range of temperatures from 0°C to 40°C. Then, a normal distribution curve was fit to the histogram (mean $\mu = 20.32$, standard deviation $\sigma = 6.89$). The normal curve is a common distribution to use for oversampling (Pan et al. 2020) that includes a smooth decline with an asymptote at 0°C. For each temperature bin between 33°C and 40°C with sample counts below the normal curve, we randomly oversampled with small added noise (0.0125/0.125 variance Gaussian noise on normalized features/unnormalized observations, respectively) until the bin height matched the mean of normal curve points at both sides of the bin. This added an additional 20,377 (6.6%) observations ranging from 33°C to 40°C to the final training dataset. For the cross-validation setup used for hyperparameter tuning and error estimation described in the following subsections, oversampling was specifically done on only the training data and no observations from the test data were duplicated.

Hyperparameter tuning

As with most deep learning models, EA-LSTM requires tuning of hyperparameters for optimal performance. In machine learning, a hyperparameter is a parameter used to control the learning process and/or the network architecture. By contrast, the values of other parameters (typically network weights) are tuned during training. Here, we tuned the hyperparameter that defined the number of epochs used to train the model and also recorded the training MSE at the optimal number of epochs as another stopping condition. The number of epochs was tuned within the inner loop of the fivefold nested cross-validation (Tibshirani et al. 2009) shown visually in Fig. 3. To ensure lake diversity representation across folds, the lakes were first divided into 16 clusters using *k*-means clustering (Lloyd 1982) on latitude, longitude, and the natural log of the surface area values that had been *z*-score normalized to a mean of 0 and standard deviation of 1. Each cluster was then equally divided among the five folds to create the final fold groupings. The oversampling method previously described was used in each training dataset. The number of epochs was found for each of the five test folds by calculating where the mean validation MSE across the remaining four training folds was the lowest. The optimal values of epochs for each of the folds 1–5 were 250, 160, 210, 280, and 280, respectively. We also computed the mean training MSE across the four folds at the optimal epoch of each instance as another measure of model fitting which were 1.98, 2.01, 1.98, and 2.02 (taken from previous studies on lake temperature prediction using LSTM; Jia et al. 2019; Read et al. 2019), and 2.10°C, respectively. Other EA-LSTM hyperparameters set were a sequence length of 350 d, 256 hidden unit size, learning rate of 0.005,

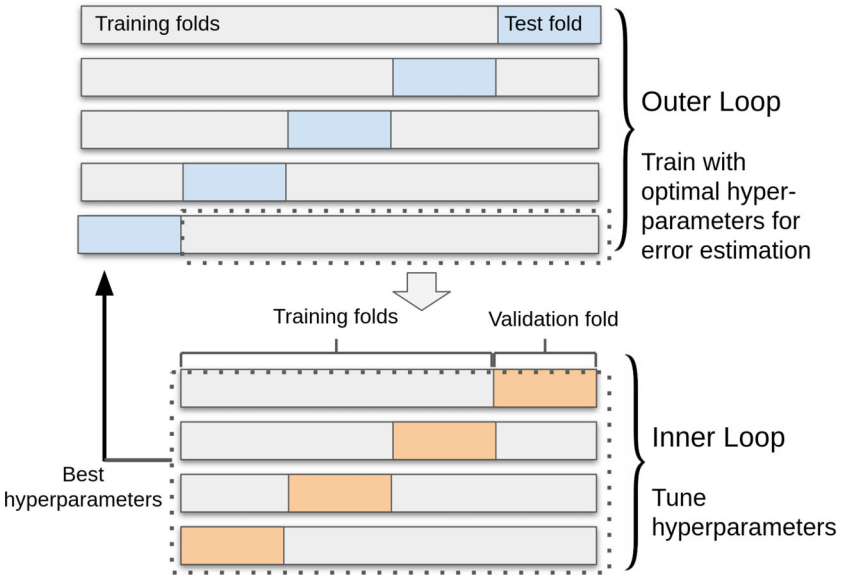

Fig. 3. Nested cross-validation process. Performance is aggregated over the fivefold outer loop where each instance of training folds also contains an inner fourfold loop for hyperparameter tuning on validation data. Hyperparameters are selected to minimize error across validation folds.

use of the Adam optimizer (Kingma and Ba 2017) and an MSE loss function, gradient clipping set to 1.0 of the 2-norm of the network weights, and a batch size of 3000 sequences. All final values are also captured in the modeling code release (https://doi.org/10.5281/zenodo.6210917).

Error estimation

To estimate model performance for the two data-driven approaches, we used the outer loop of the fivefold cross-validation shown in Fig. 3 and compared the mean out-of-fold test error across folds for each model. Each set of test data was held out of any model training or hyperparameter tuning, and also the 70 lakes not covered by ERA5 were included in training but excluded from test error calculation. Hyperparameters for each of the five models were found through the inner cross-validation loop described previously, and training data consisted of observations from the remaining $\sim 80\%$ of lakes that were not included in the test fold. The previously described oversampling method was also used to augment each training dataset with more high temperature observations, and the data splits for EA-LSTM and LM were identical. Compared to the following LM fit published in Bachmann et al. (2019),

$$\widehat{T} = 16.14 + 0.673\text{Air} - 0.0846\text{Lat} + 0.0172\text{Long} - 0.00131\text{Elev} - 0.147\text{Mon}$$

where the average over the folds of CV for each of the coefficients (air temperature [Air], latitude [Lat], longitude [Long], elevation [Elev], and month [Mon]) became the following equation:

$$\widehat{T} = 20.368 + 0.580\text{Air} - 0.159\text{Lat} + 0.0347\text{Long} - 0.0015\text{Elev} + 0.177\text{Mon}$$

For the ERA5 process-based model used for comparison, we also bias-corrected the output by adding 3.31°C to all predictions (referred to as ERA5*). This bias correction addressed a clear cold bias that currently exists in ERA5 in U.S. lakes (e.g., Betts et al. 2020 found a 4°C cold bias of ERA5 on Lake Champlain in late spring; Muñoz-Sabater et al. 2021 reported a general cold bias across many lakes). The amount of bias correction was decided based on the intercept of a linear regression with slope 1 fit to observed vs. ERA5-predicted temperatures.

Training EA-LSTM and prediction of 185,549 lakes

The final model used to generate predictions for 185,549 lakes was trained on all available surface temperature observation data from 12,227 lakes. Hyperparameter values that minimized validation error across all inner loops in the nested cross-validation were selected for the final aggregate data model (220 for the number of training epochs and 2.03 for the training MSE stopping condition). The remainder of the hyperparameters and model architecture were kept the same as during the error estimation phase, and oversampling was also applied. Using the trained model, predictions were generated for all 185,549 lakes.

Technical validation methods

We used the test data from the error estimation phase to estimate overall prediction accuracy, in addition to analyses of accuracy across geographical regions in the United States, different water temperature ranges, different years and

seasons, and different lakes. We also sought to identify potential data concerns or limitations that may affect future users of this data. All technical validation described here is transparent and reproducible using the code repositories linked at the beginning of the paper. Technical validation performed includes the error estimation for modeled temperature, assessment of model bias in various conditions, and the quality assurance and quality control (QAQC) procedures for building the in situ dataset.

The previously described error estimation method was the primary validation of overall accuracy, where all prediction errors were calculated on lakes not used for model training or hyperparameter tuning to mirror the situation of predicting on unmonitored lakes. The folds and clusters used to divide the lakes for training and validation are representative of the broader population of lakes due to (1) the *k*-means clustering grouping lakes with respect to geographical location and lake size, and (2) the even split among each cluster distributed evenly among the testing folds.

Observed temperature data were screened and unrealistic values were removed using a variety of techniques, including visual inspection, comparison to published models, and evaluating based on date or season to find likely errant data sources. While some of these steps were manual (e.g., visual inspection and contacting monitoring organizations to confirm and fix errant data entry), all alterations to the data, including unit conversions and data screening, were captured in code (*see* "lake-surface-temperature-prep" code at https://doi.org/10.5281/zenodo.6210917 for data processing). In the WQP data, numerous sites had data that were entered incorrectly for some or all measurements (*see* Sprague et al. 2017 for an overview of similar issues with nutrient metadata). Any observations that likely represented conditions from environments other than the lake water were removed, including by examining metadata fields or contacting data contributors directly. Patterns in temperature time series that suggested the data were flawed were also used to remove values and sites; sites were removed based on various visual or statistical cues (e.g., single measured values that repeated without any deviation) that suggested all site data were suspect. Additionally, the lower resolution (0.25° lat/lon) aggregated version of the ERA5 temperature estimates were used to determine extreme outliers based on exceeding 10°C above or below a bias-corrected temperature estimate (ERA5 + 3.47°C) (Hersbach and Dee 2016) and the resulting outliers were removed from the dataset. If more than one observation was reported on the same day at the same depth on the same lake, we applied the following strategy: we selected the shallower observation followed by the warmer measurement (in the case of identical depths).

Results of technical validation

After outlier removal and the selection of single values to represent a unique lake on a given date, the final dataset of observed temperatures included 306,553 near-surface (between 0 and 1 m deep, inclusive) observations from 12,227 lakes. Outliers removed include the following: (1) 7056 values were removed because "Temperature at lab" was mentioned in the "ResultCommentText" even if the other metadata indicated the measurement was made from the lake, (2) 7464 additional values were removed that included "Lab" in the "ResultAnalyticalMethod/MethodIdentifier" field as this metadata value indicated these observations of temperature were related to a laboratory measurement or extraction of another variable, (3) 3746 values from all monitoring sites prefixed with "IL_EPA" and a "CharacteristicName" of "Temperature, sample" were removed after confirmation that these temperatures were not measured directly from the lake, (4) 961 values were discarded when several monitoring sites from various agencies were removed after discovering the data were unrealistic (these sites were removed based on visual comparison to neighboring sites, because values were repeated constantly throughout the season without changing, or because reported depths were likely referenced from the bottom of the lake instead of the surface), and (5) 981 additional values were removed because they exceeded 10°C above or below the bias-corrected aggregated ERA5 temperature estimate. Despite this effort to remove errant data, it is very likely that observation errors beyond the expected range of sensor accuracy still exist in the final dataset, but we expect these issues are rare by comparison.

For the 12,227 lakes with observed temperature, 70 did not overlap ERA5 grid cells (these lakes were near coastlines), and were not included in model evaluation. The remaining 12,157 lakes and 303,579 observations had a median lake-specific RMSE (1st to 3rd quartile) for all test folds of 1.24°C (0.86–1.73°C) for EA-LSTM and 3.95°C (3.12–4.84°C) for ERA5 (Table 1). After addressing the cold bias of ERA5 by subtracting 3.31°C (denoted ERA5*), the median lake-specific RMSE of ERA5* was 1.79°C (1.25–2.57°C). The original Bachmann et al. (2019) model was constrained to periods between 01 June and 30 September, which was followed by retraining and evaluating that model only using observations from those months. The associated data released with that study was also limited to those months and was on a smaller scale than is shown here (Bachmann et al. 2019 used 1905 lakes). Here, LM predictions had a lake-specific median RMSE of 2.01°C (1.32–2.57°C), compared to 1.17°C (0.78–1.68°C) for EA-LSTM and 1.70°C (1.12–2.43°C) for ERA5* during the same months. Overall RMSE for the summer months was 1.55°C for EA-LSTM, 2.27°C for ERA5*, and 2.35°C for LM. All other presentations of LM predictions hereafter (in figures and text) are restricted to this time period as well. Five hundred thirty-four lakes had observations only outside the summer period and were excluded from the LM error calculations.

The global accuracy of each model (assessed by calculating the RMSE of all data across all test folds at once) was 1.61°C for EA-LSTM, 2.34°C for ERA5*, 4.06°C for ERA5, and 2.35°C

Table 1. Performance comparison of the three modeling approaches across the five test folds in cross-validation. Here, ERA5* is the bias-corrected version of ERA5 (an offset of +3.31°C was applied to the ERA5 data), and LM is only tested on data from June to September. From left to right, median lake-specific RMSE and overall RMSE assess overall performance, then median RMSE is shown for lakes within different size ranges, and lastly median bias of all observations in different temperature ranges is shown (all values are in°C units). Bias for bias-corrected ERA5* is not shown because observations were used in the bias correction itself, and bias in the lowest temperature range is not shown for LM due to lack of data. Numbers in parentheses represent the number of lakes (lake size) and observations (temperature group) in each data partition with the exception of the LM observations, which are lower due to their restriction to the summer months, and the ERA5 comparisons, which have 2974 fewer observations from 70 coastal lakes that are not resolved in the dataset.

			Median RMSE by lake size (ha)				Median bias by observation temperature (°C)			
	Median lake-specific RMSE	Overall RMSE	< 10 (1946)	10–100 (6707)	100–1000 (2949)	> 1000 (685)	0–10 (28,196)	10–20 (98,298)	20–30 (170,114)	30+ (15,451)
EA-LSTM	**1.24**	**1.61**	**1.24**	**1.18**	**1.27**	**1.61**	**0.38**	**0.29**	**0.10**	**−0.08**
ERA5*	1.77	2.24	1.75	1.73	1.80	2.03	NA	NA	NA	NA
ERA5	4.04	4.11	3.78	4.07	4.01	3.70	−2.37	−3.35	−3.49	−3.18
LM	2.00	2.34	1.70	1.98	2.10	2.13	NA	2.21	−0.38	−0.53

Best in column shown in bold.

for LM (Table 1; Fig. 4b,e,h). The cold bias in ERA5 is greatly reduced by applying a simple offset of +3.31°C to all ERA5 predictions (RMSE of 4.06–2.34°C; Table 1; Fig. 5e). Spatial patterns in prediction accuracy (estimated by calculating RMSE from test fold data in 1° latitude/longitude cells) showed no clear latitudinal differences for EA-LSTM and ERA5* but temperature predictions from the LM were more accurate in the southern state of Florida compared to the similarly data-rich states of Minnesota and Wisconsin (Fig. 4a,d,g). Predictive accuracy varied over time. Year-specific RMSE for EA-LSTM decreased through time; the maximum single year RMSE was 2.30°C in 1980 and minimum was 1.41°C in 2019, with a clear negative trend (Fig. 4c). Yearly ERA5* and LM RMSEs did not have a clear temporal trend (Fig. 4f,i) and ranged from 2.13°C to 2.89°C and 2.09°C to 2.76°C, respectively.

Predictions from all three models were biased for some or all data subsets (Fig. 5). Temperature predictions from the ERA5 had the greatest overall bias (specifically, the model was biased cold for all data subsets). The median bias across 10° temperature bins ranged from −0.08°C to 0.38°C for the EA-LSTM and −0.51°C to 2.29°C for LM (Table 1). Bias was greatest for all models for the coldest and warmest temperatures when finer temperature bins were used (Fig. 5b,e,h). The EA-LSTM model had a consistent warm bias across all years (Fig. 5a). When evaluated across temperature bins and seasons, predictions from EA-LSTM were most frequently warm biased, although cold biases existed for both cold/winter conditions and for extremely warm temperatures, which were substantially underpredicted by the model (Fig. 5b,c). The warmest temperatures were underpredicted by both LM and ERA5 models as well (Fig. 5e,f,h,i). The LM overpredicted temperatures at the lower end of the temperature distribution (5 h), but these temperature conditions were rare in the truncated June to September datasets that the LM model was trained and evaluated on (e.g., the 4.86°C median LM warm bias in the 10–12°C observed temperature range is based on 0.3% of test observations). Similarly, the extremely warm observations that all models struggled to reproduce were comparatively rare, as the −1.90°C, −4.03°C, −1.01°C LM, ERA5, and EA-LSTM median biases in the 32–34°C range included only 0.8% of test observations and only 0.1% of data were in the 34–36°C temperature range.

The complete set of 306,553 observations were validated against the final EA-LSTM model trained using all of the same data to see if the model was overfitting and verify prediction performance. The median lake-specific RMSE (1st to 3rd quartile) was 1.17°C (0.82–1.63°C) indicating a small decrease in error and suggesting overfitting of this model is unlikely.

Data use and recommendations for reuse

Surface water temperature estimates are useful for improving scientific outcomes in fisheries biology, limnology, and climate science. Specifically, these data (1) facilitate improved understanding of lake temperature dynamics in undermonitored and unmonitored locations, (2) enable investigators to scale up from traditional single or multisite field science to science at broad spatial scales, and (3) extend a foundational limnological data resource (LAGOS-US; Cheruvelil et al. 2021) by linking these weather and temperature predictions to numerous lake properties through common

lake identifiers. Across applications, this dataset provides the best available surface temperature accuracy at the scale of the conterminous United States. Additionally, example data access scripts for both Python and R are included in the "lakesurf-data-release" code at https://doi.org/10.5281/zenodo.6210917 to facilitate future users.

At the local to regional scale, this dataset provides essential data to parameterize models that use surface water temperature as an input (e.g., harmful algal bloom prediction [Wynne et al. 2013], gas solubility estimates [Tromans 1998], and fish bioenergetics models [Deslauriers et al. 2017]). This dataset has the potential to similarly inform improvements to other limnological data products by refining ancillary temperature estimates, including satellite derived surface temperatures (Hulley et al. 2011; Dörnhöfer et al. 2016; Schaeffer et al. 2018). When combined with additional observational data, the historical reconstruction of temperature provided here can further our understanding of how temperature correlates with overall water quality dynamics, nutrient loading (Oleksy et al. 2021), and algal bloom frequency (Larkin and Adams 2013).

This landscape-scale dataset could support a more systematic understanding of the extent of lake synchrony in response to multi-scale forcings like climate and land use change (Lottig et al. 2017). Lake temperature as a major ecological control is also important for quantifying other macroscale ecosystem properties, such as the contribution of aquatic ecosystems to continental and global carbon cycles (Raymond et al. 2013; Mendonça et al. 2017; Bartosiewicz et al. 2019). Existing approaches for quantifying lake contributions to carbon budgets rely on spatiotemporally inconsistent data (including temperature) and can be substantially improved by using comprehensive landscape-scale datasets (McDonald et al. 2013).

Across scales, surface temperature estimates can be used to estimate changes in thermal parameters related to fish spawning, growth, and abundance. Previous work has shown that population dynamics of cool- and warmwater fishes are well-predicted by surface temperature metrics (Hansen et al. 2015; Massie et al. 2021), even in stratified lakes with diverse thermal habitats. However, changes in surface water temperatures alone may be a poor proxy for estimating changes to the thermal environment of coldwater fishes or other organisms occupying the bottom waters of stratified lakes (Kraemer et al. 2015; Winslow et al. 2015). Understanding these shifts in thermal regimes will become increasingly important as climate velocities (the pace of warming compared to a species' ability to migrate to cooler habitats) increase throughout the next century (Woolway et al. 2020; Woolway and Maberly 2020). This dataset provides an essential baseline of historical temperatures upon which to compare these future changes.

These data may be suitable for evaluating the effects of climate change on lake temperatures, but caution should be applied when calculating warming trends and other metrics of change. This unprecedented collection of accurate daily surface temperature predictions, supporting in situ measurements, and prevailing weather for 1980–2020 covers a period of lake warming that has already altered the structure and function of many aquatic ecosystems (O'Reilly et al. 2015; Woolway et al. 2020). Gray et al. (2018) provide careful instruction and caution for calculating trends from the in situ data in this release. If using the EA-LSTM predictions to calculate temperature change metrics, the model's limitations (such as the absence of ice-cover) would need to be explored and found to be robust for the interpretation of the chosen metrics. Additionally, model accuracy is higher during more recent time periods (such as 2005–2020; Fig. 4c) and the

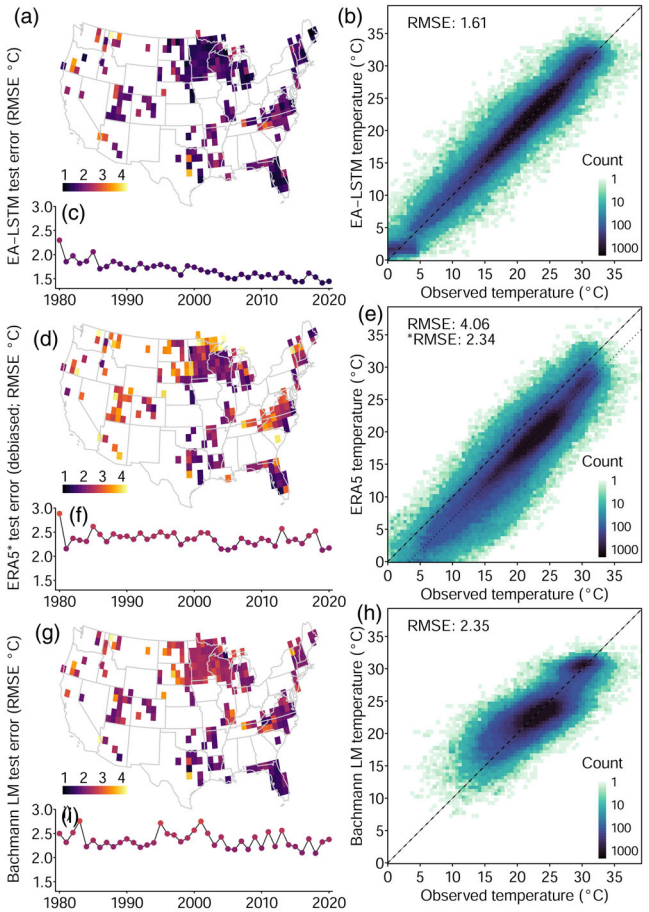

Fig. 4. RMSE for predicted compared to observed water temperatures within a single degree latitude and longitude cell for each of the three methods is shown in panels (**a**), (**d**), and (**g**). Only cells with at least 100 observations are shown. Panels (**c**), (**f**), and (**i**) show year-specific RMSE per method. Panels (**d**) and (**f**) specifically shows the bias-corrected ERA5 errors (ERA5* in Table 1). The distributions of all 306,553 observations along with a 1 : 1 line are shown in panels (**b**) and (**e**) for EA-LSTM and ERA5 respectively, and panel (**h**) shows the 188,886 summer observations predicted by LM. An additional 1 : 1 dotted line is shown in panel (**e**) with a *y*-intercept of −3.31 to represent the bias-corrected version of ERA5.

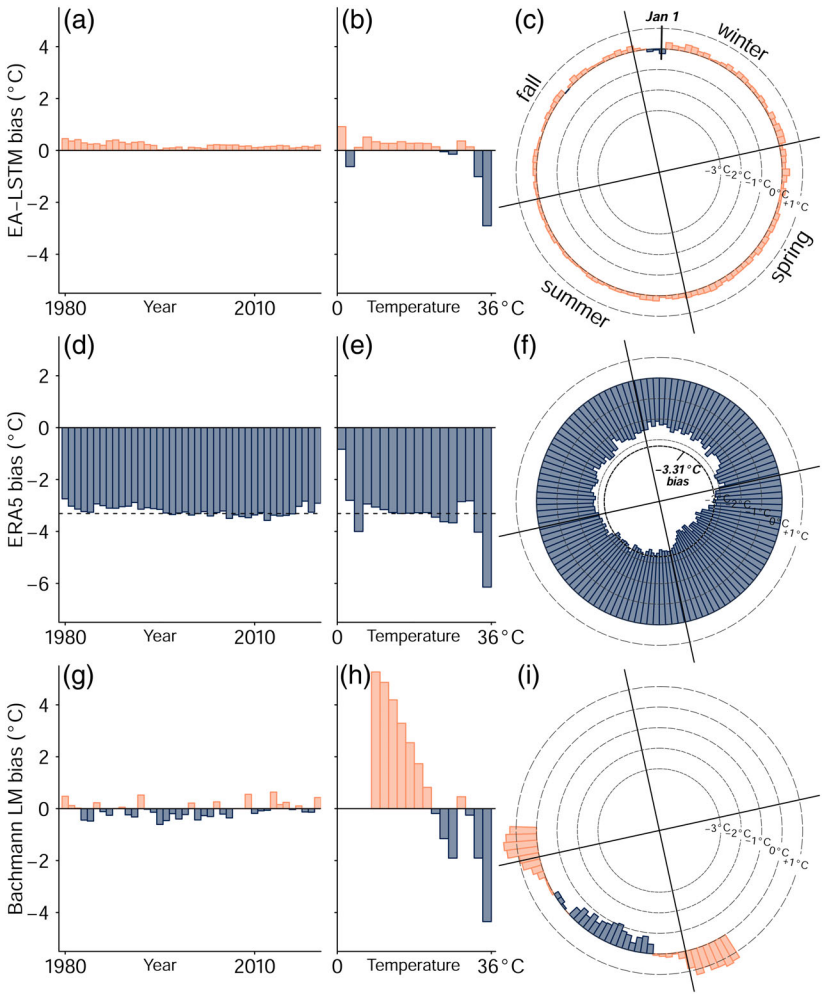

Fig. 5. Bias of predicted compared to observed water temperatures for all three approaches. Panels (**a**), (**d**), and (**g**) show median bias per year ranging from 1980 to 2020. Panels (**b**), (**e**), and (**h**) show bias per 2°C temperature bins ranging from 0°C to 36°C. Day of year median bias is shown in panels (**c**), (**f**), and (**i**) with seasonal divisions. The dotted line in panels (**d**) and (**e**) represents the −3.31°C shift for bias corrected ERA5 predictions (ERA5* in Table 1).

model shows bias in different directions across the range of predicted temperatures (Fig. 5b). These patterns in accuracy and bias warrant consideration when estimating warming trends from predictions.

While the oversampling technique used in this study increased the model's exposure to warm extremes (e.g., temperatures above 33°C), we were unable to validate the performance of predicting the timing and value of a yearly maximum water temperature. This shortcoming is because observations in this dataset are discrete and therefore missing many events that would be captured in continuous measurements. Likewise, the weather data included here offer powerful context for understanding limnological change, but are also generated by a combination of modeling and data collection (Mitchell et al. 2004), and are therefore subject to similar caution. Despite these limitations, this new dataset is the best available (in terms of accuracy and coverage) for macroscale aquatic research that can be informed by changes in lake temperatures.

Comparison with existing datasets

Daily surface water temperature predictions for lakes in the conterminous United States using the EA-LSTM are more accurate and less biased when compared to currently available models with similar or greater temporal and spatial coverage (Figs. 4, 5). The EA-LSTM outperformed ERA5 and LM temperature predictions based on the RMSE of all data subsets assessed, including global RMSE and RMSE for binned lake size classes (all observations; Table 1). Spatially, the EA-LSTM was best for 74% (63% accounting for ERA5*) of the 12,157 lakes used for model evaluation, as well as across 82% (80% including ERA5*) of the 220 1° latitude/longitude cells that

had at least 100 observations (Fig. 4). The EA-LSTM maintained the lowest RMSE across all 41 study years regardless of ERA5 debiasing.

We found a significant cold bias in ERA5 predictions that was similar for all years (Fig. 5d), but varied seasonally and across the range of temperatures (Fig. 5e,f) which is consistent with Betts et al. (2020). Bias correction may be needed for most uses of the current ERA5 mixed layer temperature predictions. The EA-LSTM outputs included in this dataset have a small warm bias that is mostly consistent seasonally and across years (Fig. 5a,c), but predictions are cold compared to the warmest observations and warm compared to the coldest observations (Fig. 5b). The Bachmann LM model had no bias across years (Fig. 5g), but was substantially biased across the range of temperatures, overpredicting colder temperatures and underpredicting warmer temperatures (5 h), and this pattern also appeared as strong seasonal model biases (Fig. 5i).

The accuracy of the EA-LSTM predictions compare favorably to other efforts on smaller numbers of lakes, including the global analysis of 235 lakes by O'Reilly et al. (2015); RMSE 1.68–2.15°C from linear regression), and regional process-based predictions of temperatures by Winslow et al. (2017); epilimnetic temperature RMSE of 1.91°C; $n = 72{,}232$). Recent summer surface temperature predictions for 2186 U.S. lakes by Kreakie et al. (2021) had similar accuracy to the EA-LSTM (1.48°C vs. 1.50°C RSME when comparing summer errors in 2007 and 2012, the 2 yr of their model) but the performance of their random forest model was not evaluated on unseen lakes and conditions (e.g., the additional 39 yr and 10,041 lakes included in this study). Satellite-based remote sensing sources of estimated surface temperature are promising, and can approach the accuracy of the EA-LSTM model presented here in certain cases (e.g., Schaeffer et al. 2018 found MAE of Landsat water pixels > 180 m from shore was 1.34°C; the EA-LSTM presented here has an MAE of 1.16°C).

The in situ measurements shared in this dataset have two orders of magnitude more observations compared to those made available in Bachmann et al. (2019); 306,553 and 2655 observations, respectively) and an unprecedented number of U.S. lakes (12,227 lakes). While the in situ data in this dataset can be accessed elsewhere, the significant effort to query, download, and screen data, in addition to the process to match temperature monitoring sites to individual lakes has resulted in a dataset that can be rapidly leveraged for future studies. Specifically, the QAQC of data from the WQP (Read et al. 2017) and site linking to lakes adds substantial value to those existing resources. A similar global compilation effort by Sharma et al. (2015) produced summer temperatures and metadata for 291 lakes that has been used extensively to quantify the effect of climate change on lake temperatures (O'Reilly et al. 2015; Kraemer et al. 2017), and we expect these in situ data to also support new aquatic science efforts. The dataset described in this article does not include data collected using automated sensors nor remotely sensed data, but either could be combined with these observations to extend the dataset.

The predicted surface temperatures for 185,549 lakes includes full coverage of lakes with surface area larger than 4 ha in the conterminous United States, which is a substantial expansion in scale or resolution compared to other available modeled temperature data products. The ERA5-simulated epilimnetic lake temperatures provide coverage of the great majority of lakes globally, but the gridded cells overlapping the lake centroids of this conterminous United States dataset have far fewer unique time series (42,354 for ERA5 vs. 185,549 here). Many of the ERA5 0.1° latitude and longitude grid cells aggregate multiple lakes into the lake tiles that are available in the ERA5 dataset. However, the ERA5 dataset does include hourly temperatures that could be useful for comparing minimum and maximum temperature ranges; our model generates a single prediction for each lake-day. Other existing process-based lake temperature predictions from Winslow et al. (2017) and Read et al. (2021) cover a smaller spatial extent, and within those regions, represent a smaller number of lakes due to a requirement parameterizing lake depth for the individual models. Semi-process-based approaches have been applied at a larger scales with good results in Gillis et al. (2021) and also with the air2water model (Piccolroaz et al. 2013; Toffolon et al. 2014; Piccolroaz 2016; Piccolroaz et al. 2018; Heddam et al. 2020). However, these approaches are also limited by the requirement of lake depth which is readily available only for a small subset ($n = 17{,}675$) of all lakes in the conterminous United States for lakes with surface area bigger than 1 ha (3.7% of 479,950) that are available in LAGOS-US (Stachelek et al. 2021). The ERA5 predictions overcome this limitation by using an estimated lake depth product that is available globally (Kourzeneva 2010). Our modeled temperatures have a similar coverage to the possible extents of the data-driven approach of Bachmann et al. (2019), but those models were not released with predictions or inputs beyond the observed lakes used for training and testing the models and are additionally limited to the summer months.

We used NHD HR permanent identifiers to enable synergistic interactions with existing datasets including LAGOS-US (Cheruvelil et al. 2021), the National Anthropogenic Barriers Dataset (Ostroff et al. 2013; Cooper et al. 2017), and the National Lakes Assessment (Pollard et al. 2018). Using GIS, the data provided can be linked to additional lake and catchment properties within the WQP (Read et al. 2017), HydroLakes (Messager et al. 2016), and the Global Lake Area, Climate, and Population dataset (Meyer et al. 2020). In combination, these macroscale datasets provide a suite of lake and catchment properties and multitemporal measurements of water quality, anthropogenic stressors, land use, and meteorological variables. This wealth of information creates novel opportunities for modeling lake systems and examining synoptic patterns in freshwater resources at the landscape scale.

While the contribution of estimated and observed water temperatures provided here is highly valuable as a stand-alone resource, the inclusion of lake-level climate and meteorological data at the daily timescale provides additional benefits not currently captured in the datasets discussed above.

Leveraging the above interconnected datasets and/or future datasets of lake properties could likely lead to modeling efforts that outperform the EA-LSTM model presented here. With future development in mind, and to maximize the utility of the provided dataset, all modeling inputs, data partitioning, training data, modeling code, and EA-LSTM predictions are accessible through this dataset. By providing this end-to-end pipeline, we aim to create continued opportunities for comparison and modeling improvements. Data such as upstream inflow, reservoir release information, and land use may allow a future model to better capture abrupt changes in temperature or to predict more accurate temperature extremes.

References

Bachmann, R. W., S. Sharma, D. E. Canfield, and V. Lecours. 2019. The distribution and prediction of summer near-furface water temperatures in lakes of the coterminous United States and southern Canada. Geosciences **9**: 296. doi:10.3390/geosciences9070296

Bartosiewicz, M., A. Przytulska, J. Lapierre, I. Laurion, M. F. Lehmann, and R. Maranger. 2019. Hot tops, cold bottoms: Synergistic climate warming and shielding effects increase carbon burial in lakes. Limnol. Oceanogr.: Letters **4**: 132–144. doi:10.1002/lol2.10117

Betts A. K., D. Reid, and C. Crossett. 2020. Evaluation of the FLake Model in ERA5 for Lake Champlain. Frontiers in Environmental Science **8**: 250. doi:10.3389/fenvs.2020.609254

Blodgett, D., and L. Winslow. 2019. ncdfgeom: "NetCDF" geometry and time series.

Cheruvelil, K. S., P. A. Soranno, I. M. McCullough, K. E. Webster, L. K. Rodriguez, and N. J. Smith. 2021. LAGOS-US LOCUS v1.0: Data module of location, identifiers, and physical characteristics of lakes and their watersheds in the conterminous U.S. Limnol. Oceanogr.: Letters **6**: 270–292. doi:10.1002/lol2.10203

Cooper, A. R., D. M. Infante, W. M. Daniel, K. E. Wehrly, L. Wang, and T. O. Brenden. 2017. Assessment of dam effects on streams and fish assemblages of the conterminous USA. Sci. Total Environ. **586**: 879–889.

Daw, A., and A. Karpatne. 2019. Physics-aware Architecture of neural networks for uncertainty quantification: application in lake temperature modeling, in: FEED Workshop at Knowledge Discovery and Data Mining Conference (SIGKDD) 2019. SIGKDD.

Deslauriers, D., S. R. Chipps, J. E. Breck, J. A. Rice, and C. P. Madenjian. 2017. Fish bioenergetics 4.0: An R-based modeling application. Fisheries **42**: 586–596.

Dörnhöfer, K., A. Göritz, P. Gege, B. Pflug, and N. Oppelt. 2016. Water constituents and water depth retrieval from Sentinel-2A—a first evaluation in an oligotrophic lake. Remote Sens. (Basel) **8**: 941.

Dutra, E., V. M. Stepanenko, G. Balsamo, P. Viterbo, P. Miranda, D. Mironov, and C. Schär. 2010. An offline study of the impact of lakes on the performance of the ECMWF surface scheme.

Edinger, J. E., D. W. Duttweiler, and J. C. Geyer. 1968. The response of water temperatures to meteorological conditions. Water Resour. Res. **4**: 1137–1143. doi:10.1029/WR004i005p01137

Estabrooks, A., T. Jo, and N. Japkowicz. 2004. A multiple resampling method for learning from imbalanced data sets. Comput. Intell. **20**: 18–36. doi:10.1111/j.0824-7935.2004.t01-1-00228.x

Fang, X., H. G. Stefan, J. G. Eaton, J. H. McCormick, and S. R. Alam. 2004. Simulation of thermal/dissolved oxygen habitat for fishes in lakes under different climate scenarios. Ecol. Model. **172**: 13–37. doi:10.1016/S0304-3800(03)00282-5

Farr, T. G., et al. 2007. The shuttle radar topography mission. Rev. Geophys. **45**: RG2004. doi:10.1029/2005RG000183

Gillis, D. P., C. K. Minns, and B. J. Shuter. 2021. Predicting open-water thermal regimes of temperate North American lakes. Can. J. Fish. Aquat. Sci. **78**: cjfas-2020-0140. doi:10.1139/cjfas-2020-0140

Gray, D. K., S. E. Hampton, C. M. O'Reilly, S. Sharma, and R. S. Cohen. 2018. How do data collection and processing methods impact the accuracy of long-term trend estimation in lake surface-water temperatures? Limnol. Oceanogr.: Methods **16**: 504–515.

Hansen, G. J. A., S. R. Carpenter, J. W. Gaeta, J. M. Hennessy, and M. J. Vander Zanden. 2015. Predicting walleye recruitment as a tool for prioritizing management actions. Can. J. Fish. Aquat. Sci. **72**: 661–672. doi:10.1139/cjfas-2014-0513

Heddam, S., M. Ptak, and S. Zhu. 2020. Modelling of daily lake surface water temperature from air temperature: Extremely randomized trees (ERT) versus Air2Water, MARS, M5Tree, RF and MLPNN. J. Hydrol. **588**: 125130.

Hersbach, H., and D. Dee. 2016. ERA5 reanalysis is in production. ECMWF Newsl. **147**: 5–6.

Hirsch, R. M., and L. A. De Cicco. 2015, *User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data*. US Geological Survey.

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

Hollister, J. 2021. Access elevation data from various APIs.

Hulley, G. C., S. J. Hook, and P. Schneider. 2011. Optimized split-window coefficients for deriving surface temperatures from inland water bodies. Remote Sens. Environ. **115**: 3758–3769. doi:10.1016/j.rse.2011.09.014

Japkowicz, N., and S. Stephen. 2002. The class imbalance problem: A systematic study1. Intell. Data Anal. **6**: 429–449. doi:10.3233/IDA-2002-6504

Jia, X., A. Karpatne, J. Willard, M. Steinbach, J. Read, P. C. Hanson, H. A. Dugan, and V. Kumar. 2018. Physics guided recurrent neural networks for modeling dynamical systems: Application to monitoring water temperature and quality in lakes. ArXiv Prepr. ArXiv181002880.

Jia, X., J. Willard, A. Karpatne, J. Read, J. Zwart, M. Steinbach, and V. Kumar. 2019. Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. Proceedings of the 2019 SIAM International Conference on Data Mining. SIAM, pp. 558–566.

Kingma, D. P., and J. Ba. 2017. Adam: A method for stochastic cptimization. ArXiv14126980 Cs.

Kourzeneva, E. 2010. External data for lake parameterization in numerical weather prediction and climate modeling.

Kraemer, B. M., and others. 2015. Morphometry and average temperature affect lake stratification responses to climate change: Lake stratification responses to climate. Geophys. Res. Lett. **42**: 4981–4988. doi:10.1002/2015GL064097

Kraemer, B. M., and others. 2017. Global patterns in lake ecosystem responses to warming based on the temperature dependence of metabolism. Global Change Biol. **23**: 1881–1890.

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

Kreakie, B. J., S. D. Shivers, J. W. Hollister, and W. B. Milstead. 2021. Predictive model of lake photic zone temperature across the conterminous United States. Front. Environ. Sci. **9**: 707874. doi:10.3389/fenvs.2021.707874

Larkin, S. L., and C. M. Adams. 2013. Economic consequences of harmful algal blooms: Literature summary. EDIS **2013**. doi:10.32473/edis-fe936-2013

Livingstone, D. M. 2008. A change of climate provokes a change of paradigm: Taking leave of two tacit assumptions about physical lake forcing. Int. Rev. Hydrobiol. **93**: 404–414.

Lloyd, S. 1982. Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**: 129–137. doi:10.1109/TIT.1982.1056489

Lottig, N. R., and others. 2017. Macroscale patterns of synchrony identify complex relationships among spatial and temporal ecosystem drivers. Ecosphere **8**: e02024.

Massie, D. L., G. J. A. Hansen, Y. Li, G. G. Sass, and T. Wagner. 2021. Do lake-specific characteristics mediate the temporal relationship between walleye growth and warming water temperatures? Can. J. Fish. Aquat. Sci. **78**: 913–923. doi:10.1139/cjfas-2020-0169

McDonald, C. P., E. G. Stets, R. G. Striegl, and D. Butman. 2013. Inorganic carbon loading as a primary driver of dissolved carbon dioxide concentrations in the lakes and reservoirs of the contiguous United States. Global Biogeochem. Cycles **27**: 285–295.

Mendonça, R., R. A. Müller, D. Clow, C. Verpoorter, P. Raymond, L. J. Tranvik, and S. Sobek. 2017. Organic carbon burial in global lakes and reservoirs. Nat. Commun. **8**: 1–7.

Messager, M. L., B. Lehner, G. Grill, I. Nedeva, and O. Schmitt. 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat. Commun. **7**: 1–11.

Meyer, M. F., S. G. Labou, A. N. Cramer, M. R. Brousil, and B. T. Luff. 2020. The global lake area, climate, and population dataset. Sci. Data **7**: 1–12.

Mironov, D., E. Heise, E. Kourzeneva, B. Ritter, N. Schneider, and A. Terzhevik. 2010. Implementation of the lake parameterisation scheme FLake into the numerical weather prediction model COSMO. Boreal Env. Res. **15**: 13.

Mitchell, K. E., et al. 2004. The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. J. Geophys. Res. Atmos. **109**: D07S90.

Moore, R. B. and others. 2019. User's guide for the National Hydrography Dataset plus (NHDPlus) high resolution. Open-File Report. US Geological Survey.

Muñoz-Sabater, J., and others. 2021. ERA5-land: A state-of-the-art global reanalysis dataset for land applications. Earth Syst. Sci. Data **13**: 4349–4383.

Oleksy, I. A., J. S. Baron, and W. S. Beck. 2021. Nutrients and warming alter mountain lake benthic algal structure and function. Freshw. Sci. **40**: 88–102. doi:10.1086/713068

O'Reilly, C. M., and others. 2015. Rapid and highly variable warming of lake surface waters around the globe: Global lake surface warming. Geophys. Res. Lett. **42**: 10773–10781. doi:10.1002/2015GL066235

Ostroff, A., D. Wieferich, A. Cooper, and D. Infante. 2013. Data release: National Anthropogenic Barrier Dataset (NABD) 2012 U.S. Geological Survey data release.

Pan, T., J. Zhao, W. Wu, and J. Yang. 2020. Learning imbalanced datasets based on SMOTE and Gaussian distribution. Inform. Sci. **512**: 1214–1233.

Pebesma, E. 2018. Simple features for R: Standardized support for spatial vector data. R J. **10**: 439–446. doi:10.32614/RJ-2018-009

Piccolroaz, S. 2016. Prediction of lake surface temperature using the air2water model: Guidelines, challenges, and future perspectives. Adv. Oceanogr. Limnol. **7**.

Piccolroaz, S., N. C. Healey, J. D. Lenters, S. G. Schladow, S. J. Hook, G. B. Sahoo, and M. Toffolon. 2018. On the predictability of lake surface temperature using air temperature in a changing climate: A case study for Lake Tahoe (U.S.A.): On the predictability of lake surface temperature. Limnol. Oceanogr. **63**: 243–261. doi:10.1002/lno.10626

Piccolroaz, S., M. Toffolon, and B. Majone. 2013. A simple lumped model to convert air temperature into surface

water temperature in lakes. Hydrol. Earth Syst. Sci. **17**: 3323–3338.

Pollard, A. I., S. E. Hampton, and D. M. Leech. 2018. The promise and potential of continental-scale limnology using the US Environmental Protection Agency's National Lakes Assessment. Limnol. Oceanogr. Bull. **27**: 36–41.

Raymond, P. A., and others. 2013. Global carbon dioxide emissions from inland waters. Nature **503**: 355–359. doi:10.1038/nature12760

Read, E. K., and others. 2017. Water quality data for national-scale aquatic research: The Water Quality Portal. Water Resour. Res. **53**: 1735–1745. doi:10.1002/2016WR019993

Read, J. S., and others. 2021. Data release: Process-based predictions of lake water temperature in the Midwest US: U.S. Geological Survey data release.

Read, J. S., and others. 2019. Process-guided deep learning predictions of lake water temperature. Water Resour. Res. **55**: 9173–9190.

Rew, R., and G. Davis. 1990. NetCDF: An interface for scientific data access. IEEE Comput. Graph. Appl. **10**: 76–82.

Rose, K. C., L. A. Winslow, J. S. Read, and G. J. Hansen. 2016. Climate-induced warming of lakes can be either amplified or suppressed by trends in water clarity. Limnol. Oceanogr.: Letters **1**: 44–53.

Sabater, M. 2019. ERA5-Land hourly data from 1981 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi:10.24381/cds.e2161bac

Schaeffer, B. A., J. Iiames, J. Dwyer, E. Urquhart, W. Salls, J. Rover, and B. Seegers. 2018. An initial validation of Landsat 5 and 7 derived surface water temperature for U.S. lakes, reservoirs, and estuaries. Int. J. Remote Sens. **39**: 7789–7805. doi:10.1080/01431161.2018.1471545

Schmid, M., and J. Read. 2021. Heat budget of lakes. *In Reference module in earth systems and environmental sciences*. Elsevier. doi:10.1016/B978-0-12-819166-8.00011-6

Sharma, S., and others. 2015. A global database of lake surface temperatures collected by in situ and satellite methods from 1985–2009. Sci. Data **2**: 150008. doi:10.1038/sdata.2015.8

Sharma, S., S. C. Walker, and D. A. Jackson. 2008. Empirical modelling of lake water-temperature relationships: A comparison of approaches. Freshw. Biol. **53**: 897–911.

Shen, C. 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resour. Res. **54**: 8558–8593.

Sit, M., B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and I. Demir. 2020. A comprehensive review of deep learning applications in hydrology and water resources. Water Sci. Technol. **82**: 2635–2670. doi:10.2166/wst.2020.369

Sola, J., and J. Sevilla. 1997. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Trans. Nucl. Sci. **44**: 1464–1468. doi:10.1109/23.589532

Soranno, P. A., and others. 2017. LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. GigaScience **6**: gix101.

Spacio, U. G., A. Meteorologia, G. Wodnej, A. R. per la Protezione, A. dell Emilia-Romagna, S. I. Meteo, C. I. Ricerche, and A. für GeoInformationswesen. 2008. Parameterization of lakes in numerical weather prediction. Description of a lake model.

Sprague, L. A., G. P. Oelsner, and D. M. Argue. 2017. Challenges with secondary use of multi-source water-quality data in the United States. Water Res. **110**: 252–261.

Stachelek, J., and others. 2021. LAGOS-US DEPTH v1.0: Data module of observed maximum and mean lake depths for a subset of lakes in the conterminous U.S. EDI Data Portal. doi:10.6073/pasta/64ddc4d04661d9aef4bd702dc5d8984f

Sun, A. Y., and B. R. Scanlon. 2019. How can Big Data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. Environ. Res. Lett. **14**: 073001. doi:10.1088/1748-9326/ab1b7d

Tibshirani, R., J. Friedman, and T. Hastie. 2009, *The elements of statistical learning: Data mining, inference, and prediction*. Springer series in statistics, 2nd Edition. Springer.

Toffolon, M., S. Piccolroaz, B. Majone, A.-M. Soja, F. Peeters, M. Schmid, and A. Wüest. 2014. Prediction of surface temperature in lakes with different morphology using air temperature. Limnol. Oceanogr. **59**: 2185–2202.

Tromans, D. 1998. Temperature and pressure dependent solubility of oxygen in water: A thermodynamic analysis. Hydrometallurgy **48**: 327–342.

Vanhellemont, Q. 2020. Automated water surface temperature retrieval from Landsat 8/TIRS. Remote Sens. Environ. **237**: 111518. doi:10.1016/j.rse.2019.111518

Willard, J. D., J. S. Read, S. Topp, G. J. A. Hansen, and V. Kumar. 2022. Daily surface temperature predictions for 185,549 U.S. lakes with associated observations and meteorological conditions (1980-2020): U.S. Geological Survey data release. doi:10.5066/P9CEMS0M

Winslow, L. A., S. Chamberlain, A. P. Appling, and J. S. Read. 2016. sbtools: A package connecting R to cloud-based data for collaborative online research. R J. **8**: 387–398.

Winslow, L. A., G. J. A. Hansen, J. S. Read, and M. Notaro. 2017. Large-scale modeled contemporary and future water temperature estimates for 10774 Midwestern U.S. lakes. Sci. Data **4**: 170053. doi:10.1038/sdata.2017.53

Winslow, L. A., J. S. Read, G. J. A. Hansen, and P. C. Hanson. 2015. Small lakes show muted climate change signal in deepwater temperatures. Geophys. Res. Lett. **42**: 355–361. doi:10.1002/2014GL062325

Woolway, R. I., and others. 2016. Diel surface temperature range scales with lake size. PLoS One **11**: e0152466. doi:10.1371/journal.pone.0152466

Woolway, R. I., B. M. Kraemer, J. D. Lenters, C. J. Merchant, C. M. O'Reilly, and S. Sharma. 2020. Global lake responses to climate change. Nat. Rev. Earth Environ. **1**: 388–403.

Woolway, R. I., and S. C. Maberly. 2020. Climate velocity in inland standing waters. Nat. Clim. Change **10**: 1124–1129.

Wynne, T. T., R. P. Stumpf, M. C. Tomlinson, G. L. Fahnenstiel, J. Dyble, D. J. Schwab, and S. J. Joshi. 2013. Evolution of a cyanobacterial bloom forecast system in western Lake Erie: Development and initial evaluation. J. Great Lakes Res. **39**: 90–99.

Acknowledgments

Jeff Holister helped with elevation data retrieval. Samantha Oliver helped assemble midwestern temperature data. R. Iestyn Woolway helped assemble resources to consider for dataset comparison. David Blodgett helped design file formats for compliance with existing standards. Shad Mahlum assisted with error estimation. Ellen Bechtel designed Fig. 1. Kathy Webster helped us better understand lake filtering done in compiling the LAGOS-US spatial dataset, which provides access to additional data for most of the lakes in this dataset. The Minnesota Supercomputing Institute (MSI) at the University of Minnesota provided resources that contributed to the research results reported within this paper. Jared Smith and Evan Goldstein provided early review of this manuscript and we thank *L&O:L* journal staff and two anonymous reviewers that helped improve the paper and underlying dataset. We thank data contributors who have made future research more efficient by sharing their data through prior data releases or the Water Quality Portal. This work was sponsored by the Department of the Interior United States Geological Survey Midwest Climate Adaptation Science Center, and also NSF grant 1934721 under the Harnessing the Data Revolution (HDR) program under PI Vipin Kumar. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Conflict of Interest

The authors declare no conflicts of interest.

Submitted 24 September 2021
Revised 23 February 2022
Accepted 28 February 2022