Received: 30 April 2024 | Revised: 8 October 2024 | Accepted: 11 December 2024

DOI: 10.1002/eap.70004

ARTICLE

ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA

What can we learn from 100,000 freshwater forecasts? A synthesis from the NEON Ecological Forecasting Challenge

Freya Olsson[1,2] | Cayelan C. Carey[1,2] | Carl Boettiger[3] | Gregory Harrison[2] | Robert Ladwig[4,5] | Marcus F. Lapeyrolerie[3] | Abigail S. L. Lewis[1] | Mary E. Lofton[1,2] | Felipe Montealegre-Mora[3] | Joseph S. Rabaey[6] | Caleb J. Robbins[7,8] | Xiao Yang[9] | R. Quinn Thomas[1,2,10]

[1]Department of Biological Sciences, Virginia Tech, Virginia, USA

[2]Center for Ecosystem Forecasting, Virginia Tech, Virginia, USA

[3]Department of Environmental Science, Policy, and Management, University of California Berkeley, California, USA

[4]Department of Ecoscience, Aarhus University, Aarhus, Denmark

[5]Center for Limnology, University of Wisconsin-Madison, Madison, Wisconsin, USA

[6]Large Lakes Observatory, University of Minnesota, Duluth, Minnesota, USA

[7]Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, Alaska, USA

[8]Center for Reservoir and Aquatic Systems Research, Baylor University, Waco, Texas, USA

[9]Department of Earth Sciences, Southern Methodist University, Dallas, Texas, USA

[10]Department of Forest Resources and Environmental Conservation, Virginia Tech, Virginia, USA

Correspondence
Freya Olsson
Email: freyao@vt.edu

Funding information
Division of Environmental Biology, Grant/Award Numbers: 1753639, 1926050, 1926388; Office of Advanced Cyberinfrastructure, Grant/Award Number: 2209866; Division of Graduate Education, Grant/Award Number: 1840995; Division of Biological Infrastructure, Grant/Award Number: 1933016

Handling Editor: Brice X. Semmens

Abstract

Near-term, iterative ecological forecasts can be used to help understand and proactively manage ecosystems. To date, more forecasts have been developed for aquatic ecosystems than other ecosystems worldwide, likely motivated by the pressing need to conserve these essential and threatened ecosystems and increasing the availability of high-frequency data. Forecasters have implemented many different modeling approaches to forecast freshwater variables, which have demonstrated promise at individual sites. However, a comprehensive analysis of the performance of varying forecast models across multiple sites is needed to understand broader controls on forecast performance. Forecasting challenges (i.e., community-scale efforts to generate forecasts while also developing shared software, training materials, and best practices) present a useful platform for bridging this gap to evaluate how a range of modeling methods perform across axes of space, time, and ecological systems. Here, we analyzed forecasts from the aquatics theme of the National Ecological Observatory Network (NEON) Forecasting Challenge hosted by the Ecological Forecasting Initiative. Over 100,000 probabilistic forecasts of water temperature and dissolved oxygen concentration for 1–30 days ahead across seven NEON-monitored lakes were submitted in 2023. We assessed how forecast performance varied among models with different structures, covariates, and sources of uncertainty relative to baseline null models. A similar proportion of forecast models were skillful across both variables (34%–40%), although more individual models outperformed the baseline models in forecasting water temperature (10 models out of 29) than dissolved oxygen (6 models out of 15). These top performing models came from a range of classes and structures. For water temperature, we found that forecast skill degraded with increases in forecast horizons, process-based models, and models that included air temperature as a covariate generally exhibited the highest forecast performance, and that the most skillful forecasts often accounted for more sources of uncertainty than the lower performing models. The most skillful forecasts were for sites where observations were most divergent from historical conditions (resulting in poor baseline model performance). Overall, the NEON Forecasting Challenge provides an exciting opportunity for a model intercomparison to learn about the relative strengths of a diverse suite of models and advance our understanding of freshwater ecosystem predictability.

KEYWORDS
ecological forecasting, forecasting challenge, freshwater, near-term forecast, NEON, uncertainty, water quality

INTRODUCTION

Ecological forecasting is a growing field that leverages predictions of future ecological states to help understand and manage ecosystems (Dietze et al., 2018; Lewis et al., 2023; Tulloch et al., 2020). Here, we define forecasts as predictions of future conditions with specified uncertainty (Lewis et al., 2022). As environmental conditions increasingly change in response to altered climate and land use (Arias et al., 2021), ecological forecasts have considerable potential for improving management to support ecosystem services now and in the future (Bradford et al., 2018; Dietze et al., 2018). Moreover, forecasting future conditions that have yet to occur inherently requires out-of-sample implementation of models, which can lead to insights into optimal modeling approaches (Lewis et al., 2023).

In freshwater ecosystems, rapid environmental change has led to conditions that are both more variable and outside of historically observed states, motivating a particular need for near-term, iterative ecological forecasts (e.g., Carey, 2023; Richardson et al., 2024; Siam & Eltahir, 2017). Near-term (i.e., subdaily to decadal) forecasts allow researchers to evaluate models within management-relevant timescales (Dietze et al., 2018), and iteratively updating and evaluating forecasts enables rapid improvement in forecast performance by integrating observational data and updating parameters (Dietze et al., 2018; Loescher et al., 2017). These near-term iterative ecological forecasts will help protect critical provisioning, regulating, supporting, and cultural services (Dodds et al., 2013; Lofton et al., 2023; Sterner et al., 2020) that these highly threatened systems provide (Carrizo et al., 2017; Dudgeon et al., 2006; Reid et al., 2019), thereby improving management and mitigation (e.g., Carey et al., 2022; Huang et al., 2011; Zwart et al., 2023).

Although the number of near-term, iterative water quality forecasts of freshwater ecosystems is growing (Lofton et al., 2023), challenges remain in producing reliable and accurate predictions of changes in these environments. To date, researchers have implemented many classes of models to forecast freshwater variables (reviewed by Lofton et al., 2023), including process-based (PB) models (Baracchini et al., 2020; Clayer et al., 2023; Page et al., 2018; Thomas et al., 2020), machine learning (ML) models (Cheng et al., 2020; Di Nunno et al., 2023; Read et al., 2019; Zwart et al., 2023), statistical models (Caissie et al., 2017; McClure et al., 2021; Woelmer et al., 2021), and multimodel and hybrid approaches (Olsson, Moore, et al., 2024; Qu et al., 2017; Saber et al., 2020). In addition, forecasts have been generated using a range of model covariates (i.e., driver variables). In many cases, weather forecasts are used as covariates because meteorology is a key driver of many ecosystem processes in freshwater ecosystems (Hipsey et al., 2019; Livingstone & Padisák, 2007; Rousso et al., 2020). Additionally, some models include autoregressive terms as covariates (e.g., ARIMA models). While forecasting methods have demonstrated promise at individual freshwater sites or a handful of sites (e.g., Barrachini et al., 2020; Chen et al., 2024; Ouellet-Proulx et al., 2017; Page et al., 2018; Thomas et al., 2020; Zwart et al., 2023), to date there has yet to be a comprehensive analysis of the performance of forecasting models across a large range of model classes and model covariates across multiple sites.

Forecasting challenges present a useful platform for bridging this gap and learning about how a range of modeling methods perform across axes of space, time, and ecological systems (Humphries et al., 2018; Thomas et al., 2023). Forecasting challenges typically entail an open call to the research community with a "challenge" to forecast a specific variable, standardized requirements, and formal evaluation of out-of-sample time steps. Some challenges have aimed to identify a "winner" or best approach, while others have focused more on community and knowledge building (Humphries et al., 2018; Makridakis et al., 2020; Thomas et al., 2023). By bringing together individuals and teams from broad backgrounds, challenges provide opportunities for innovation and community-building, and the development of community cyberinfrastructure can accelerate discipline-wide progress (Fer et al., 2021). Altogether, this collaborative effort can facilitate the development of new methods, standardization of forecasting targets and formats, and tools and templates that expand the training and education to improve accessibility of forecasting (Thomas et al., 2023). While forecasting challenges are common in the fields of finance, business, demography (Bojer & Meldgaard, 2021; Makridakis et al., 2020), and epidemiology (Biggerstaff et al., 2018; Johansson et al., 2019; Viboud et al., 2018), few have existed in ecology until recently (e.g., Humphries et al., 2018; Wheeler et al., 2024), providing new opportunities for advancing the discipline. For example, previous efforts to compare outcomes among ecological forecasting methods have been hindered by differences in evaluation metrics, sites, and variables being forecasted (e.g., Rousso et al., 2020), which can be addressed by a standardized forecasting challenge framework.

The National Ecological Observatory Network (NEON) Forecasting Challenge (hereafter NEON Challenge), hosted by the Ecological Forecasting Initiative (EFI) Research Coordination Network, was designed to initiate these advances in ecological forecasting. The NEON Challenge is "an open platform for the ecological and data science communities to forecast NEON data before they are collected" (Thomas et al., 2023). The challenge aims to galvanize the forecasting community around a common framework, with the goals of improving forecasting tools (e.g., Dietze et al., 2023), learning about ecological predictability (e.g., Wheeler et al., 2024), and advancing training (e.g., Willson et al., 2023).

The NEON Challenge provides a unique case study for examining the performance of freshwater forecasts across space, time, and ecological systems. Ecological time series present specific complexities compared with previous forecasting challenges given the variability in ecological data collection, irregularities in data resolution, and the inherent variability of the observations (Farley et al., 2018; Michener & Jones, 2012). Moreover, unlike previous forecasting challenges, the NEON Challenge is ongoing and accepts submissions of as-yet-unmeasured conditions on a rolling basis, with scoring occurring continuously as new data are collected and made available in near real time (Thomas et al., 2023). In the aquatics lake theme of the NEON Challenge, participants were invited to submit 1- to 30-day-ahead probabilistic forecasts of daily surface mean water temperature (hereafter, T_w) and dissolved oxygen concentration (DO) of seven NEON lake sites, with new forecasts accepted daily (Thomas et al., 2023). Due to issues relating to data quality, submitted forecasts of chlorophyll *a* were omitted from our analysis. Forecasts were solicited across a range of sites, dates, and variables to understand how skill varies across these three axes. Forecasts could be generated using any method but had to include an estimate of uncertainty.

The inclusion of, and emphasis on, uncertainty was a novel component of the NEON Challenge, as uncertainty has been rarely included in previous forecasting challenges. Meaningful representations of uncertainty are critical to forecast interpretation and comparison, but uncertainty quantification is still not ubiquitous across ecological forecasts (reviewed by Lewis et al., 2022), and freshwater forecasts in particular. In a review of freshwater forecasts by Lofton et al. (2023), only 16 out of 61 near-term (subdaily to decadal) forecasts of water quality variables included an estimate of the uncertainty associated with a prediction. Uncertainty can arise from a variety of sources: model process, model parameters, model initial conditions, model drivers, and observations (Table 1). The relative importance of each source is often dependent on the ecosystem process or state being forecasted and the forecast horizon (Lofton et al., 2022; Ouellet-Proulx et al., 2017; Thomas et al., 2020). In addition, the predictability of an ecological process or state depends on the magnitude of the forecast spread (forecast uncertainty) and the rate at which uncertainty increases across the forecast horizon. Predictability is low when forecast spread is large enough that it cannot distinguish between consequential differences in ecosystem processes

TABLE 1 Definitions of forecast uncertainty sources included in the submitted models, modified from Dietze (2017), Lofton et al. (2023), and Thomas et al. (2020).

Source of uncertainty	Definition	Example of how the uncertainty source could be quantified
Process	Uncertainty from the inability of the model to replicate the dynamics of the forecasted state.	Calculating the error from the residuals of the model fit to historical data.
Parameter	Uncertainty in the parameter values of a fitted model.	Sampling from a distribution of parameter values and assigning different parameter values to each ensemble member.
Initial condition	Uncertainty in estimates of current conditions at the time of forecast generation (e.g., as a result of observation uncertainty, missing observations, and data assimilation).	Quantifying the spread in updated states following data assimilation or the previous day's forecast.
Driver	Uncertainty from driver data (e.g., future air temperature).	Using an ensemble of weather forecasts as drivers to the model.
Observation	Uncertainty from measurement error in the state being forecasted (difference between actual state and measured state).	Calculating the standard deviation of replicate water temperature observations.

or states, or when it is no different from random chance. Forecast spread in turn depends on the sources of uncertainty in the forecast model (Dietze, 2017) and the model sensitivity to these sources. For example, the predictability of ecosystem processes that are sensitive to meteorological drivers (e.g., air temperature) depends on the uncertainty in the weather forecasts used as inputs to the ecological forecast model (Dietze, 2017).

We were specifically focused on uncertainty in our analysis because forecasts that include well-quantified uncertainty, in addition to being accurate, have been shown to improve decision-making outcomes (Mylne, 2002; Nadav-Greenberg & Joslyn, 2009; Ramos et al., 2013). NEON forecast submissions were thus evaluated in two ways that captured different attributes of accuracy and precision: the continuous rank probability score (CRPS), a CRPS comparison with a baseline (null) model that acted as a benchmark to assess relative gains in forecast performance (forecast skill; Murphy, 1992; Pappenberger et al., 2015), and an evaluation of how well the forecast CIs capture the observation (CI reliability; e.g., if 90% of the observations in the 90% forecast CI).

In this study, we analyzed a year of submissions to the aquatics theme of the NEON Challenge and assessed how model performance varied among model class, model covariates, and forecast sites. We used the forecast analysis to answer the following research questions: Q1: How does model class and inclusion of covariates affect forecast performance? Q2: To what extent is relative forecast skill affected by the inclusion of different sources of uncertainty? Q3: How consistent are the patterns in forecast performance across sites? We included all T_w and DO forecasts in the analysis of Q1 but focused primarily on T_w forecasts for Q2 and Q3 due to the much higher number of submissions for that variable (see below). To the best of our knowledge, our study is the first analysis that investigates the performance of freshwater forecasts across multiple model classes, model covariates, and sites using genuine forecasts of the future.

METHODS

NEON challenge overview

The NEON Challenge has five forecasting themes that cover a range of ecological populations, communities, and ecosystems across the NEON network of monitored freshwater and terrestrial sites. Our coauthor team represents a group of the Challenge organizers, cyberinfrastructure developers, and/or forecast submitters.

Submissions were accepted to the aquatics theme of the NEON Challenge starting in 2021 and continuing to the present (>3 years) for forecasts of water quality. Here, we focus on the forecasts of T_w and DO submitted to lake sites within the aquatics theme of the NEON Challenge during 2023, which represented the first full year with sufficient submissions for a robust intermodel comparison.

Challenge design

NEON data

Water quality data were collected at seven lakes across the United States (Figure 1). T_w and DO were collected

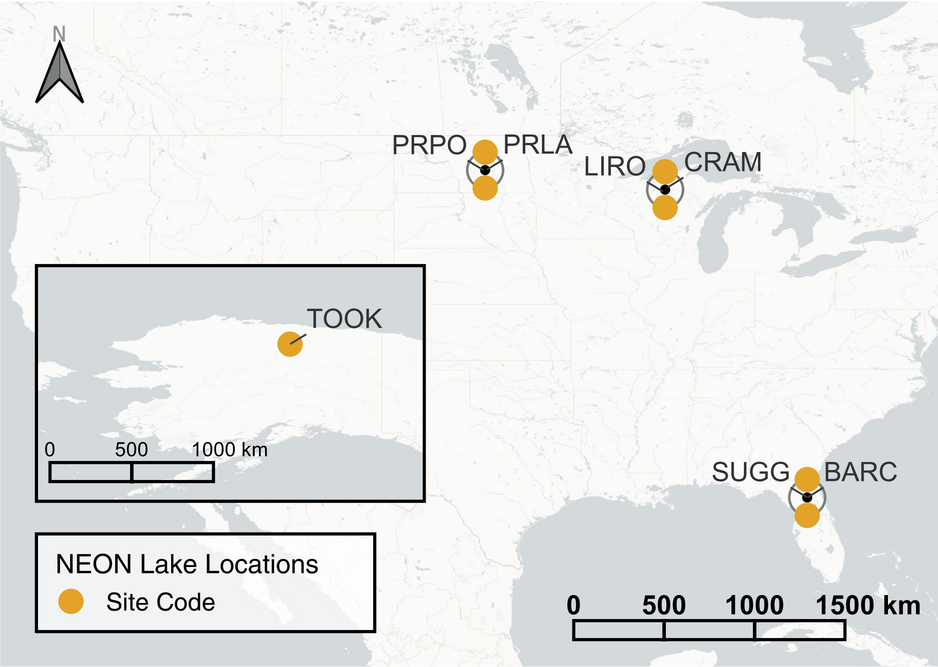

FIGURE 1 Map of National Ecological Observatory Network (NEON) lake sites located across the contiguous United States, with map inset showing Alaska. Co-occurring sites are shown by the black centroid and the colored points are offset from this location. The points are labeled with their four-character NEON site code: BARC, Barco Lake; CRAM, Crampton Lake; LIRO, Little Rock Lake; PRLA, Prairie Lake; PRPO, Prairie Pothole Lake; SUGG, Suggs Lake; TOOK, Toolik Lake.

using in situ sensors. Full descriptions of the sensors and protocol are included in the data product metadata provided by NEON (DP1.20264.001, NEON TSD) for T_w and DP1.20288.001 for DO (NEON water quality). At each lake, data were only available at one location (generally at the center, near the deepest point). For the purposes of the Challenge, unpublished data were made available to participants by NEON at a data latency of 2–3 days after collection. The T_w and DO NEON data products extend back to 2016, but their temporal coverage varies across sites in three ways. First, there is variability in the duration of time-series data available for each site and variable (Appendix S1: Figure S1), ranging from 3.1 to 6.6 years (up to 1 January 2023, the beginning of our focal forecasting period). Second, at five lake sites, sensors are removed during winter due to ice formation. Finally, maintenance issues resulted in data gaps at some sites. Consequently, total data availability varied between 167 and 2154 days for each site/variable combination (Appendix S1: Figure S1).

Data processing and targets generation

We, as challenge organizers, converted the T_w and DO data supplied by NEON in near-real time to “targets”—observations specific to the challenge—by subsetting the sensor locations, performing additional quality control, and aggregating 30-min sensor data to daily means. We used daily mean temperature to focus the Challenge on predicting day-to-week dynamics in water quality, rather than subdaily dynamics. The data were subset to include only the surface measurements (top 1 m of the water column). Using only surface measurements, rather than full water column profiles, enabled intercomparison across the seven lakes, which had varying maximum depths that ranged from 3.2 to 27 m. Second, we filtered the data using the existing NEON flags (see metadata) and applied additional quality control measures (e.g., additional filtering for maximum and minimum allowable values for each variable; see Olsson, Carey, et al., 2024b). The targets data could then be used by teams to calibrate and train models and were used for forecast evaluation.

These processed target data were publicly available to all Challenge teams at a persistent URL location and were updated daily as new data became available. To further support modeling efforts by the teams, we also provided supplementary hourly water temperature profile data collected by NEON at each of the lake sites (derived from NEON DP1.20264.001, see Olsson, Carey, et al., 2024b). These supplemental data were available to teams to use in model development and training but were not used in forecast evaluation.

Ancillary driver data

NOAA's Global Ensemble Forecasting System (GEFS; Hamill et al., 2022) weather forecast data were made available to forecast teams via functions in the custom R package *neon4cast* (Boettiger & Thomas, 2024). NOAA weather data for all NEON sites were downloaded each day and standardized to be used as driver data and covariates in forecast models. Teams were not required to use weather covariates, but providing standardized NOAA weather forecasts ensured that the teams that used weather covariates had consistent data, and weather forecast performance was therefore not the primary driver of differences in aquatic forecast performance among model submissions. Two NOAA data products were used by forecast teams: an ensemble forecast of future weather and a historic weather product. The ensemble weather forecast consisted of 31 ensemble members up to 35 days into the future at each of the seven sites. The historic product consisted of stacked 1-day-ahead forecasts from each day as an estimate of observed historical conditions that was consistent with the ensemble weather forecast data available to teams to forecast (i.e., having similar biases, compared with observational weather data) and could be used to calibrate models. Teams were also able to use any other openly-available covariate data in their forecasts, although none chose to do this.

Forecast submission guidelines

Challenge teams were invited to forecast T_w and DO in all of the lakes or in any subset of sites or variables. Forecast submissions were required to have a daily time step of the focal variable(s) over a forecast horizon of at least 1–30 days into the future and include an estimate of uncertainty in the forecast. Uncertainty could be represented by submitting a probabilistic forecast (Gneiting & Katzfuss, 2014), either in the form of a mean and a SD for a normally distributed forecast or as an ensemble forecast for which the uncertainty was represented as a series of predictions that represent a range of future conditions (Gneiting & Katzfuss, 2014). Submissions were required to follow a standardized format (Dietze et al., 2023; Thomas et al., 2023) to enable automated evaluation and processing. New forecasts were accepted every day and evaluated as new observational data became available (see Forecast evaluation).

During 2022 and 2023, we ran multiple workshops to introduce the Challenge to a cross-section of aquatic and data scientists and managers to increase forecast submissions to this theme (Meyer et al., 2023; Olsson, Boettiger et al., 2024). In total, more than 300 people attended the workshops in person or online. Workshop materials were also available online for individuals or groups to use independently (Olsson, Boettiger et al., 2024).

Baseline model

Following forecast evaluation best practices (Harris et al., 2018; Lewis et al., 2022), we generated a baseline model that represents a limited (naive) understanding of the system for comparison with the submitted forecast models. It can be helpful to compare submitted forecasts with forecasts generated from baseline models as part of forecast evaluation to identify whether new methods provide additional, useful information beyond uninformed models (Jolliffe & Stephenson, 2012; Makridakis et al., 2020; Pappenberger et al., 2015). Specifically, we generated a model that assumes the forecast for a particular day-of-year (DOY) is equal to the mean of historical data on that DOY. The DOY baseline model assumes dynamics will follow the mean conditions for that date in previously observed years (Hyndman & Athanasopoulos, 2021; Jolliffe & Stephenson, 2012). The uncertainty in this DOY forecast was generated by calculating the SD of the past observations (see Appendix S1: Text S1). The SD of the daily average for the forecast period was used to represent the uncertainty for the whole horizon. The DOY forecast was assumed to follow a normal distribution, given by a mean and SD for each day of year calculated separately for each site and variable.

The baseline model was selected based on the observed dynamics of the variable of interest (Jolliffe & Stephenson, 2012; Pappenberger et al., 2015) as well as being a common baseline for ecological forecasts (e.g., Lewis et al., 2022; Thomas et al., 2020; Wheeler et al., 2024). The DOY model is particularly useful as a baseline when the target variable's dynamics follow a seasonal cycle (Pappenberger et al., 2015), such as variables primarily driven by meteorological forcing. A second baseline model that assumes a forecast is equal to the last observation (persistence; Jolliffe & Stephenson, 2012) was also included in submissions.

Forecast evaluation

Initially, forecasts were evaluated against observations using the CRPS, as implemented in the *scoringRules* R package (Jordan et al., 2019). CRPS evaluates the probability distribution of the forecast and assesses both the accuracy and precision of the forecast relative to observations and is calculated as follows:

$$\mathrm{CRPS} = \int (F(y) - H(y - y_{\mathrm{obs}}))^2 dy \quad (1)$$

where y is the value for the forecasted variable, y_{obs} is the observation, and $F(y)$ is the cumulative distribution function of the probabilistic forecast at the value of y. H is the indicator or step function, which is zero if $y < y_{\mathrm{obs}}$ and one otherwise (Jordan et al., 2019). CRPS is a generalization of mean absolute error for probabilistic forecasts and is expressed in the same units as the variable, and ranges from an optimal value of zero to infinity (Pappenberger et al., 2015). In addition, we used a *relative forecast skill* (hereafter, $\mathrm{CRPS}_{\mathrm{skill}}$ or skill) metric to describe how much additional information is gained in each model over a naive baseline model. $\mathrm{CRPS}_{\mathrm{skill}}$ was calculated based on the difference in CRPS score between the submitted forecast and the DOY baseline model, following Equation (2):

$$\mathrm{CRPS}_{\mathrm{Skill}} = \mathrm{forecast\_score} - \mathrm{DOY\_score} \quad (2)$$

with positive values indicating a submitted forecast showing lower skill and higher error, relative to the DOY model, and negative values indicating that the submitted model performed better with lower error rates, as quantified using CRPS. We used this convention to ease comparison with other papers that synthesized submissions to the NEON Ecological Forecasting Challenge (e.g., Wheeler et al., 2024). We opted to focus on $\mathrm{CRPS}_{\mathrm{skill}}$ relative to the DOY model rather than the persistence baseline model as the DOY model had lower average CRPS and was the better performing of the baseline models for more of the 30-day-ahead forecast horizons (27 out of the 30 days; Appendix S1: Figure S2).

Analyses

We assessed the performance of the forecast models across different horizons and sites by aggregating raw $\mathrm{CRPS}_{\mathrm{skill}}$ metrics at different temporal and spatial scales. To identify the best performing models per variable, we calculated the mean $\mathrm{CRPS}_{\mathrm{skill}}$ aggregated across all forecast submission dates, horizons, and sites. To ensure that the comparisons among models were based on a similar number of submissions, we only included models in the analysis that had submissions for 80% of evaluated days (i.e., days with observations). We allowed teams to “catch-up” their forecasts (i.e., submit forecasts that were not “real time” but “retroactive forecasts” following Jolliffe & Stephenson, 2012) when they missed submissions due to any issues with automated cyberinfrastructure. Retroactive forecasts could only use target data and forecasted covariates that would have been available if the forecast was generated in real time (i.e., a retroactive forecast of water temperature for 1 July 2023 only used observations before this date for model training and was driven by NOAA weather forecasts generated on 30 June 2023 or earlier). No model was represented only by retroactive forecasts. In our analysis, we removed the 16-day-ahead horizon from evaluation because of processing issues when downloading NOAA weather forecasts. The 16-day-ahead horizon had an artificially low variance in the forecast that was not present in the other horizons due to an error in the post-processing of the weather forecast from the 6- to 1-h time resolution. The 1- to 16-day-ahead forecast becomes available for download from NOAA earlier than the 17- to 35-day-ahead forecast. When combining the two sets of forecasts and temporally downscaling to a 1-h time step, ensemble members were not matched correctly, resulting in reduced variance at the concatenation point. The processing issue was resolved during the period of evaluation, but we excluded the affected horizon, regardless, so that we could compare forecasts throughout all of 2023.

The reliability of the CIs was calculated by estimating the percentage of observations that fell within a specified CI. Reliability refers to the statistical agreement of forecast probabilities with observed relative frequencies of events (Gneiting et al., 2007; Schepen et al., 2016; see also *calibration* and *coverage*). A forecast that has perfectly reliable CIs will have the equivalent proportion of the observations falling within the CI (Jolliffe & Stephenson, 2012; Thomas et al., 2020): for example, 80% of observations falling within the 80% CI and 95% of observations falling within the 95% CI. “Underconfident” forecasts are represented by CIs that are too wide and result in more observations falling within them (e.g., 90% of observations falling within an 80% CI), whereas “overconfident” forecasts have CIs that are too narrow and fail to capture the observations (e.g., only 40% of observations falling inside an 80% CI) (following Ouellet-Proulx et al., 2017; Thomas et al., 2020; Zwart et al., 2023). We opted to look at the 80% and 95% CIs as the 80% CI covers the bulk of the forecast distribution, and the 95% CI shows the ability of the forecast to represent the values in the tails of the distribution.

RESULTS

Forecast inventory

Individuals and teams submitted a total of 100,475 daily forecasts for 1- to 30-day-ahead horizons using 28 different models (Olsson, Carey, et al., 2024a) to the aquatics lake theme of the NEON Challenge in 2023. Here, we define one forecast as a collection of predictions for 1–30 days in the future for a unique combination of forecast starting date, forecast site, forecasted variable, and forecasting model. The 28 models were used in addition to the two baseline models (persistence and DOY models) submitted by Challenge organizers ($n = 30$ models total). The forecasted variables were unevenly represented in the submissions: 14 models (plus two baselines) were used to submit forecasts for both variables (T_w, DO), 14 models were used to submit forecasts for only T_w, and no models submitted forecasts for only DO (total model submissions for each variable: $T_w = 30$, DO $= 16$). Across all submissions, forecasts of water temperature for the lake sites were the most numerous ($n = 63{,}189$; 63% of total lake forecasts) and had a greater diversity of model classes and covariates.

The 30 T_w models included a range of model classes and exogenous covariates. The self-reported model classes included empirical models (statistical and time series), ML, and PB models, as well as multimodel ensembles (MME; i.e., predictions were based on an aggregation of other model forecast submissions). Within the MMEs, forecasts were generated by combining process models, baseline and process models, empirical and baseline models, and an MME of the two baselines (Table 2; Appendix S1: Table S1). Forecast models included a range of exogenous covariates from the NOAA GEFS weather forecasts, with forecasted air temperature being the most commonly used covariate ($n = 19$; Appendix S1: Table S1). No other exogenous covariates (i.e., non-NOAA GEFS weather covariates) were included in any model. Details of all of the models that submitted forecasts in 2023 that met the criteria for inclusion in this analysis are provided in Appendix S1: Text S1 and Olsson, Carey, et al. (2024a).

The 16 DO models represented less diversity in model classes and covariates than the T_w models (Figure 2). The model classes for the DO models included only empirical and ML models (in addition to the baseline models), and air temperature was used as a covariate in six of the 16 DO models (38%).

How does model class affect forecast performance across all variables?

More T_w forecast models ($n = 10$) outperformed the DOY baseline than DO forecast models ($n = 6$; Figure 2). Only six of the submitted DO models outperformed the DOY baseline model across all forecast dates and sites (i.e., models had mean negative $\text{CRPS}_{\text{skill}}$, with a mean between −0.01 and −0.08 mg/L aggregated across the 1- to 30-day-ahead horizon; Figure 2c). These six highest performing DO models included both ML and empirical models, of which the highest performing models were ML models that used air temperature as a covariate (Random Forest, Lasso, and XGBoost). The models that did not outperform the baseline were all empirical, and no PB models were used to forecast DO in lakes.

Unlike DO, the best performing models for water temperature (T_w) were from the full range of model classes (Figure 2a,b). Of the 30 submitted models, 10 T_w forecast models outperformed the DOY baseline model when forecasts were aggregated across all sites and horizons for

TABLE 2 Representation of uncertainty within the best performing water temperature (T_w) models (sorted in descending order) that had negative mean $\text{CRPS}_{\text{skill}}$ (i.e., outperformed the day-of-year baseline) over the 1- to 30-day-ahead forecast horizon.

		Source of uncertainty represented				
Model	**Model classification**	**Driver**	**Parameter**	**Process**	**Initial conditions**	**Observation**
FLARE-GLM	PM	×	×	×	×	×
FLARE-GLM-noDA	PM	×	×	×	×	×
FLARE-GOTM	PM	×	×	×	×	×
XGBoost	ML	×		×		
Random Forest	ML	×				
LER-Baselines MME	MME (PB, Baseline)	×	×	×	×	×
FLARE-LER MME	MME (PB)	×	×	×	×	×
FLARE-GOTM-noDA	PM	×	×	×	×	×
Prophet	ML		×	×		
Lasso	ML	×				

Note: See Table 1 for definitions of uncertainty types. Model type classification as categorized by model teams: ML, machine learning model; MME, multimodel ensemble; PM, process model. For the MME model type, the constituent model class is shown in parentheses. For the comprehensive list of uncertainty sources for all submitted models and all variables, see Appendix S1: Table S1.

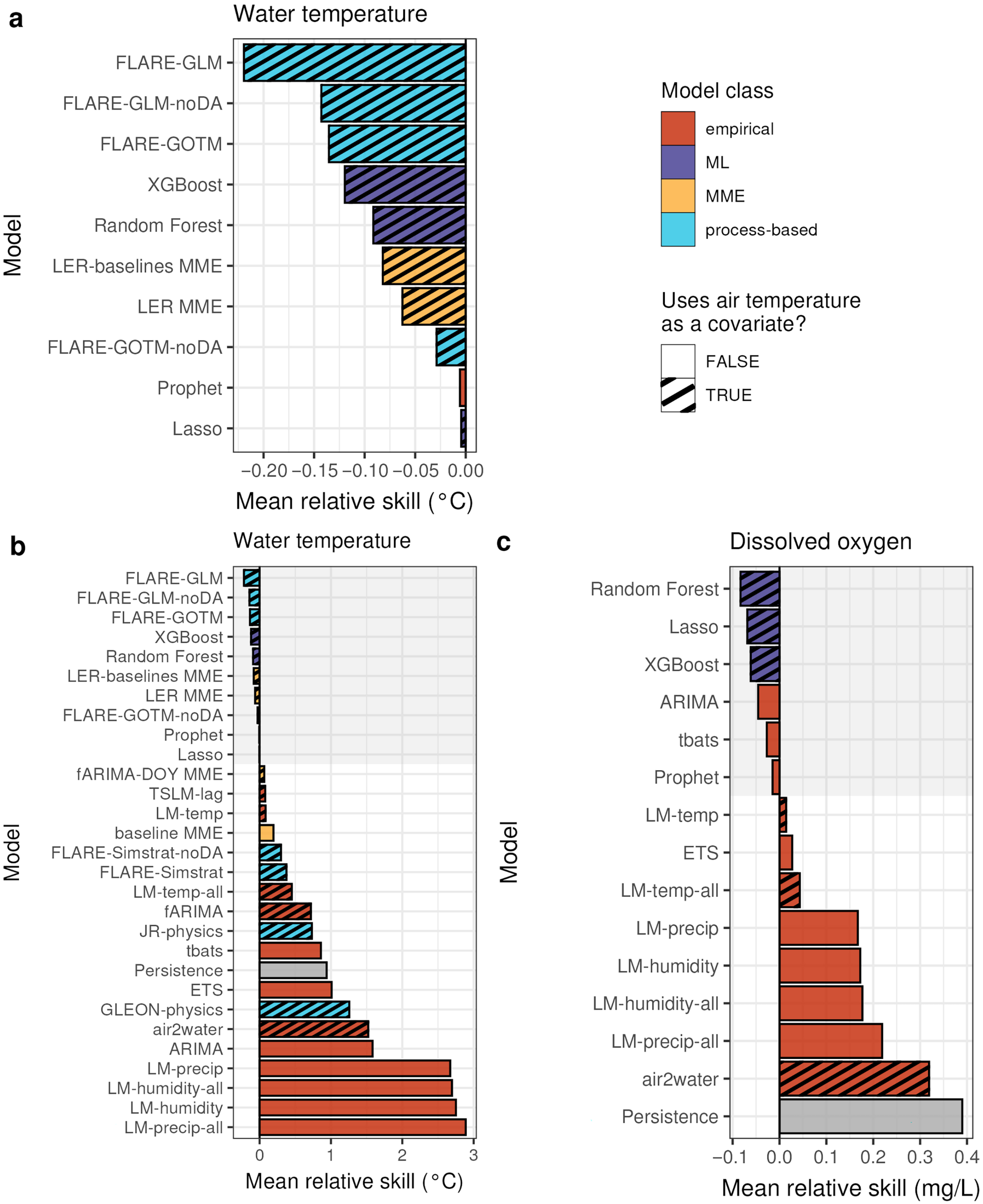

FIGURE 2 Legend on next page.

the year of forecasts. Across all sites and forecasts, a PB model had the best skill (Figure 2a), with a mean $\text{CRPS}_{\text{skill}}$ of −0.22°C aggregated across the 1- to 30-day-ahead horizon. Although the overall top three models were PB models, not all PB models were high performing, as four PB models had a positive mean $\text{CRPS}_{\text{skill}}$ (Figure 2b).

Altogether, of the different model classes used to submit forecasts of T_w, 4 of the 8 PB models, 1 of the 13 empirical models, 2 of the 4 MME, and all 3 of the ML models outperformed the baseline DOY model on average over the year (Figure 2a). Machine learning models accounted for three models in the top 10 T_w forecast models, as XGBoost, Random Forest, and Lasso models all had negative $\text{CRPS}_{\text{skill}}$. Empirical models exhibited the worst performance among the model classes, as only one (the Prophet model, Figure 2a) outperformed the DOY baseline model across all forecasts. Given the better performance of forecasts for T_w (10 models beating the baseline), as well as the higher diversity of model classes represented in these higher performing models ($n = 4$), further analyses for addressing Q1, Q2, and Q3 were conducted on the T_w forecasts only.

Among T_w models, how does model class and inclusion of covariates affect performance across the forecast horizon?

Nine out of the 10 T_w models that outperformed the baseline model included air temperature as a covariate (Figure 2a). The specific inclusion of air temperature as a covariate appeared to confer some skill, as it was not included in any of the five lowest performing models (Figure 2b). However, the inclusion of exogenous covariates did not guarantee high performance of a model, as 10 of the models exhibiting positive $\text{CRPS}_{\text{skill}}$ included air temperature as a covariate, as well as other NOAA weather covariates such as humidity and precipitation (Appendix S1: Text S1). There was only one model that outperformed the baseline model, the empirical Prophet model, which was based solely on observations and included no exogenous covariates (Figure 2a).

Focusing on T_w, $\text{CRPS}_{\text{skill}}$ in the most skillful forecasts generally worsened across the forecast horizon (Figure 3a), and, on average, were unable to outperform the DOY baseline at horizons of 15–25 days ahead. The exceptions to this pattern were the Lasso and Random Forest ML models, which showed improvement in skill for the first 7 or 8 days ahead and then decreases in skill at horizons longer than 8 days. Generally, the PB models and MME forecasts showed larger rates of degradation compared with the ML and empirical models (Figure 3a). The Prophet ML model exhibited the smallest degradation in skill (−0.16–0.08°C) across the 30 days, although its skill at 7- to 16-day-ahead horizons was the worst of any model that outperformed the baseline (Figure 3a). In comparison, two MME forecasts showed the largest rates of degradation (LER baselines MME and FLARE-LER MME), from high performance at short horizons (−0.58 and −0.64°C) to poor performance at the longest horizons (0.24 and 0.32°C). Only one model had negative $\text{CRPS}_{\text{skill}}$ across the full forecast horizon, the XGBoost ML model, which had a low rate of skill degradation across the 30 days (0.32°C; Figure 3a). The models that exhibited poorer skill throughout the 30-day forecast horizon generally showed consistently worsening performance into the future (Appendix S1: Figure S3), although the worst performing models had poor performance irrespective of forecast horizon.

Of the 10 models that outperformed the DOY model on average, 7 models also outperformed the persistence model at all forecast horizons. Only the empirical Prophet model and the ML Lasso and Random Forest models did not outperform the persistence model at all horizons; the persistence model was better performing during the first 3 days of the forecast (Appendix S1: Figure S4). The persistence model had its highest performance at the shortest horizons (1–3 days-ahead) and was the best performing baseline model at these horizons (Appendix S1: Figure S2).

Out of all T_w models that outperformed the DOY baseline (as determined by the aggregation of skill over the full forecast horizon; Figure 2a), XGBoost had negative $\text{CRPS}_{\text{skill}}$ for the full forecast horizon, outperforming the DOY and persistence models at all horizons

FIGURE 2 Mean relative skill ($\text{CRPS}_{\text{skill}}$, compared with day-of-year [DOY] baseline model) of water temperature (T_w) and dissolved oxygen (DO) forecasts for the submitted models (averaged across sites, submission dates, and 1- to 30-day-ahead horizons). Negative values indicate that a submitted model performed better, on average, than the DOY baseline and positive values indicate that the baseline performed better. (a) The T_w models that outperformed the DOY baseline as defined by $\text{CRPS}_{\text{skill}}$; (b) all T_w models; (c) $\text{CRPS}_{\text{skill}}$ for DO models. The shading of the bars indicates the model structure; color indicates model class (empirical, machine learning [ML], multimodel ensemble [MME], process), and pattern indicates the inclusion of air temperature as a covariate. A second baseline model (persistence) is shown in gray (b, c) and models that outperformed the DOY baseline are highlighted by the gray background shading. Constituent model classes of the multimodel ensemble models are given in Table 2 and Appendix S1: Table S1. CRPS, continuous rank probability score.

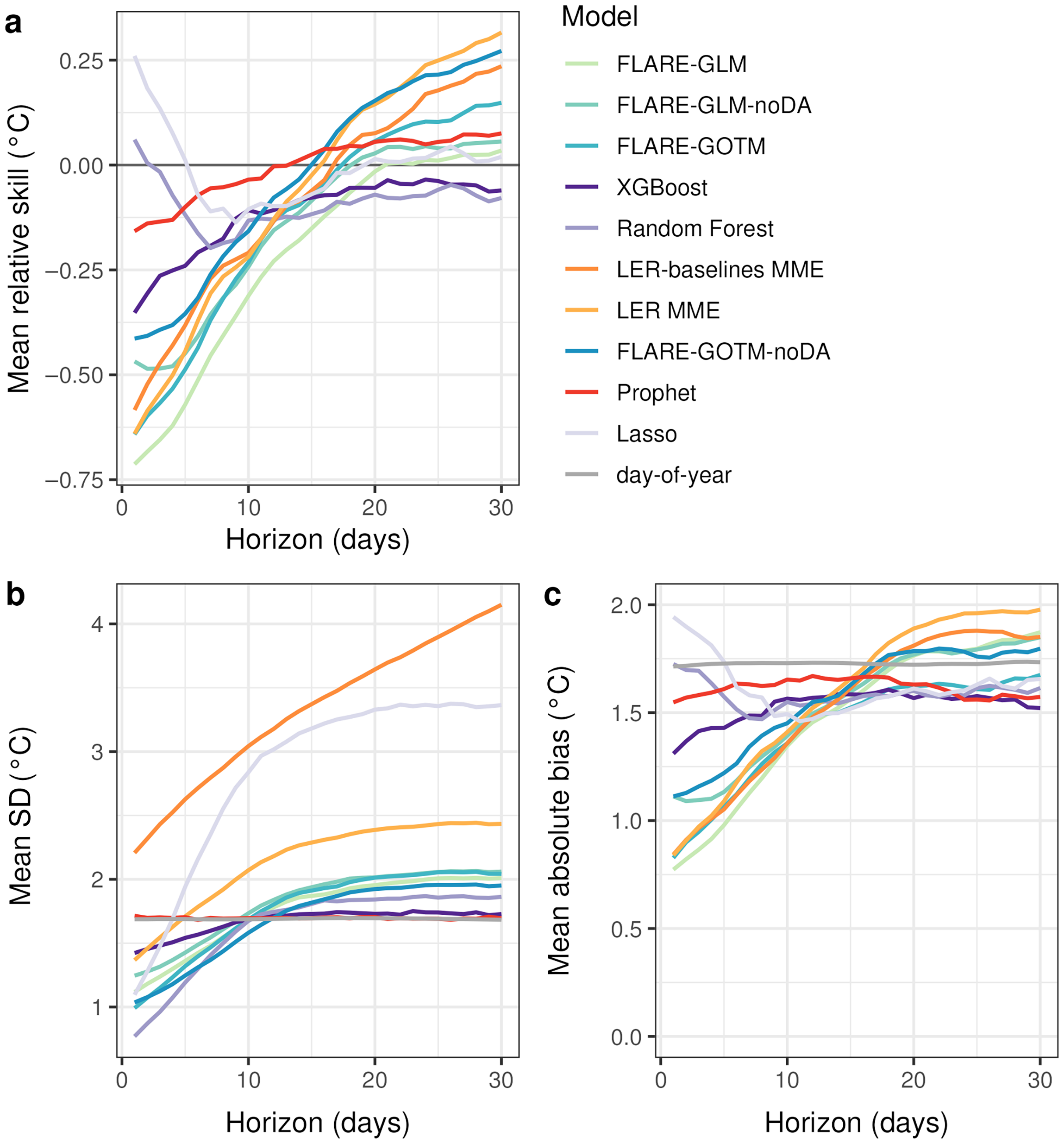

FIGURE 3 (a) Relative skill, (b) mean standard deviation (SD), and (c) mean absolute bias across the 30-day-ahead forecast horizon for the models that outperformed the day-of-year (DOY) baseline for water temperature. Relative skill was calculated as the difference in continuous rank probability score between the focal model and the DOY baseline, with negative values indicating that a submitted model performed better, on average, than the DOY baseline and positive values indicating that the baseline performed better. The metrics in each panel were averaged across all sites and forecast submission dates. Models are listed in the key in ascending order of mean skill aggregated over the forecasting period.

(Appendix S1: Figure S5). The FLARE-GLM PB model and the Random Forest ML model had the next longest durations, where they outperformed the DOY (i.e., 19 and 27 days, respectively, over the 30-day forecast horizon), but differed in the timing of these days. The Random Forest model had positive $\text{CRPS}_{\text{skill}}$ at the start of the forecast horizon and FLARE-GLM had positive relative skill at the end of the forecast period (Figure 3a), although both were only marginally worse-performing than the baseline on the days when their $\text{CRPS}_{\text{skill}}$ was positive. FLARE-GLM was the most skillful model for the first 16 days of the forecast horizon, dropping only to the fourth highest performer overall at other horizons. In contrast, the best performing model at 30 days ahead, the Random Forest model, was the second worst-performing model at 1–4 days ahead.

To what extent is relative forecast skill affected by the inclusion of different sources of uncertainty?

Although submissions were required to include an estimate of forecast uncertainty (Thomas et al., 2023), the sources of uncertainty varied among the models. The most commonly represented source of uncertainty in T_w models was driver uncertainty ($n = 22$; Appendix S1: Table S1), with 13 models including only one source of uncertainty, seven models including two sources, one model including three sources, and eight models including all five sources of uncertainty (defined in Table 1).

Of the 10 T_w models that had mean negative skill aggregated over the forecast horizon for T_w (Figure 2a), seven included at least three sources of uncertainty and six included five sources (Table 2). All but one model ($n = 9$) included driver uncertainty (in the form of the NOAA GEFS weather ensembles as covariates), with parameter and process uncertainty the next most common uncertainty source included with these top models ($n = 8$ models represented this source of uncertainty). In comparison, T_w models that failed to outperform the baseline rarely included sources of uncertainty other than driver data uncertainty (Appendix S1: Table S1).

The degradation in relative skill for the majority of T_w models at longer horizons was concurrent with an increase in bias (i.e., lower accuracy; Figure 3c) and SD (i.e., lower precision; Figure 3b). The improvement in relative skill exhibited by two ML models (Lasso and Random Forest) across the first 7 days of the forecast horizon (Figure 3a) was concurrent with reductions in absolute bias (Figure 3c). Across the first 10 days, the PB models (FLARE-GLM, FLARE-GOTM) and MMEs that included the PB models (FLARE-LER MME and LER baselines MME) exhibited the lowest absolute bias, which increased steadily across the horizon up to ~20 days ahead. In comparison, the forecast accuracy and to a certain extent, precision, in the Prophet, XGBoost, and Random Forest ML models degraded less, resulting in lower bias and SD at longer horizons (Figure 3c).

Increased SD (i.e., greater uncertainty) across the forecast horizon may indicate a reduction in precision in the forecasts, which can degrade $\text{CRPS}_{\text{skill}}$ and reliability of the forecast CIs. The top performing T_w models were primarily underconfident (Figure 4a) for the 80% CIs, meaning that >80% of observations fell within the 80% CIs. Generally, the confidence of the forecasts changed little over the horizon, especially beyond the first 5 days (Figure 4a). Beyond this horizon, only the Random Forest and Lasso ML models showed shifts in confidence beyond 5 days, becoming less overconfident and eventually becoming underconfident at horizons greater than 8 days (Figure 4). The XGBoost ML model yielded the most reliable forecasts, with 80.4% of observations in the 80% CI when averaged across horizons (Figure 4). The Prophet model was the only model that outperformed the baseline that was overconfident for the whole forecast horizon, with its uncertainty changing little across the forecast horizon (74%–79% of observations in the 80% CI; Figure 4a). The two MME models showed the highest rates of underconfidence, with 91.5% and 96.2% points falling on average into the 80% CI (Figure 4). Among the poorer performing T_w models, there was a greater rate of overconfidence, especially at horizons less than 7 days ahead, with 9 out of the 18 models overconfident. The rate of overconfidence increased among all models at the 95% CI (Figure 4b,d), demonstrating poor calibration for models when forecasting observations at the tails of the distribution.

Are the patterns in performance consistent across sites?

Within model classes, T_w forecast $\text{CRPS}_{\text{skill}}$ showed similar patterns among sites, with the exception of empirical models (Figure 5a). Generally, ML, PB models, and MMEs had negative $\text{CRPS}_{\text{skill}}$ at PRLA, PRPO, and TOOK, although the latter had a limited number of forecasts given its much shorter buoy deployment duration (Appendix S1: Figure S1). In comparison, ML models, PB models, and MMEs generally exhibited positive $\text{CRPS}_{\text{skill}}$ at SUGG, BARC, and CRAM (Figure 5a).

Mean $\text{CRPS}_{\text{skill}}$ (from the T_w models that outperformed the baseline, as shown in Figure 2a) degraded across the forecast horizon for all sites, but remained negative at PRPO and PRLA for the full 30-day horizon and at TOOK for the first 18 days (Figure 5b). In contrast, at CRAM, LIRO, BARC, and SUGG, $\text{CRPS}_{\text{skill}}$ was negative between 1 and 12 days ahead. This better $\text{CRPS}_{\text{skill}}$ at PRPO, PRLA, and TOOK is likely due to the relative gains against more poorly performing DOY baseline forecasts at these sites (Appendix S1: Figure S6). Focusing on the 4 months when all lakes had data availability (i.e., when all lakes had

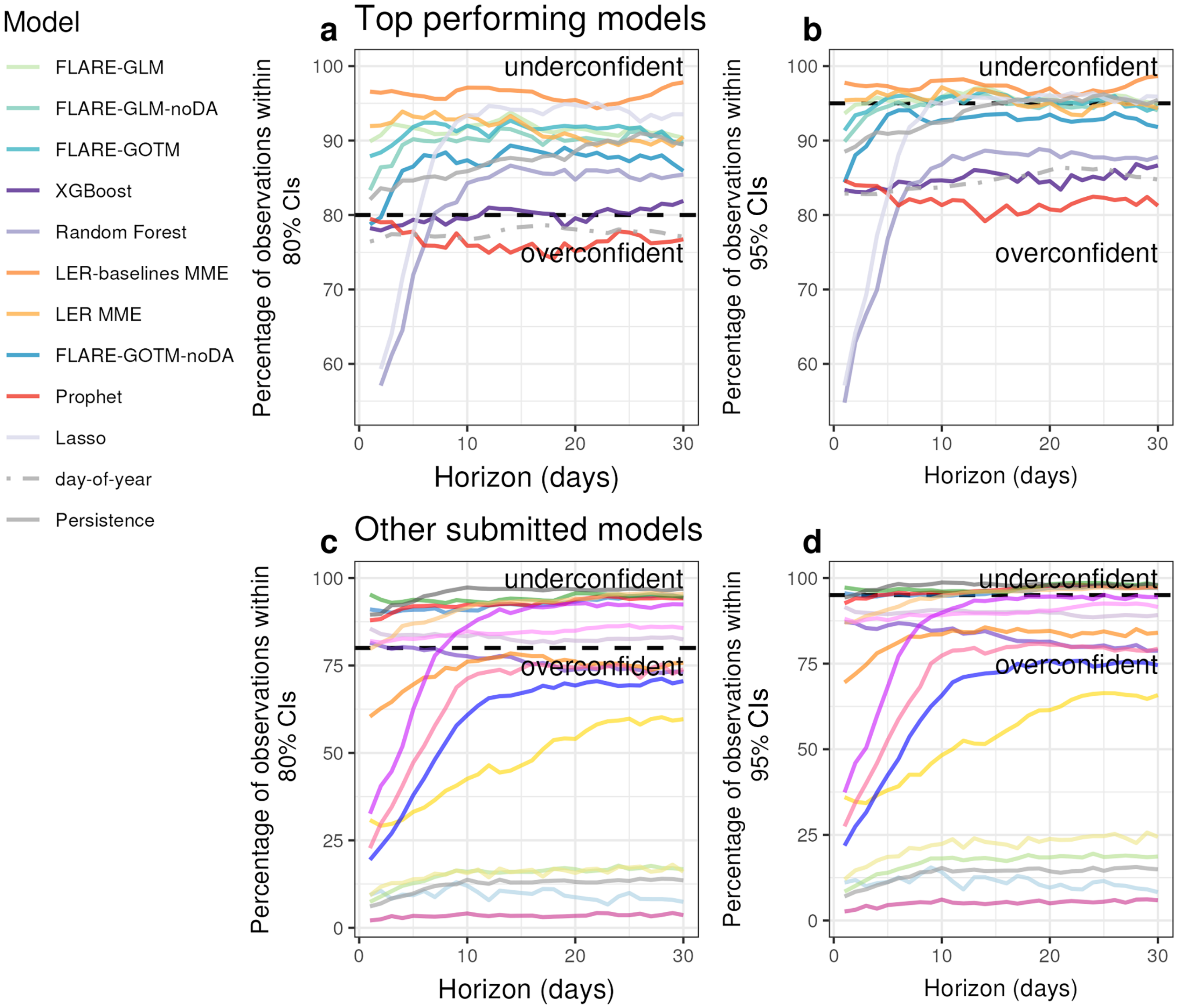

FIGURE 4 Reliability plot (percentage of observations falling within the 95% and 80% confidence intervals (CI for the water temperature models that (a, b) outperformed the day-of-year baseline and (c, d) those that did not (gray lines). Perfectly confident forecasts would have an equal percentage of observations within the CI as the percentage covered by the CI. Values above the dashed line indicate that the forecast is underconfident (forecast precision is too wide) and values below the line indicate that the forecast is overconfident (forecast precision is too narrow). Values above the dotted threshold indicate that the forecast is underconfident (i.e., there are too many observations falling within the specified CI) and values below the line indicate that the forecast is overconfident. Note the differences in scale between Panels (a, b) and (c, d) that show the 80% and 95% CIs, respectively.

buoys deployed) versus longer time periods did not substantially alter the differences in $\mathrm{CRPS}_{\mathrm{skill}}$ observed among lakes (Appendix S1: Figure S7).

Climate variability may have influenced why some models performed better than others in forecasting out-of-sample conditions. Observations for water temperatures in 2023 show that PRPO and PRLA were warmer than historical conditions represented in the DOY model, especially in May and June (Figure 6). In comparison, CRAM and LIRO, for which models performed worse than the baseline on average, exhibited water temperatures generally within around 2°C of historical conditions (Figure 6). BARC and SUGG exhibited a smaller range of water temperatures that fell within 2°C of historical conditions for all months except March (Figure 6).

DISCUSSION

Among the 29 models that forecasted water quality variables across seven lakes, 10 models outperformed the baseline model for T_w, and 6 for DO (Figure 2). Of

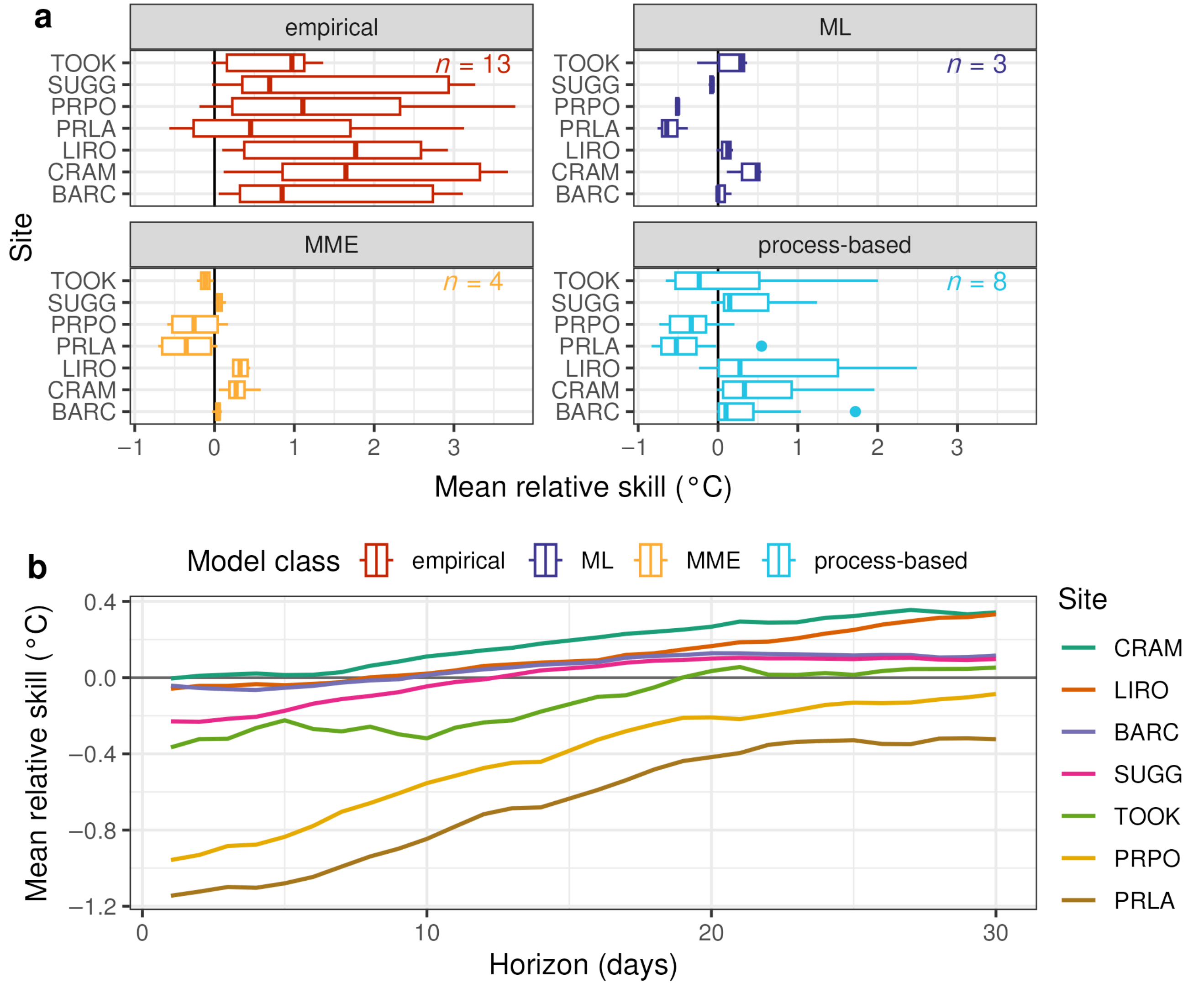

FIGURE 5 (a) Relative skill of water temperature forecasts compared with the baseline (day-of-year) for each site compared among model classes: Empirical, machine learning (ML), multimodel ensemble (MME), and process-based. Negative values indicate the submitted model performed better, on average, than the baseline and positive values indicate that the baseline performed better. The *n* value indicates the number of models represented in each model class. (b) Mean relative skill for the top 10 performing models among sites across the forecast horizon.

the 10 best performing T_w models, there were 4 PB models that included multiple exogenous weather covariates, 3 ML models, 2 multimodel ensembles, and 1 empirical model, demonstrating that multiple different model classes can yield skillful forecasts for lake water temperature. Our uncertainty analysis showed that poorly performing T_w models were generally more overconfident, likely due to insufficient representation of uncertainty in the forecasts. Finally, model skill was inconsistent across sites for the best performing lake temperature forecast models, which may be related to site-to-site differences in weather. Below, we discuss how our findings addressed our research questions, with a focus on the T_w models.

How do model class and model covariates affect forecast performance?

No individual model submitted to the challenge was the best performing model for both variables, although four models outperformed the baselines for both T_w and DO. These four models—the ML models XGBoost, Random Forest, and Lasso and the empirical model Prophet—show that a range of model types were useful for a range of variable forecasts. High-performing models for DO were in both empirical and ML categories, although no PB or MME models were submitted for DO, necessitating further investigation of both model types to potentially improve forecast performance (Hagedorn et al., 2005; Olsson, Moore, et al., 2024).

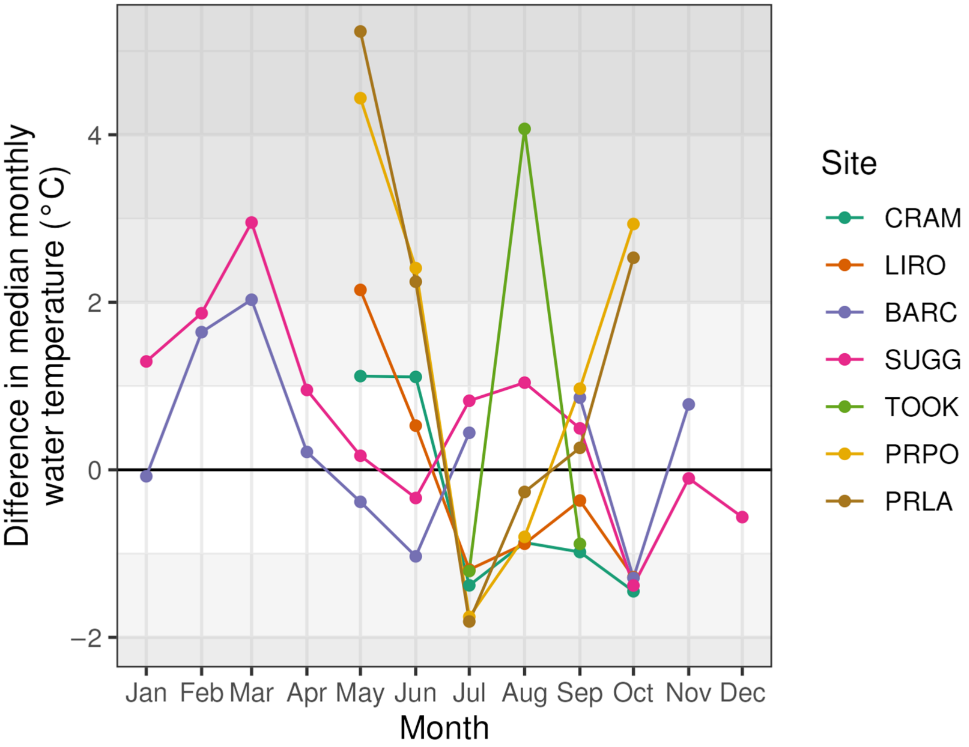

FIGURE 6 Difference in median monthly surface water temperature (depths < 1 m) between 2023 and historical observations (2015–2022) at the seven lake sites. Shaded regions show delta values that exceed 1°C from median historical conditions. Not all lakes have historical observations for the full 8-year historical period or observations during all months.

In contrast, models outperforming the baseline for T_w came from four model classes (ML, PB, empirical, and MMEs).

In an analysis of T_w models specifically (because of the higher diversity of model classes that were submitted for this variable), we found that PB models that included air temperature as a covariate performed best across all sites (Figure 2a). Air temperature is likely a key covariate for high-performing surface water temperature forecasts because T_w dynamics are primarily driven by, and tightly related to, processes at the air–water interface of lakes (Piccolroaz et al., 2024; Schmid & Read, 2022). Air temperature is a causal forcing variable and is highly correlated with other key meteorological drivers (Livingstone & Padisák, 2007). PB models that used additional meteorological parameters (e.g., incoming short-wave radiation, relative humidity, wind speed) to calculate heat fluxes to mechanistically derive water temperatures had even higher performing forecasts (Figure 2), although at some horizons the PB models were outperformed by ML models, which did not include physical processes (Figure 3a). One exception was a simple-physics PB model that included fewer sources of uncertainty and was not able to outperform the baseline model (Appendix S1: Text S1). Altogether, our results strongly support that including the dominant drivers of water temperature (namely, air temperature) unsurprisingly improved the performance of lake water temperature forecasts.

In contrast to the T_w PB models, the domain-agnostic models (i.e., models that do not include any mechanistic information about lake functioning; ML and empirical models) showed less degradation across the forecast horizon, which may be potentially due to the nondynamic nature of the methods (Appendix S1: Table S1). In comparison, the PB models were more skillful at short horizons, suggesting that forecasters might choose different T_w models based on the horizon needed. XGBoost, Lasso, and Random Forest ML models and the empirical Prophet model were less skillful than the PB models and PB-MMEs in the first 10 days, but become more skillful than the PB models at horizons >10 days due to their low rates of degradation. XGBoost was the only model that outperformed the baseline across the full forecast horizon (on average for all forecasts and sites), highlighting a robust method for forecasting T_w at any site in our study. Our results are similar to other ecological forecasting studies: for example, domain-agnostic models outperformed PB models in a penguin population forecasting competition in which annual populations were forecasted up to 3 years ahead (Humphries et al., 2018). Similarly, simple time-series models have shown promise in other ecological population forecasts (Ward et al., 2014). In the NEON

Challenge, the same ML and empirical models that performed well for T_w also performed well for DO forecasts, on average outperforming the DOY baseline, and thereby representing robust methods across multiple variables.

Reduction in the skill of T_w forecasts over the forecast horizon may be linked to a reduction in skill of the air temperature forecasts being used as model driver data. The Prophet model, which was the only model that outperformed the baseline that did not include air temperature as a covariate (or any covariates at all), showed less degradation in forecast performance than the overall better performing PB models, although this represents only a single model. The PB models, generally, benefit from high weather forecast skill at shorter horizons (Petchey et al., 2015; Zhou et al., 2022) but degrade in performance along with the performance of their covariates. Beyond 10 days ahead, when the weather forecasts are less skillful (Zhou et al., 2022), the PB models' performance also degraded, suggesting that ecological forecasting models requiring weather drivers may be restricted by the skill of weather forecasts. Future analyses that quantify the contribution of the weather driver accuracy and uncertainty to ecological forecast skill could determine whether this decline in skill is due to degradation in weather forecast skill or the accumulation of uncertainty from other sources.

The differences in the forecast horizons at which each T_w model was most skillful may present opportunities for generating MMEs or hybrid models (e.g., combining domain-agnostic models with PB models) to exploit the strengths of multiple model types across the forecast horizon. Hybrid model approaches have shown high performance in other forecasting challenges and competitions (Clark et al., 2022; Makridakis et al., 2020), and MMEs are most successful when the individual model structures are more diverse (Dormann et al., 2018; Olsson, Moore, et al., 2024; Petropoulos et al., 2022). The performance of the MMEs in this NEON Challenge synthesis was not consistent with previous studies and other forecasting challenges, in which MMEs showed the best performance (Clark et al., 2022; Makridakis et al., 2020). For example, in forecasts of tick disease incidence, the simple model average of four individual models was better than any individual model (Clark et al., 2022), and the winner of the M4 forecasting competition (a wide-ranging time series forecasting challenge) was a combination of statistical and empirical models (Makridakis et al., 2020). Similarly, in a recent single-site lake study, forecasts generated by an MME composed of three PB and two baseline models outperformed the individual models across 2 years (Olsson, Moore, et al., 2024). Conversely, in this analysis, the same MME had lower relative skill, higher bias, and higher uncertainty than some of the individual models from which it was derived (Figure 2). This discrepancy in MME performance could be caused by poor calibration in the individual models at some of the lake sites. The individual models included in this study were almost all underconfident (Figure 4), which resulted in very large uncertainty in the MMEs and likely contributed to their poor performance, as MME forecasts have been shown to be most successful when the individual constituent models are slightly overconfident (Hagedorn et al., 2005; Wang et al., 2023). Methods such as trimming, where distributions are narrowed, could help constrain MME uncertainty, increasing the overall skill of these forecasts (Howerton et al., 2023).

Finally, the differences among forecasting models, especially within model classes, can be interpreted in the context of model relatedness. For example, six of the seven PB models were 1-D hydrodynamic models that share meteorological drivers (Olsson, Moore, et al., 2024). These six PB models included three unique 1-D hydrodynamic models (GLM, GOTM, and Simstrat, see Appendix S1) with and without an ensemble Kalman filter data assimilation method. As a result of the PB models using similar equations for modeling some components of lake ecosystems (e.g., well-established surface energy balance equations) and the shared data assimilation approaches, the PB models are not entirely independent representations. Similarly, many of the empirical models were based on the same structures with differing drivers. Here, we focused on analyzing the results for broad classes of models (PB, empirical, and ML) rather than within-model classes to reduce the impact of any of model relatedness on the analysis. Future work can build on lessons learned in the climate modeling community to interpret multimodel analyses in the context of quantifying model independence and similarity (Pathak et al., 2023; Pennell & Reichler, 2011).

To what extent is relative forecast skill affected by the inclusion of different sources of uncertainty?

Our synthesis suggests that representation of forecast uncertainty is important for determining the overall forecast performance of probabilistic T_w forecasts. The top performing T_w models often included multiple sources of uncertainty (up to $n = 5$; Table 2), unlike the lower performing models, which frequently only included driver uncertainty. Consequently, many poor performing models were overconfident in their predictions, suggesting there was insufficient uncertainty included in those forecasts. By omitting parameter and process uncertainty, the forecasts fail to acknowledge the inability of the models to completely capture the ecological and stochastic processes being modeled and that even complex process models are an approximation of reality (Dietze, 2017). These results suggest that driver uncertainty alone is not a sufficient representation of the total uncertainty, especially given that weather forecasts are themselves often overconfident at shortest forecast horizons (1–7 days; Zhou et al., 2022). When these weather forecasts are used as driver data for overfitted lake models (Zwart et al., 2023), overconfidence in water quality forecasts is even more likely to occur. Overconfidence of forecasts was also reported in a forest phenology forecast synthesis, in which forecasts that included covariates were overconfident at shorter horizons (Wheeler et al., 2024). In our analysis, the Lasso and Random Forest ML models, which only included driver uncertainty, showed performance improvements from 1 to 8 days ahead (Figure 3), as the uncertainty from the weather forecasts increased and the water temperature forecasts became less overconfident (Figure 4). Furthermore, the ML XGBoost model, which included process uncertainty in addition to driver uncertainty, outperformed the other ML models at shorter horizons. Improving the representation of uncertainty for many of the models that failed to outperform the baseline could be achieved by additionally quantifying: (1) the uncertainty from the chosen model through the inclusion of parameter and/or process uncertainty; or (2) from the measurements through the inclusion of initial conditions or observational uncertainty (see Table 1).

Improving the representation of uncertainty in forecasts, as quantified by the reliability of forecast CIs, is important for management (Crochemore et al., 2021; Ramos et al., 2013). The use of ecological forecasts by decision makers is likely to improve if forecast uncertainty is well quantified and CIs are appropriate (Buizza, 2008; Nadav-Greenberg & Joslyn, 2009; Ramos et al., 2013). Underconfidence and overconfidence limit the use of forecasts for management, as underconfident forecasts provide too wide of a range of potential future conditions and overconfident forecasts underestimate the possible range of conditions, with both leading to inappropriate management actions (Crochemore et al., 2021). Consequently, our results suggest that including more than one source of uncertainty may help increase the usability of forecasts as decision support tools.

Is model forecast performance consistent across sites?

T_w forecast performance varied among sites, with the relative gain in skill likely due to the lower performance of baseline models at some lakes, especially at PRPO and PRLA, two lakes in North Dakota. The DOY baseline model had the lowest performance at PRPO and PRLA, potentially because 2023 conditions in these two lakes were substantially different from historical observations, resulting in a lower performing baseline forecast (Figure 6; Appendix S1: Figure S6). This is consistent with a previous single-model forecasting study (FLARE-GLM) that also showed improved performance above a DOY baseline for these two sites, especially at shorter horizons (Thomas et al., 2023). Differences from historical conditions that exceeded 3°C resulted in poor DOY baseline performance in that study. Our results suggest that if there is a divergence of water temperature of this magnitude, using a PB or ML model provides a much stronger forecasting approach than a baseline model. All model classes except the empirical model class showed better performance compared with the DOY baseline at PRLA and PRPO as well as at TOOK, to a lesser extent. As environmental conditions further exceed historical means due to global change, models that only consider patterns from long-term historical observations may be less valuable than models that are able to infer ecological processes or use recently-observed data in generating forecasts.

Value and refinements for forecasting challenges

Forecasting challenges provide a compelling opportunity to learn about ecological predictability over gradients of time, space, ecological level of organization, and forecasting methods. The submissions from 30 models (including two baselines) to the aquatics lake theme of the NEON Challenge covered a range of model classes and approaches. However, since the NEON Challenge was open to the community and we did not specifically guide the types of submissions, the breadth of models was not exhaustive and therefore some questions remain. Specifically, quantifying the value of different covariates to different models (e.g., XGBoost, linear models, Random Forest) would be best done by comparing forecasts with the same modeling approach but with differing covariates and quantitatively seeing how forecast skill changes with their addition or removal. It is possible that this “model selection” was done by teams before forecasts were submitted and that the final model submitted to the Challenge was the optimal structure, but we cannot know from the submitted metadata whether these models represent each team’s “best” attempt at producing a forecast.

We also saw uneven representation in the variables being forecasted, with more submitted forecasts of T_w than DO. We identified several potential factors that contributed to this uneven representation. First, NEON Challenge training materials were focused on lake temperature forecasting, which may have skewed submissions to this variable because participants in workshops may have been more likely to modify pre-existing code for submitting a new model type to T_w, rather than develop new code for DO submissions. Second, water temperature may have been an easier, more "introductory" forecast target variable as there are well-established mechanistic processes linked to driver datasets (e.g., meteorology) that were made readily available for teams to use. Conversely, the drivers of DO concentrations are much more complex, drawing from physical, chemical, and biological processes (Carey, 2023; Hanson et al., 2006; Langman et al., 2010) that vary by timescale (Hanson et al., 2006; Langman et al., 2010) and are likely to be more or less important depending on lake mixing (Robbins et al., 2024), trophic status (Steinsberger et al., 2020), and lake size (Langman et al., 2010). To use these additional driver data to forecast model lake DO processes, forecasts of those drivers must first be generated before they can be used in a model submitted to the Challenge.

Overall, our conclusions about the best performing model are limited to mean surface water temperature (the target variable chosen by the Challenge organizers), as forecasts at other depths or temporal aggregations may lead to different conclusions. For example, in contrast to our findings about surface temperature, Thomas et al. (2020) found persistence forecasts of bottom water temperature performed better than a process model because of the low variability in temperature below the thermocline. Our analyses motivate future work that focuses on different depths and temporal aggregations, as motivated by the needs of forecast users. When using forecasting methods for environmental management, the appropriateness of the forecast target (e.g., surface water temperature vs. chance of lake mixing), in addition to the chosen models should be evaluated to ensure that models, and forecast output are fit for their management purpose (Bokulich & Parker 2021; Parker, 2020). Applying methods and approaches from one application in a new situation, without accessing the fitness-for-purpose, could result in misplaced confidence or harmful outcomes (Parker, 2020).

Nonetheless, the forecasting approaches shown in this synthesis could provide a valuable starting point for developing forecasts for management decision-making or as inputs into other models and decision-support tools (e.g., Carey et al., 2022), for example, using a water temperature forecast as an input into an algal bloom risk model. For the NEON lake sites specifically, although not actively managed, water temperature forecasts of these lakes may help to optimize NEON sampling protocols, for example, by forecasting the lake ice-on dates and therefore maximizing the deployment of the water quality buoys that have to be removed during winter ice cover or to anticipate a water quality impairment event for higher frequency spatial sampling.

The NEON Challenge also sets the stage for future forecasting model analyses. For example, future work could address whether the inclusion of exogenous covariates in models produces forecasts that are overconfident at shorter horizons for other ecological variables, which could be corrected using multiple sources of uncertainty. Similarly, it would be useful to investigate whether the domain-agnostic models that outperformed the baseline for DO and T_w perform similarly well when forecasting other ecological variables. The spatial and temporal extent of NEON data, as well as the range of ecological variables on which data are collected, provides a suite of opportunities to continue to investigate these questions and as a platform to grow the field of ecological forecasting.

CONCLUSION

Our synthesis of more than 100,000 submissions to the NEON Forecasting Challenge demonstrates that several model classes were able to outperform a DOY baseline model to forecast water temperature and dissolved oxygen across seven lake sites, providing insight into optimal forecasting approaches for different contexts. Water temperature models that included air temperature as an exogenous covariate and those that included multiple sources of uncertainty generally performed well and came from PB, empirical, ML, and multimodel ensemble model classes. The relative skill of these models was shown to be highest at sites that exhibited conditions outside of historical observations. These forecasting methods are likely to become increasingly valuable for guiding decision-making in a world in which ecosystems are become more variable and continue to move outside of historically observed conditions. Overall, our results highlight the value of forecasting challenges to advance the development of ecological forecasts for both theory and management.

AUTHOR CONTRIBUTIONS

R. Quinn Thomas, Freya Olsson, and Cayelan C. Carey designed and developed the NEON EFI Aquatics Challenge. R. Quinn Thomas, Carl Boettiger, and Freya Olsson developed cyberinfrastructure. All coauthors contributed to forecasts. Freya Olsson, Cayelan C. Carey, and R. Quinn Thomas developed the synthesis and analysis approach with feedback from all coauthors. Freya Olsson led the manuscript writing, supported by Cayelan C. Carey and R. Quinn Thomas. All coauthors contributed to the text describing the forecasting approaches, provided feedback, and approved the final manuscript.

ACKNOWLEDGMENTS

The authors wish to thank the Ecological Forecasting Initiative Research Coordination Network, funded by the U.S. National Science Foundation (NSF) grant DEB-1926388, and the original aquatics theme design team members for their work in providing a foundation for the ongoing NEON Challenge. We also wish to thank NEON staff, especially Bobby Hensley and Kaelin Cawley, for their help and guidance obtaining and troubleshooting aquatics data for the Challenge. Freya Olsson was financially supported by NSF grant DBI-1933016; Cayelan C. Carey and Mary E. Lofton were supported by DEB-1926050; R. Quinn Thomas was financially supported by OAC-2209866; and Abigail S. L. Lewis was financially supported by an NSF graduate research fellowship (DGE-1840995), DEB-1753639, and funding from the Virginia Tech College of Science Roundtable.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Model code used to generate forecasts (Olsson, Carey, et al., 2024a), data analyzed in the synthesis (Olsson, Carey, et al., 2024b), and the code used to generate figures and results in this manuscript (Olsson, Thomas, et al., 2024) are available on Zenodo: https://doi.org/10.5281/zenodo.13750779, https://doi.org/10.5281/zenodo.11087208, https://doi.org/10.5281/zenodo.11093206, respectively.

ORCID

Freya Olsson https://orcid.org/0000-0002-0483-4489
Cayelan C. Carey https://orcid.org/0000-0001-8835-4476
Abigail S. L. Lewis https://orcid.org/0000-0001-9933-4542
Mary E. Lofton https://orcid.org/0000-0003-3270-1330
Caleb J. Robbins https://orcid.org/0000-0001-9579-295X
R. Quinn Thomas https://orcid.org/0000-0003-1282-7825

REFERENCES

Arias, P. A., N. Bellouin, E. Coppola, R. G. Jones, G. Krinner, J. Marotzke, et al. 2021. "Technical Summary." In *Climate Change 2021 – The Physical Science Basis. Contributions of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al., 33–144. Cambridge, United Kingdom: Cambridge University Press. https://doi.org/10.1017/9781009157896.002.

Baracchini, T., A. Wüest, and D. Bouffard. 2020. "Meteolakes: An Operational Online Three-Dimensional Forecasting Platform for Lake Hydrodynamics." *Water Research* 172: 115529. https://doi.org/10.1016/j.watres.2020.115529.

Biggerstaff, M., M. Johansson, D. Alper, L. C. Brooks, P. Chakraborty, D. C. Farrow, S. Hyun, et al. 2018. "Results from the Second Year of a Collaborative Effort to Forecast Influenza Seasons in the United States." *Epidemics* 24: 26–33.

Boettiger, C., and R. Q. Thomas. 2024. "neon4cast: Helper Utilities for the EFI NEON Forecast Challenge." R Package Version 0.1.0.

Bojer, C. S., and J. P. Meldgaard. 2021. "Kaggle Forecasting Competitions: An Overlooked Learning Opportunity." *International Journal of Forecasting* 37: 587–603.

Bokulich, A., and W. Parker. 2021. "Data Models, Representation and Adequacy-for-Purpose." *European Journal for Philosophy of Science* 11(1). https://doi.org/10.1007/s13194-020-00345-2.

Bradford, J. B., J. L. Betancourt, B. J. Butterfield, S. M. Munson, and T. E. Wood. 2018. "Anticipatory Natural Resource Science and Management for a Changing Future." *Frontiers in Ecology and the Environment* 16: 295–303.

Buizza, R. 2008. "The Value of Probabilistic Prediction." *Atmospheric Science Letters* 9: 36–42.

Caissie, D., M. E. Thistle, and L. Benyahya. 2017. "River Temperature Forecasting: Case Study for Little Southwest Miramichi River (New Brunswick, Canada)." *Hydrological Sciences Journal* 62: 683–697.

Carey, C. C. 2023. "Causes and Consequences of Changing Oxygen Availability in Lakes: Kilham Plenary Lecture Article." *Inland Waters* 13: 316–326.

Carey, C. C., W. M. Woelmer, M. E. Lofton, R. J. Figueiredo, B. J. Bookout, R. S. Corrigan, V. Daneshmand, et al. 2022. "Advancing Lake and Reservoir Water Quality Management with Near-Term, Iterative Ecological Forecasting." *Inland Waters* 12: 107–120.

Carrizo, S. F., S. C. Jähnig, V. Bremerich, J. Freyhof, I. Harrison, F. He, S. D. Langhans, K. Tockner, C. Zarfl, and W. Darwall. 2017. "Freshwater Megafauna: Flagships for Freshwater Biodiversity under Threat." *BioScience* 67: 919–927.

Chen, C., Q. Chen, S. Yao, M. He, J. Zhang, G. Li, and Y. Lin. 2024. "Combining Physical-Based Model and Machine Learning to Forecast Chlorophyll-A Concentration in Freshwater Lakes." *Science of The Total Environment* 907: 168097. https://doi.org/10.1016/j.scitotenv.2023.168097.

Cheng, M., F. Fang, T. Kinouchi, I. M. Navon, and C. C. Pain. 2020. "Long Lead-Time Daily and Monthly Streamflow Forecasting Using Machine Learning Methods." *Journal of Hydrology* 590: 125376.

Clark, N. J., T. Proboste, G. Weerasinghe, and R. J. S. Magalhães. 2022. "Near-Term Forecasting of Companion Animal Tick Paralysis Incidence: An Iterative Ensemble Model." *PLoS Computational Biology* 18: e1009874.

Clayer, F., L. Jackson-Blake, D. Mercado-bettín, M. Shikhani, A. French, D. Mercado, M. Shikhani, et al. 2023. "Sources of Skill in Lake Temperature, Discharge and Ice-off Seasonal Forecasting Tools." *Hydrology and Earth System Sciences* 27: 1361–81.

Crochemore, L., C. Cantone, I. G. Pechlivanidis, and C. S. Photiadou. 2021. "How Does Seasonal Forecast Performance Influence Decision-Making? Insights from a Serious Game." *Bulletin of the American Meteorological Society* 102: E1682–E1699.

Di Nunno, F., S. Zhu, M. Ptak, M. Sojka, and F. Granata. 2023. "A Stacked Machine Learning Model for Multi-Step Ahead Prediction of Lake Surface Water Temperature." *Science of the Total Environment* 890: 164323.

Dietze, M. C., R. Q. Thomas, J. Peters, C. Boettiger, G. Koren, A. N. Shiklomanov, and J. Ashander. 2023. "A Community Convention for Ecological Forecasting: Output Files and Metadata Version 1.0." *Ecosphere* 14(11). https://doi.org/10.1002/ecs2.4686.

Dietze, M. C. 2017. "Prediction in Ecology: A First-Principles Framework." *Ecological Applications* 27: 2048–60.

Dietze, M. C., A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T. H. Keitt, et al. 2018. "Iterative Near-Term Ecological Forecasting: Needs, Opportunities, and Challenges." *Proceedings of the National Academy of Sciences of the United States of America* 115: 1424–32.

Dodds, W. K., J. S. Perkin, and J. E. Gerken. 2013. "Human Impact on Freshwater Ecosystem Services: A Global Perspective." *Environmental Science & Technology* 47(16): 9061–68. https://doi.org/10.1021/es4021052.

Dormann, C. F., J. M. Calabrese, G. Guillera-Arroita, E. Matechou, V. Bahn, K. Bartoń, C. M. Beale, et al. 2018. "Model Averaging in Ecology: A Review of Bayesian, Information-Theoretic, and Tactical Approaches for Predictive Inference." *Ecological Monographs* 88: 485–504.

Dudgeon, D., A. H. Arthington, M. O. Gessner, Z. I. Kawabata, D. J. Knowler, C. Lévêque, R. J. Naiman, et al. 2006. "Freshwater Biodiversity: Importance, Threats, Status and Conservation Challenges." *Biological Reviews of the Cambridge Philosophical Society* 81: 163–182.

Farley, S. S., A. Dawson, S. J. Goring, and J. W. Williams. 2018. "Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions." *BioScience* 68(8): 563–576. https://doi.org/10.1093/biosci/biy068.

Fer, I., A. K. Gardella, A. N. Shiklomanov, E. E. Campbell, E. M. Cowdery, M. G. De Kauwe, A. Desai, et al. 2021. "Beyond Ecosystem Modeling: A Roadmap to Community Cyberinfrastructure for Ecological Data-Model Integration." *Global Change Biology* 27: 13–26.

Gneiting, T., F. Balabdaoui, and A. E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 69: 243–268.

Gneiting, T., and M. Katzfuss. 2014. "Probabilistic Forecasting." *Annual Review of Statistics and Its Application* 1: 125–151.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer. 2005. "The Rationale behind the Success of Multi-Model Ensembles in Seasonal Forecasting – I. Basic Concept." *Tellus A: Dynamic Meteorology and Oceanography* 57: 219.

Hamill, T. M., J. S. Whitaker, A. Shlyaeva, G. Bates, S. Fredrick, P. Pegion, E. Sinsky, et al. 2022. "The Reanalysis for the Global Ensemble Forecast System, Version 12." *Monthly Weather Review* 150: 59–79.

Hanson, P. C., S. R. Carpenter, D. E. Armstrong, E. H. Stanley, and T. K. Kratz. 2006. "Lake Dissolved Inorganic Carbon and Dissolved Oxygen: Changing Drivers from Days to Decades." *Ecological Monographs* 76: 343–363.

Harris, D. J., S. D. Taylor, and E. P. White. 2018. "Forecasting Biodiversity in Breeding Birds Using Best Practices." *PeerJ* 6: e4278.

Hipsey, M. R., L. C. Bruce, C. Boon, B. Busch, C. C. Carey, D. P. Hamilton, P. C. Hanson, et al. 2019. "A General Lake Model (GLM 3.0) for Linking with High-Frequency Sensor Data from the Global Lake Ecological Observatory Network (GLEON)." *Geoscientific Model Development* 12: 473–523.

Howerton, E., M. C. Runge, T. L. Bogich, R. K. Borchering, H. Inamine, J. Lessler, L. C. Mullany, et al. 2023. "Context-Dependent Representation of Within- and Between-Model Uncertainty: Aggregating Probabilistic Predictions in Infectious Disease Epidemiology." *Journal of the Royal Society Interface* 20: 20220659.

Huang, B., C. Langpap, and R. M. Adams. 2011. "Using Instream Water Temperature Forecasts for Fisheries Management: An Application in the Pacific Northwest." *Journal of the American Water Resources Association* 47: 861–876.

Humphries, G. R. W., C. Che-Castaldo, P. J. Bull, G. Lipstein, A. Ravia, B. Carrión, T. Bolton, A. Ganguly, and H. J. Lynch. 2018. "Predicting the Future is Hard and Other Lessons From a Population Time Series Data Science Competition." *Ecological Informatics* 48: 1–11. https://doi.org/10.1016/j.ecoinf.2018.07.004.

Hyndman, R. J., and G. Athanasopoulos. 2021. *Forecasting: Principles and Practice*. Melbourne: OTexts.

Johansson, M. A., K. M. Apfeldorf, S. Dobson, J. Devita, A. L. Buczak, B. Baugher, L. J. Moniz, et al. 2019. "An Open Challenge to Advance Probabilistic Forecasting for Dengue Epidemics." *Proceedings of the National Academy of Sciences of the United States of America* 116: 24268–74.

Jolliffe, I. T., and D. B. Stephenson. 2012. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, (2nd ed.) ed., edited by I. T. Jolliffe and D. B. Stephenson. Oxford: Wiley Blackwell.

Jordan, A., F. Krüger, and S. Lerch. 2019. "Evaluating Probabilistic Forecasts With Scoring Rules." *Journal of Statistical Software* 90(12). https://doi.org/10.18637/jss.v090.i12.

Langman, O., P. Hanson, S. Carpenter, and Y. Hu. 2010. "Control of Dissolved Oxygen in Northern Temperate Lakes over Scales Ranging from Minutes to Days." *Aquatic Biology* 9: 193–202.

Lewis, A. S. L. L., W. M. Woelmer, H. L. Wander, D. W. Howard, J. W. Smith, R. P. McClure, M. E. Lofton, et al. 2022. "Increased Adoption of Best Practices in Ecological Forecasting Enables Comparisons of Forecastability." *Ecological Applications* 32: e02500.

Lewis, A. S. L., C. R. Rollinson, A. J. Allyn, J. Ashander, S. Brodie, C. B. Brookson, E. Collins, et al. 2023. "The Power of Forecasts to Advance Ecological Theory." *Methods in Ecology and Evolution* 14: 746–756.

Livingstone, D. M., and J. Padisák. 2007. "Large-Scale Coherence in the Response of Lake Surface-Water Temperatures to Synoptic-Scale Climate Forcing during Summer." *Limnology and Oceanography* 52: 896–902.

Loescher, H. W., E. F. Kelly, and R. Lea. 2017. "National Ecological Observatory Network: Beginnings, Programmatic and Scientific Challenges, and Ecological Forecasting." In *Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities*, edited by A. Chabbi and H. W.Loescher, 27–52. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781315368252-3

Lofton, M. E., D. W. Howard, R. Q. Thomas, and C. C. Carey. 2023. "Progress and Opportunities in Advancing Near-Term Forecasting of Freshwater Quality." *Global Change Biology* 29(7): 1691–1714. https://doi.org/10.1111/gcb.16590.

Lofton, M. E., J. A. Brentrup, W. S. Beck, J. A. Zwart, R. Bhattacharya, L. S. Brighenti, S. H. Burnet, et al. 2022. "Using Near-Term Forecasts and Uncertainty Partitioning to Inform

Prediction of Oligotrophic Lake Cyanobacterial Density." *Ecological Applications* 32(5). https://doi.org/10.1002/eap.2590.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos. 2020. "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods." *International Journal of Forecasting* 36: 54–74.

McClure, R. P., R. Q. Thomas, M. E. Lofton, W. M. Woelmer, and C. C. Carey. 2021. "Iterative Forecasting Improves Near-Term Predictions of Methane Ebullition Rates." *Frontiers in Environmental Science* 9. https://doi.org/10.3389/fenvs.2021.756603.

Meyer, M. F., M. E. Harlan, R. T. Hensley, Q. Zhan, C. C. Barbosa, N. S. Börekçi, J. J. Borrelli, et al. 2023. "Hacking Limnology Workshops and DSOS23: Growing a Workforce for the Nexus of Data Science, Open Science, and the Aquatic Sciences." *Limnology and Oceanography Bulletin* 33(1): 35–38. https://doi.org/10.1002/lob.10607.

Michener, W. K., and M. B. Jones. 2012. "Ecoinformatics: Supporting Ecology as a Data-Intensive Science." *Trends in Ecology & Evolution* 27: 85–93.

Murphy, A. H. 1992. "Climatology, Persistence, and their Linear Combination as Standards of Reference in Skill Scores." *Weather and Forecasting* 7: 692–98.

Mylne, K. R. 2002. "Decision-Making from Probability Forecasts Based on Forecast Value." *Meteorological Applications* 9: 307–315.

Nadav-Greenberg, L., and S. L. Joslyn. 2009. "Uncertainty Forecasts Improve Decision Making among Nonexperts." *Journal of Cognitive Engineering and Decision Making* 3: 209–227.

Olsson, F., C. Boettiger, C. C. Carey, M. E. Lofton, and R. Q. Thomas. 2024. "Can you predict the future? A tutorial for the National Ecological Observatory Network Ecological Forecasting Challenge." *Journal of Open Source Education* 7(82): 259. https://doi.org/10.21105/jose.00259.

Olsson, F., C. C. Carey, C. Boettiger, G. Harrison, R. Ladwig, M. Lapeyrolerie, A. S. L. Lewis, et al. 2024a. "What Can We Learn from 100,000 Freshwater Forecasts? A Synthesis from the NEON Ecological Forecasting Challenge: Model Archive." Zenodo. https://doi.org/10.5281/zenodo.13750779.

Olsson, F., C. C. Carey, C. Boettiger, G. Harrison, R. Ladwig, M. Lapeyrolerie, A. S. L. Lewis, et al. 2024b. "What Can We Learn from 100,000 Freshwater Forecasts? A Synthesis from the NEON Ecological Forecasting Challenge: Scores and Targets." Zenodo. https://doi.org/10.5281/zenodo.11087208.

Olsson, F., R. Q. Thomas, and C. C. Carey. 2024. "What Can We Learn from 100,000 Freshwater Forecasts? A Synthesis from the NEON Ecological Forecasting Challenge: Scripts." Zenodo. https://doi.org/10.5281/zenodo.11093206.

Olsson, F., T. N. Moore, C. C. Carey, A. Breef-Pilz, and R. Q. Thomas. 2024. "A Multi-Model Ensemble of Baseline and Process-Based Models Improves the Predictive Skill of Near-Term Lake Forecasts." *Water Resources Research* 60: e2023WR035901. https://doi.org/10.1029/2023WR035901.

Ouellet-Proulx, S., A. St-Hilaire, and M. A. Boucher. 2017. "Water Temperature Ensemble Forecasts: Implementation Using the CEQUEAU Model on Two Contrasted River Systems." *Water* 9(7): 457.

Page, T., P. J. Smith, K. J. Beven, I. D. Jones, J. A. Elliott, S. C. Maberly, E. B. Mackay, M. De Ville, and H. Feuchtmayr. 2018. "Adaptive Forecasting of Phytoplankton Communities." *Water Research* 134: 74–85.

Pappenberger, F., M. H. Ramos, H. L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, and P. Salamon. 2015. "How Do I Know If My Forecasts Are Better? Using Benchmarks in Hydrological Ensemble Prediction." *Journal of Hydrology* 522: 697–713.

Parker, W. S. 2020. "Model Evaluation: An Adequacy-for-Purpose View." *Philosophy of Science* 87(3): 457–477. https://doi.org/10.1086/708691.

Pathak, R., H. P. Dasari, K. Ashok, and I. Hoteit. 2023. "Effects of Multi-Observations Uncertainty and Models Similarity on Climate Change Projections." *Npj Climate and Atmospheric Science* 6(1). https://doi.org/10.1038/s41612-023-00473-5.

Pennell, C., and T. Reichler. 2011. "On the Effective Number of Climate Models." *Journal of Climate* 24(9): 2358–67. https://doi.org/10.1175/2010jcli3814.1.

Petchey, O. L., M. Pontarp, T. M. Massie, S. Kéfi, A. Ozgul, M. Weilenmann, G. M. Palamara, et al. 2015. "The Ecological Forecast Horizon, and Examples of its Uses and Determinants." *Ecology Letters* 18: 597–611.

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, C. Bergmeir, et al. 2022. "Forecasting: Theory and Practice." *International Journal of Forecasting* 38: 705–871.

Piccolroaz, S., S. Zhu, R. Ladwig, L. Carrea, S. Oliver, A. P. Piotrowski, M. Ptak, et al. 2024. "Lake Water Temperature Modeling in an Era of Climate Change: Data Sources, Models, and Future Prospects." *Reviews of Geophysics* 62: e2023RG000816.

Qu, B., X. Zhang, F. Pappenberger, T. Zhang, and Y. Fang. 2017. "Multi-Model Grand Ensemble Hydrologic Forecasting in the Fu River Basin Using Bayesian Model Averaging." *Water* 9: 74.

Ramos, M. H., S. J. Van Andel, and F. Pappenberger. 2013. "Do Probabilistic Forecasts Lead to Better Decisions?" *Hydrology and Earth System Sciences* 17: 2219–32.

Read, J. S., X. Jia, J. Willard, A. P. Appling, J. A. Zwart, S. K. Oliver, A. Karpatne, et al. 2019. "Process-Guided Deep Learning Predictions of Lake Water Temperature." *Water Resources Research* 55: 9173–90.

Reid, A. J., A. K. Carlson, I. F. Creed, E. J. Eliason, P. A. Gell, P. T. J. Johnson, K. A. Kidd, et al. 2019. "Emerging Threats and Persistent Conservation Challenges for Freshwater Biodiversity." *Biological Reviews* 94: 849–873.

Richardson, D. C., A. Filazzola, R. I. Woolway, M. A. Imrit, D. Bouffard, G. A. Weyhenmeyer, J. Magnuson, and S. Sharma. 2024. "Nonlinear Responses in Interannual Variability of Lake Ice to Climate Change." *Limnology and Oceanography* 69: 789–801.

Robbins, C. J., J. M. Sadler, D. Trolle, A. Nielsen, N. D. Wagner, and J. T. Scott. 2024. "Does Polymixis Complicate Prediction of High-Frequency Dissolved Oxygen in Lakes and Reservoirs?" *Limnology and Oceanography*. https://doi.org/10.1002/lno.12650.

Rousso, B. Z., E. Bertone, R. Stewart, and D. P. Hamilton. 2020. "A Systematic Literature Review of Forecasting and Predictive Models for Cyanobacteria Blooms in Freshwater Lakes." *Water Research* 182: 115959.

Saber, A., D. E. James, and D. F. Hayes. 2020. "Long-Term Forecast of Water Temperature and Dissolved Oxygen Profiles in Deep Lakes Using Artificial Neural Networks Conjugated with Wavelet Transform." *Limnology and Oceanography* 65: 1297–1317.

Schepen, A., T. Zhao, Q. J. Wang, S. Zhou, and P. Feikema. 2016. "Optimising Seasonal Streamflow Forecast Lead Time for Operational Decision Making in Australia." *Hydrology and Earth System Sciences* 20: 4117–28.

Schmid, M., and J. Read. 2022. “Heat Budget of Lakes.” *Encyclopedia of Inland Waters*: 467–473. https://doi.org/10.1016/b978-0-12-819166-8.00011-6.

Siam, M. S., and E. A. B. Eltahir. 2017. “Climate Change Enhances Interannual Variability of the Nile River Flow.” *Nature Climate Change* 7: 350–54.

Steinsberger, T., R. Schwefel, A. Wüest, and B. Müller. 2020. “Hypolimnetic Oxygen Depletion Rates in Deep Lakes: Effects of Trophic State and Organic Matter Accumulation.” *Limnology and Oceanography* 65(12): 3128–38. https://doi.org/10.1002/lno.11578.

Sterner, R. W., B. Keeler, S. Polasky, R. Poudel, K. Rhude, and M. Rogers. 2020. “Ecosystem Services of Earth's Largest Freshwater Lakes.” *Ecosystem Services* 41: 101046. https://doi.org/10.1016/j.ecoser.2019.101046.

Thomas, R. Q., C. Boettiger, C. C. Carey, M. C. Dietze, L. R. Johnson, M. A. Kenney, J. S. McLachlan, et al. 2023. “The NEON Ecological Forecasting Challenge.” *Frontiers in Ecology and the Environment* 21: 112–13.

Thomas, R. Q., R. J. Figueiredo, V. Daneshmand, B. J. Bookout, L. K. Puckett, and C. C. Carey. 2020. “A Near-Term Iterative Forecasting System Successfully Predicts Reservoir Hydrodynamics and Partitions Uncertainty in Real Time.” *Water Resources Research* 56(11). https://doi.org/10.1029/2019wr026138.

Tulloch, A. I. T., V. Hagger, and A. C. Greenville. 2020. “Ecological Forecasts to Inform Near-Term Management of Threats to Biodiversity.” *Global Change Biology* 26: 5816–28.

Viboud, C., K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, and A. Vespignani. 2018. “The RAPIDD Ebola Forecasting Challenge: Synthesis and Lessons Learnt.” *Epidemics* 22: 13–21.

Wang, X., R. J. Hyndman, F. Li, and Y. Kang. 2023. “Forecast Combinations: An over 50-Year Review.” *International Journal of Forecasting* 39: 1518–47.

Ward, E. J., E. E. Holmes, J. T. Thorson, and B. Collen. 2014. “Complexity Is Costly: A Meta-Analysis of Parametric and Non-Parametric Methods for Short-Term Population Forecasting.” *Oikos* 123: 652–661.

Wheeler, K. I., M. C. Dietze, D. LeBauer, J. A. Peters, A. D. Richardson, A. A. Ross, R. Q. Thomas, et al. 2024. “Predicting Spring Phenology in Deciduous Broadleaf Forests: NEON Phenology Forecasting Community Challenge.” *Agricultural and Forest Meteorology* 345: 109810.

Willson, A. M., H. Gallo, J. A. Peters, A. Abeyta, N. Bueno Watts, C. C. Carey, T. N. Moore, et al. 2023. “Assessing Opportunities and Inequities in Undergraduate Ecological Forecasting Education.” *Ecology and Evolution* 13: 1–16.

Woelmer, W. M., L. M. Bradley, L. T. Haber, D. H. Klinges, A. S. L. Lewis, E. J. Mohr, C. L. Torrens, K. I. Wheeler, and A. M. Willson. 2021. “Ten Simple Rules for Training yourself in an Emerging Field.” *PLoS Computational Biology* 17: e1009440.

Zhou, X., Y. Zhu, D. Hou, B. Fu, W. Li, H. Guan, E. Sinsky, et al. 2022. “The Development of the NCEP Global Ensemble Forecast System Version 12.” *Weather and Forecasting* 37: 1069–84.

Zwart, J. A., S. K. Oliver, W. D. Watkins, J. M. Sadler, A. P. Appling, H. R. Corson-Dosch, X. Jia, V. Kumar, and J. S. Read. 2023. “Near-Term Forecasts of Stream Temperature Using Deep Learning and Data Assimilation in Support of Management Decisions.” *JAWRA Journal of the American Water Resources Association* 59: 317–337.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Olsson, Freya, Cayelan C. Carey, Carl Boettiger, Gregory Harrison, Robert Ladwig, Marcus F. Lapeyrolerie, Abigail S. L. Lewis, et al. 2025. “What Can we Learn from 100,000 Freshwater Forecasts? A Synthesis from the NEON Ecological Forecasting Challenge.” *Ecological Applications* 35(1): e70004. https://doi.org/10.1002/eap.70004