A multi-model ensemble of baseline and process-based models improves the predictive skill of near-term lake forecasts

Freya Olsson[1*], Tadhg Moore[1,2], Cayelan C. Carey[1], Adrienne Breef-Pilz[1], R. Quinn Thomas[1,3]

Freya Olsson: 0000-0002-0483-4489, freyao@vt.edu
Tadhg Moore: 0000-0002-3834-8868, tadhgm@vt.edu; tadhg.moore6@gmail.com
Cayelan C. Carey: 0000-0001-8835-4476, cayelan@vt.edu
Adrienne Breef-Pilz: 0000-0002-6759-0063, abreefpilz@vt.edu
R. Quinn Thomas: 0000-0003-1282-7825, rqthomas@vt.edu

[1]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, USA 24061
[2]Independent researcher
[3]Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, Virginia, USA 24061
*Corresponding author (freyao@vt.edu)

Keywords

forecasting; multi-model ensemble; water temperature; process-based models; baseline models

Key points (140 characters)

- Aggregated lake temperature forecast skill was higher for multi-model ensemble forecasts than individual model forecasts
- Including baseline empirical models (climatology, persistence) with process models improved multi-model ensemble forecast performance
- Multi-model ensemble forecasts improved forecast skill by ‘hedging’, as no individual model performed best at all horizons or depths

Abstract

Water temperature forecasting in lakes and reservoirs is a valuable tool to manage crucial freshwater resources in a changing and more variable climate, but previous efforts have yet to identify an optimal modelling approach. Here, we demonstrate the first multi-model ensemble (MME) reservoir water temperature forecast, a forecasting method that combines individual model strengths in a single forecasting framework. We developed two MMEs: a three-model process-based MME and a five-model MME that includes process-based and empirical models to forecast water temperature profiles at a temperate drinking water reservoir. Our results showed that the five-model MME improved forecast performance by 8-30% relative to individual models and the process-based MME, as quantified using an aggregated probabilistic skill score. This increase in performance was due to large improvements in forecast bias in the five-model MME, despite increases in forecast uncertainty. High correlation among the process-based models resulted in little improvement in forecast performance in the process-based MME relative to the individual process-based models. The utility of MMEs is highlighted by two results: 1) no individual model performed best at every depth and horizon (days in the future), and 2) MMEs avoided poor performances by rarely producing the worst forecast for any single forecasted period (<6% of the worst ranked forecasts over time). This work presents an example of how existing models can be combined to improve water temperature forecasting in lakes and reservoirs and discusses the value of utilising MMEs, rather than individual models, in operational forecasts.

1 Introduction

In the face of increased ecosystem variability, researchers are developing new methods for forecasting freshwater quality and quantity (Lofton et al., 2023). Here, we define a forecast as a prediction of a future state of a variable with quantified uncertainty (Lewis et al., 2022). Forecasts of freshwater variables have considerable potential for improving management and guiding ecosystem service provision as environmental conditions increasingly exceed the historical envelope due to climate and land use change (Bradford et al., 2020; Dietze et al., 2018; IPCC, 2023). Despite the urgent need for freshwater forecasts, however, the optimal modelling approach for developing forecasts remains unresolved across different spatial and temporal scales. One promising forecasting approach that has emerged from other disciplines is multi-model ensembles (MMEs), in which more than one model is used to simultaneously forecast the same variable into the future (Chandler, 2013; Clark et al., 2022; Humphries et al., 2018; Kirtman et al., 2014; Long et al., 2021; Velázquez et al., 2011). To date, MMEs have not been applied to freshwater forecasting (reviewed by Lofton et al., 2023), motivating the need to understand how an MME forecast performs relative to individual models, as well as how the structure of the different models in the MME influences forecast performance.

Water temperature forecasting in lakes and reservoirs is an ideal application for testing the performance of MMEs. First, water temperature forecasts can be useful for the management of inland waters (Lofton et al., 2023). For example, water temperature forecasts are used to optimise downstream water release from reservoirs (Huang et al., 2011; Jackson-Blake et al., 2022; Weber et al., 2017; Zwart et al., 2023), guide water quality management related to lake mixing events (Carey, Woelmer, et al., 2022; Thomas et al., 2020), as well as underpin the development of other water quality and ecological forecasts (Huang et al., 2011; Page et al., 2018; Weber et al., 2017), given the importance of water temperature for determining metabolism, water chemistry, and biological growth (Carey et al., 2012; Kraemer et al., 2017; Yvon-Durocher et al., 2015). Second, a wide range of models have been developed to predict lake and reservoir water temperatures, thereby providing an excellent opportunity for examining the sensitivity of an MME's performance to the identity and structure of multiple component models.

To date, process-based models (Baracchini et al., 2020; Clayer et al., 2023; Mercado-Bettín et al., 2021; Thomas et al., 2020), machine learning and data-driven models (Read et al., 2019; Zhu et al., 2020; Zwart et al., 2023), as well as a range of "hybrid" approaches (e.g. Graf et al., 2019; Zhu et al., 2020) have been used to forecast near-term dynamics (days to seasons ahead) in lake and reservoir water temperatures, with varying levels of performance (reviewed by Lofton et al., 2023). Of these modelling approaches, process-based models (hereafter, process models) have shown substantial promise, especially in near-term forecast horizons (Baracchini et al., 2020; Carey, Woelmer, et al., 2022; Mercado-Bettín et al., 2021; Thomas et al., 2020), with a performance of 0.4 - 1.4 °C RMSE (root mean square error) for reservoir water temperature forecasted 1-16 days-ahead (Thomas et al., 2020). However, the skill of these models is often limited by the skill of other forecasts (e.g., weather and inflow discharge) needed as model driver data (Mercado-Bettín et al., 2021; Thomas et al., 2020). Moreover, process models also often demonstrate substantial differences in skill among forecasted sites (Thomas et al., 2023) and depths (Thomas et al., 2020), as well as at different times of year (e.g., in thermally-stratified vs mixed conditions; Thomas et al., 2020; Wander et al., 2023).

Despite their simplicity, simple empirical models such as persistence and climatology (historical day-of-year mean and variance) models can also provide useful forecasts (Ward et al., 2014). Often used as null models to test the skill of emerging forecasting approaches (Lofton et al., 2023; Pappenberger et al., 2015), these simple baseline models include information on current conditions and seasonal trends that influence lake temperature dynamics. For example, a persistence model can be useful for forecasting dynamics in systems with high inertia that exhibit small changes across the forecast horizon (i.e., time into the future; Ward et al., 2014), which is common in lakes and reservoirs that exhibit seasonal thermal stratification. Additionally, climatology forecasts exhibit high performance at longer horizons (e.g., months to years), for which repeatable seasonal cycles dominate the dynamics (Pappenberger et al., 2015).

Multi-model ensembles (MMEs) that integrate both process models and these simple baseline models may be particularly effective for forecasting lake and reservoir water temperatures. This type of MME may be able to overcome the limitations of individual process and baseline models that are unable to consistently forecast all environmental conditions with high accuracy across space (i.e., multiple depths in a lake), time (i.e., different seasons within a year), and forecast horizons. Implementation in other disciplines has overwhelmingly found that MMEs often produce more skillful forecasts, on average, than individual model forecasts (Atiya, 2020; Clark et al., 2022; Humphries et al., 2018; Velázquez et al., 2011). Using MMEs also leads to greater diversity in forecast predictions, potentially increasing decision-making success (Boettiger, 2022). Although predictions from individual models can outperform the aggregated prediction from the MME locally, at a specific depth, time, or horizon (Abrahart & See, 2002; Atiya, 2020), it is often not known *a priori* which forecast model will be best at any given future timestep, especially for forecasts of sites with substantial spatial and temporal heterogeneity. MMEs are ideally suited for these situations, because they integrate information from different model structures into a single forecast, enabling the forecaster to ‘hedge’ (i.e., minimise risk of incorrect forecasts by assigning non-zero probability to a wide range of possible outcomes) and provide a more comprehensive and accurate representation of the potential forecasted outcomes than individual models (Abrahart & See, 2002; Atiya, 2020). MMEs have been successfully applied to a diverse range of ecological and environmental forecasting applications, including ticks (Clark et al., 2022), sea level (Long et al., 2021), penguins (Humphries et al., 2018), and river flow (Abrahart & See, 2002; Velázquez et al., 2011), suggesting that their application for forecasting freshwater ecosystems has promise.

To the best of our knowledge, no one has applied an MME approach to forecasting lake and reservoir temperatures with specified uncertainty. While MMEs for water temperatures have been applied to long-term projections (Almeida et al., 2022; Feldbauer et al., 2022; La Fuente et al., 2022; Wynne et al., 2023), or as model inter-comparisons (Golub et al., 2022), the utility of MMEs for real-time water temperature forecasting remains unknown. This gap may exist because ensemble near-term forecasts have, to date, focused on using ensembles of multiple driver datasets (e.g., weather forecasts; Mercado-Bettín et al., 2021) and parameter sets (e.g.

Thomas et al., 2020) to partition and quantify uncertainty (Clayer et al., 2023; Thomas et al., 2020), rather than using multiple models to generate more skillful operational forecasts.

Here, we developed a near-term forecasting system that integrates an MME of lake process models, baseline empirical models, and data assimilation algorithms in an automated forecasting approach. We used this MME to produce weekly, 1-14 day-ahead forecasts of water temperature profiles for two years in a small, temperate, drinking water reservoir. We aimed to understand how MME approaches may improve near-term forecast performance and how forecast performance varies over different spatial scales and forecast horizons. We used the MME forecasts to answer the research questions: 1) How does the forecast performance of the process model MME compare to the individual process models?, 2) How does the addition of the baseline models into the MME affect forecast performance?, and 3) How does the forecast performance of the individual models and MMEs vary across horizons and depths? Our goal was to determine if MMEs can improve freshwater water quality forecasting to guide the development of operational forecasting workflows.

2 Methods

2.1 Overview of forecasting system

Here, we summarise the automated MME forecasting framework (Figure 1) that leverages the state-of-the-art FLARE (Forecasting Lake And Reservoir Ecosystems) water forecasting system (Thomas et al., 2020). FLARE uses *in situ* water temperature sensor data, which are wirelessly transmitted directly from the waterbody to the cloud, in a data assimilation algorithm to update model initial conditions and to calibrate model parameters (Figure 1; Daneshmand et al., 2021). FLARE's ensemble-based forecasting algorithm generates forecasts using process hydrodynamic models that quantify the uncertainty from driver data (weather forecasts), initial conditions, model process, and model parameters and then samples from these sources of uncertainty to generate probability distributions for water temperature at multiple lake or reservoir depths (see Thomas et al., 2020).

Instead of using a single process model, as has been done in previous implementations of FLARE (Carey, Woelmer, et al., 2022; Thomas et al., 2020, 2023), we used three different process models, implemented via integration with *LakeEnsemblR* R software (LER; Moore et al., 2021), to answer question 1. These process models were run inside the FLARE framework to generate a multi-model ensemble (MME) from the output (Figure 1), hereafter referred to as the process model MME forecast (PM MME hereafter). To answer questions 2 and 3, two baseline models were also included to produce the full MME forecast (full MME hereafter), which consisted of five individual models (n=3 process models and n=2 baseline models). Finally, these forecasts are evaluated using the in-situ water temperature observations (Figure 1) via a suite of metrics, described below.

2.2 Site description and data collection

We generated water temperature forecasts for Falling Creek Reservoir (FCR), a eutrophic reservoir located in Vinton, Virginia, USA (37.30°N, 79.84°W). FCR is managed by the Western Virginia Water Authority as a drinking water source. The reservoir has a mean depth of 4 m and a maximum depth of 9.3 m, with a surface area of 0.12 km^2 (Carey, Lewis, et al., 2022). A dimictic system, FCR generally stratifies from May to October and has intermittent ice-cover from December to March (Carey & Breef-Pilz, 2023). The reservoir has one primary inflow and water level is maintained to be generally constant over time.

FCR is monitored by a series of high-frequency sensors deployed at fixed depths in the water column at its deepest site near the dam. Water temperature data were collected using T-Node FR thermistors (NexSens, Fairborn, OH, USA) from March 2019 to March 2023 (Carey et al., 2023; Olsson et al., 2023a), with minor data gaps due to sensor maintenance (see metadata in Carey et al., 2023), across ten depths in the water column (0.0, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0 m). Additional temperature data were collected at 1.6 m with a YSI EXO2 sonde (Xylem Inc., Yellow Springs, OH, USA), and at 5.0 and 9.0 m using RDO PRO-X Dissolved Oxygen Sensors (In-Situ Inc., Fort Collins, CO, USA). All measurements were collected at a 10-minute frequency and averaged to an hourly timestep. Observations were then binned into 0.25 m intervals, so that they could be matched with the process model output. When multiple measurements were collected at the same depths, a mean value was calculated. These data were used in FLARE data assimilation and process model parameter tuning, as well as inputs to the two baseline models (see section 2.3), and in forecast evaluation (Figure 1, section 2.4).

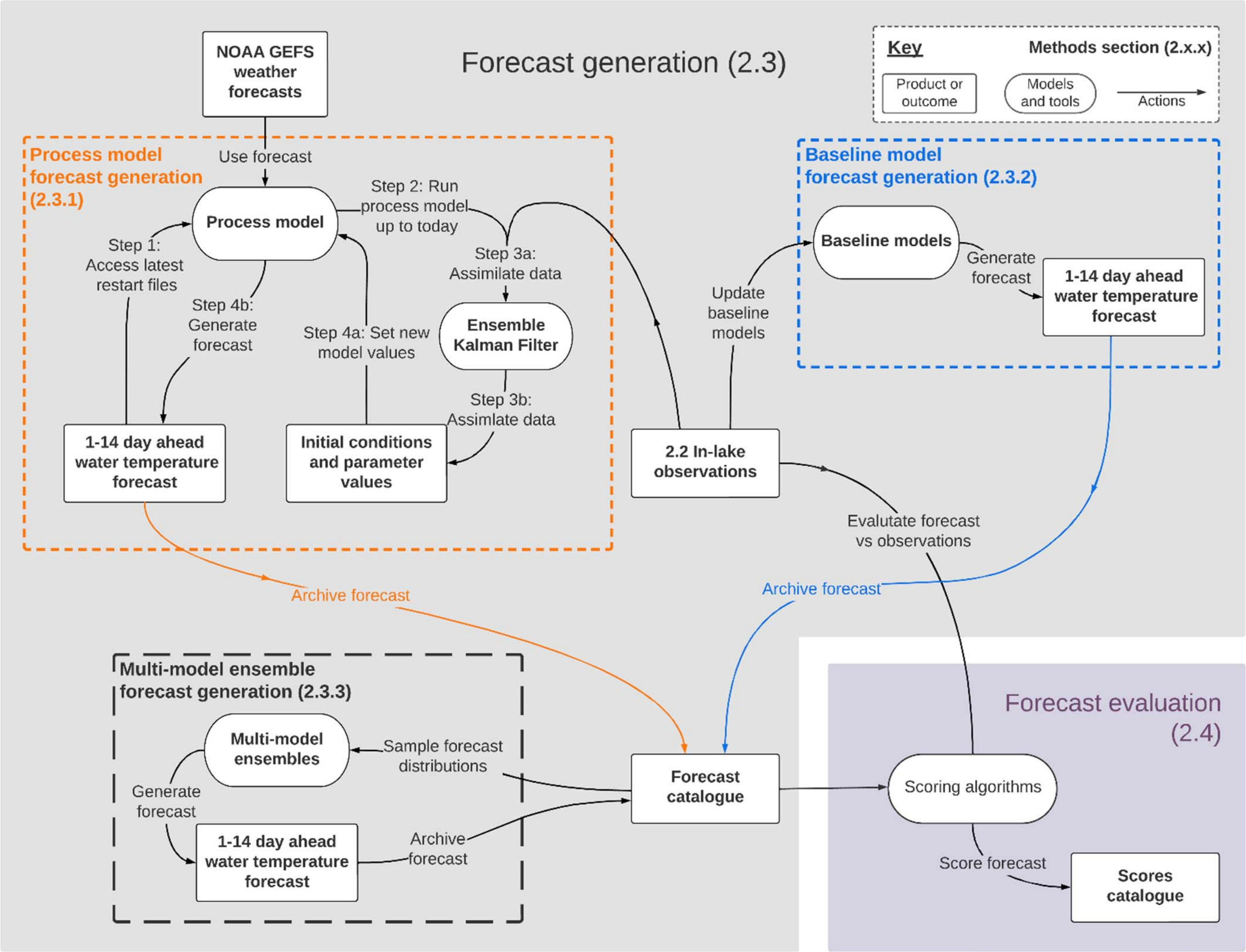
Forecast generation (2.3)
NOAA GEFS weather forecasts
Key
Methods section (2.x.x)
Product or outcome
Models and tools
Actions
Process model forecast generation (2.3.1)
Use forecast
Process model
Step 1: Access latest restart files
Step 2: Run process model up to today
Step 3a: Assimilate data
Ensemble Kalman Filter
Step 4a: Set new model values
Step 4b: Generate forecast
Step 3b: Assimlate data
1-14 day ahead water temperature forecast
Initial conditions and parameter values
2.2 In-lake observations
Baseline model forecast generation (2.3.2)
Baseline models
Update baseline models
Generate forecast
1-14 day ahead water temperature forecast
Evalutate forecast vs observations
Archive forecast
Archive forecast
Multi-model ensemble forecast generation (2.3.3)
Multi-model ensembles
Sample forecast distributions
Generate forecast
1-14 day ahead water temperature forecast
Archive forecast
Forecast catalogue
Forecast evaluation (2.4)
Scoring algorithms
Score forecast
Scores catalogue

Figure 1. Multi-model ensemble and individual forecast generation (grey shading) and forecast evaluation workflow (purple shading), with each corresponding text section number in parentheses (e.g., 2.3). Boxes represent *tools, objects, and/or products* and lines represent *actions.* The parallel workflows of individual model forecast generation are shown in the orange (process models) and blue (baseline models) boxes. Within the process model forecast workflow, the *steps* correspond to the text in section 2.3.1. Each individual model forecast is archived into the *Forecast Catalogue* from which the distributions are sampled and combined in the multi-model ensemble (black dashed box). The multi-model ensemble forecasts are also archived in the Forecast Catalogue. From this catalogue, forecasts are evaluated against in-lake observations via several scoring algorithms to generate a *Scores Catalogue*, which is subsequently analysed (section 2.4).

Dates of the mixed and stratified periods were calculated based on the density difference between 1 m from surface (1 m) and ~1 m above the bottom (8 m) of the lake, with a minimum density difference of 0.1 kg m^{-2} indicating that the lake was stratified (Wilson et al., 2020). The stratified period was defined as the summer period when continuous stratified conditions occurred and the mixed period as any time outside of this. In addition, the observed thermocline depth was estimated using the *LakeAnalyzer* R package (Read et al., 2011).

2.3 Forecast generation

2.3.1 Process model forecasts

In this application, FLARE generated forecasts for each of the three process models every seven days for 1 to 14 day-ahead horizons over 2 years (March 2021 - March 2023), resulting in a total of n=104 forecasts per model. For each process model, the weekly forecasts were generated using the following steps (Figure 1): Step 1) access the individual FLARE forecasts (Figure 1, step 1) for 1-week ago in a prior FLARE run (or, in the case of the first forecast, following a spin-up described below); Step 2) use this prediction to initialise each process model FLARE run that starts 1-week ago and runs to current day (Figure 1, step 2); Step 3) use a data assimilation algorithm (the ensemble Kalman filter; Evensen, 2003) to assimilate new observations collected over the past week (Figure 1, step 3) to update that model's states and parameters (Figure 1, step 3); and Step 4) use the updated states and parameters as initial conditions for a 1- to 14-day ahead forecast that starts today (Figure 1, step 4). Each forecast included 256 simulations (ensemble members) that quantified the uncertainty from driver data (weather forecasts), initial conditions, model process, and model parameters. Additional information about FLARE configuration can be found in Thomas et al., (2020) and Thomas et al. (2023).

Within the FLARE framework, three process models were implemented using the *LakeEnsemblR* R package (LER; Moore et al., 2021), and underwent data assimilation within FLARE as described above. This R package facilitates the running of up to five one-dimensional hydrodynamic lake models simultaneously using the same driving data and configuration files (see Moore et al. 2021). The three process models we included in the PM MME were the General Lake Model (GLM; Hipsey et al., 2019), General Ocean Turbulence Model (GOTM; Umlauf et al., 2005), and Simstrat (Goudsmit et al., 2002), hereafter referred to as PM1, PM2, and PM3, respectively. The other two process models implemented in LER (FLake, MyLake) were not included because our aim was to apply data assimilation and iteratively forecast full water column temperature profiles. Specifically, FLake simulates lake systems using a two-layer representation (Mironov, 2021) that does not simulate a full profile, and MyLake is not able to “restart” daily (Saloranta & Andersen, 2007), as needed for iterative forecasting with data assimilation.

All three process models require forecasted meteorological driving data to produce water temperature forecasts. To make near-term predictions of water temperature, we used weather forecasts for FCR from the National Oceanic and Atmospheric Administration's (NOAA's) Global Ensemble Forecast System (GEFS; Hamill et al., 2022). The NOAA GEFS weather forecast consists of a set of 31 simulations and a forecast horizon of 1 to 16 days-ahead, which we used to produce 1-14 day-ahead water temperature forecasts from the midnight UTC data product.

We followed the standardised FLARE configuration for forecasting (Thomas et al., 2023). All process models were run at an hourly time step with the midnight output as the daily forecast. A spin-up of all models was run from 1 October 2020 to 1 March 2021, the date of the first forecast. During this spin-up, each model’s parameters were individually tuned by the ensemble Kalman filter within FLARE (see Supplementary Information, Table S1, Figure S1). Each model used default parameters to initialise the forecast run and two sensitive parameters were tuned in the data assimilation process of FLARE (See Supplementary Information, Table S1, Figure S1). The sensitive parameters selected, based on initial investigation and configuration in other lakes, were the sediment temperature and incoming shortwave radiation scaling factor for GLM, and the wind scaling and incoming shortwave radiation factors for Simstrat and GOTM (see Supplementary Information).

2.3.2 Baseline models

Two simple, empirical baseline models were also used to generate forecasts (Figure 1). The persistence model uses the last observation for each specific depth as a prediction of future conditions and the climatology model uses a long-term day-of-year mean as the daily forecast (Hyndman & Athanasopoulos, 2021; Jolliffe & Stephenson, 2012); both are described in detail below.

2.3.2.1 Persistence model

A persistence model assumes that, on average, the forecasted state (in this case, water temperature) on average does not change over the forecast horizon, with uncertainty driven by a random walk process (Hyndman & Athanasopoulos, 2021):

$$y_{T+1} = y_T + e_{T+1} \quad \text{(Eqn. 1)}$$

where y_T is today's observation or forecast, e_{T+1} is random noise, and y_{T+1} is the next day's forecast. The uncertainty (e_{T+1}) in the persistence model forecasts were generated using a bootstrapping method, as a normal distribution could not be assumed. The bootstrap method calculates the distribution of residuals from the fit and samples from that distribution for a value of e_{T+1}. We used bootstrapping to generate a set of n=256 ensemble members to match the number of simulations as the process models. The persistence model forecasts were generated using the *RW* (random walk) function in the *fable* R package (version 0.3.2; O'Hara-Wild et al., 2022).

2.3.2.2 Climatology model

A climatology model, also based on historic observations, was used to generate a forecast, assuming forecasted mean conditions are equal to the historic day-of-year mean. We used two years of observations (March 2019 - March 2021) from FCR to calculate a day-of-year mean water temperatures at each depth. We chose this period because the thermistor sensors were deployed in summer 2018 and we wanted to ensure that each day-of-year mean water temperature was derived from the same number of historical observations. To obtain uncertainty around these climatology forecasts, we fitted a linear model between the two years of observations and calculated the standard deviation of the residuals, at each depth independently. We generated the probabilistic climatology forecasts by sampling from a normal distribution with the obtained mean and standard deviation, generating n=256 ensemble members.

2.3.3 Multi-model ensembles (MMEs)

As described above, we generated two MMEs: the PM MME (containing PM1, PM2, PM3; n = 3 models total) and the full MME that also included the two baseline models (persistence, climatology; n = 5 models total). To create the full MME forecasts, the n=256 ensemble members from each of the three individual process models and two baseline models were combined into a new MME (Figure 1). As the number of simulations generated from forecast can affect forecast skill (Machete & Smith, 2016), we sampled from the pool of individual model simulations to generate MMEs with n=256 ensemble members, with each model equally represented. The number of simulations from each individual model in the MME forecasts is given as 256/*n*, where *n* is the number of models in the MME. For example, in the full MME there were 5 models represented in the forecast, giving 51 simulations (256/5) from each individual model.

2.4 Forecast evaluation

Forecasts from both the individual models and MMEs were evaluated using four evaluation metrics calculated on each forecast-observation pair. We used multiple evaluation metrics because each metric provides complementary information about the performance of the forecast. First, we calculated the mean bias (difference in mean forecasted water temperature and observed water temperature). Forecasts with lower bias indicate increased forecast accuracy. Second, we calculated the standard deviation (SD) of the forecasts to understand uncertainty in the forecasts. We expect uncertainty to increase across the forecast horizon as confidence in future conditions decreases. We also expect to see larger SD in the MME forecasts than individual model forecasts as they reflect a greater diversity of predictions. Both metrics are useful for determining how the forecast accuracy (bias) and precision (using SD as a metric of uncertainty) vary independently and are commonly calculated metrics for forecast performance (Jolliffe & Stephenson, 2012).

Third, we evaluated the models using the ignorance score (IGN), which uses both the accuracy and the precision of the forecasts in its evaluation, and describes the probability placed by the forecast on the observed outcome (Smith et al., 2015). IGN was calculated using the *scoringRules* R package (Jordan et al., 2019), in which larger values represent a lower probability placed on the observed outcome and lower forecast performance. IGN, originally proposed by Good (1952), is defined as:

$$IGN(p(x), X) = -\log_2 (p(X)) \quad \text{(Eqn. 2)}$$

where $p(x)$ is the density assigned to the outcome X.

IGN penalises forecasts that place very low probabilities on the observed outcome and gives an infinitely large score if a forecast places zero probability on an outcome that is ultimately observed (Smith et al., 2015). We selected the IGN score as a focal evaluation metric because differences in scores between models represent the additional probability placed on the observed outcome in the more skillful forecast (Smith et al., 2015). The difference in IGN scores between two models can be used as the exponent of base two to calculate the probability difference between the models (Smith et al., 2015). For example, an IGN score difference of 0.5 units between two models corresponds to the better model placing $2^{0.5}$, or 1.41 times more probability, on the more skillful forecast. Thus, in this example, there is a confidence gain of 41% in the better model compared to the other model (Smith et al., 2015).

Finally, we calculated shadowing time, which quantifies the time that the forecast is able to "shadow" the observations, given an estimate of observational uncertainty (Gilmour & Smith, 1997; Smith et al., 2010). The shadowing time is the maximum number of consecutive days, starting from forecast initiation, that at least one simulation (ensemble member) tracks the mean observation, within a specified observation uncertainty. Here, we define a simulation as shadowing when it falls within the 95% confidence interval of each observation (assuming a normal distribution centred on the observation). Observational uncertainty (standard deviation) was estimated at 0.2°C, based on an analysis of the variation in observations within each day and depth (see Supplementary Information, Figure S2). Shadowing time is a useful metric to determine how well the forecast models can replicate the dynamics of a system, rather than the statistics of the forecast (Gilmour & Smith, 1997; Smith et al., 2010).

2.5 Analyses

First, to address question 1, we compared the evaluation metrics among the individual process model forecasts and the PM MME forecast. Second, to address question 2, we compared the full MME with the PM MME, and the performance of the five individual process models and baseline models. To understand how and why the MME forecasts might be able to outperform individual models, we also calculated the Pearson correlation coefficient (r) on forecast bias. Third, to address question 3, we compared the forecast metrics at different depths and forecast horizons. We also determined each model's rank (out of the 7 forecasts from the 5 individual and 2 MMEs) for each individual forecast-observation pair across depth and horizon using the

IGN score. All analyses were conducted using R statistical software (v.4.2.1; R Core Team, 2021).

2.6 Archiving

All data and code are archived and available in the Zenodo repositories (Olsson et al., 2023a, 2023b) or the Environmental Data Initiative repositories (Carey et al., 2023; Carey & Breef-Pilz, 2023). Instructions on reproducing the individual model forecasts as well as the multi-model ensemble are available in (Olsson et al., 2023b). In addition, the forecasts and scores can be accessed here to enable the manuscript figures to be reproduced (Olsson et al., 2023a).

3 Results

3.1 Observed and forecasted temperature dynamics at FCR

FCR exhibited typical seasonal dynamics during the two-year forecasting period. Continuous summer thermal stratification lasted from 11 March - 3 November 2021 and 31 March - 19 October 2022. Outside of these periods, there were transient periods of mixing and stratification during spring and autumn (Figure 2). Ice cover was observed intermittently during the periods of 11 January - 8 February 2022 and 23 December 2022 - 6 February 2023. As ice cover was short and intermittent, we hereafter refer to the period outside of the summer stratified period as ‘mixed,’ despite brief periods with inverse thermal profiles (Figure 2). Mean thermocline depth during the summer stratified period was 2.7 m in 2022 and 3.1 m in 2023.

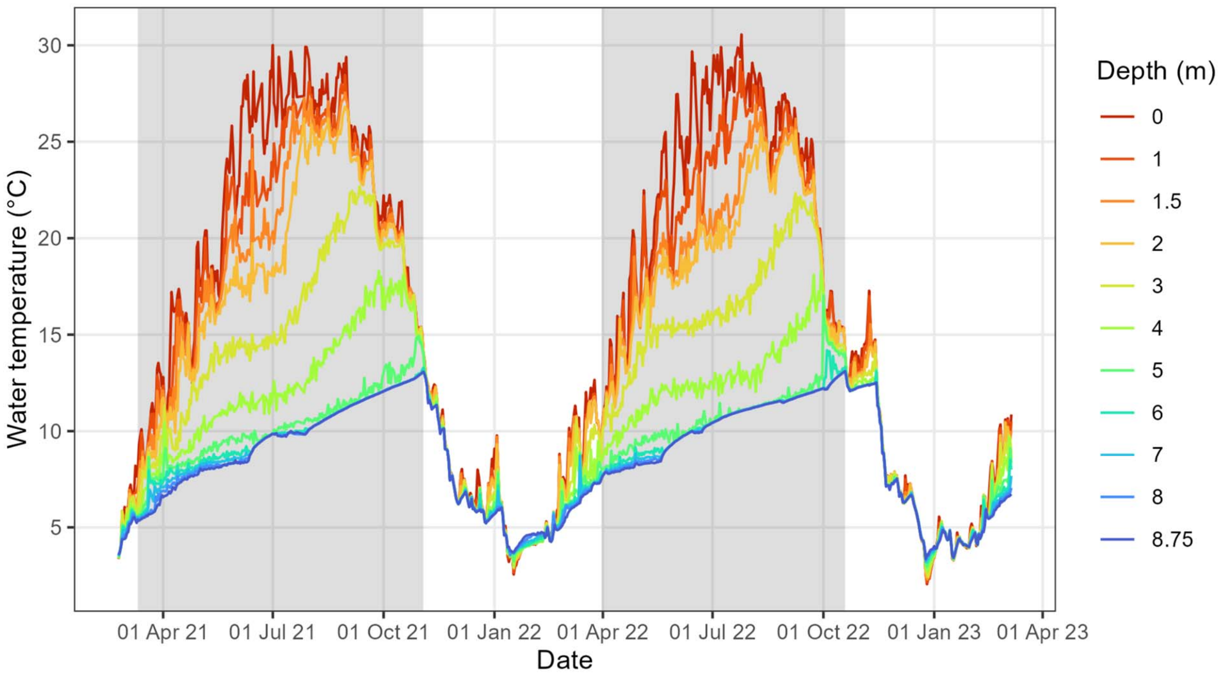

Figure 2. Observed high-frequency water temperatures across eleven depths at Falling Creek Reservoir from March 2021 to March 2023. The grey shaded areas show the periods of continuous summer stratification and white shaded areas show the mixed periods.

Our workflow (Figure 1) was able to successfully produce weekly 1-14 day-ahead forecasts for the two-year forecasting period for all five individual models and the two MMEs (Figure S3). In general, mean forecast performance was highest at the beginning of the forecast horizon and decreased further into the 14-day-horizon (Figure S3). Across all depths and horizons, the IGN score of the individual models (other than climatology) increased by 80-170% from 1 to 14 days-ahead, representing lower performance. Forecast uncertainty also increased across the 14-day horizon for all models except for climatology (by >100%).

Two examples highlight how the forecasts generated by the individual models exhibited differences in how well they reproduced observations across depths and times (Figure 3). First, forecasts generated during the mixed period at 1 m depth (20 February 2023) show that PM1 and PM3 forecasts closely followed observations throughout the 1 to 14-day ahead horizon, with PM2 diverging from observations after the 8th day of the forecast horizon. In contrast, the climatology and persistence baseline models consistently underestimated water temperature. In a second period with stratified water temperatures (1 August 2022), forecasts generated at 8 m depth show that PM3, and to a lesser extent PM1 and PM2, underestimated the water temperature. The two baseline models were most skillful for this particular forecast.

The variable performances of the individual models are reflected in the performance of the two MME forecasts (Figure 3). For example, in the first example forecasts for 20 February 2023 at 1 m depth, the PM MME performed better than the full MME because of the superior performance of the process models than the baseline models. Likewise, in the second example forecasts for 1 August 2022 at 8 m, the full MME performed better than the PM MME because of the strength of the baseline models.

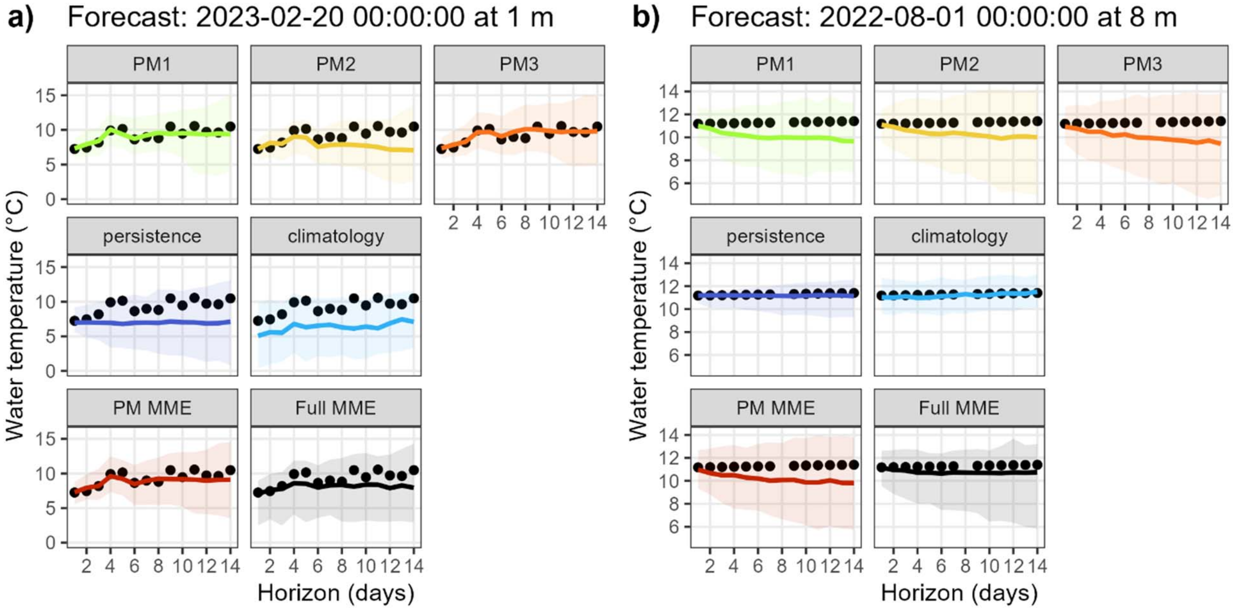

Figure 3. Two example water temperature forecasts from the five individual models and the two multi-model ensembles: one generated on 20 February 2023 at 1 m depth (mixed period; left panels) and one generated on 1 August 2022 at 8 m depth (stratified period; right panels). The top row shows the individual forecasts from the three process models (PMs), the middle row shows the individual forecasts from the two baseline models, and the bottom row shows the multi-model ensembles (PM and full MMEs). Shaded areas show the 95% confidence interval around the median forecast (line) and the filled points are the observed water temperatures. The colours for the different forecasts are consistent throughout.

3.2 Question 1: How does the performance of the process model MME compare to the individual process models?

Overall, the PM MME exhibited a higher aggregated performance, as determined by the lowest absolute bias and ignorance score, than the individual PMs (Table 1; Figure 4). When aggregated across all forecast dates, horizons, and depths, the bias of the PM MME was similar to PM1, highlighting how the addition of the other two PMs with slightly higher absolute bias did not increase bias in the MME (Table 1). The bias increased over the 1-14 day forecast horizon for all PMs and the PM MME, with bias increasing less for the PM MME and PM1 (Figure 4a). In contrast to bias, the PM MME had a larger mean forecast uncertainty (SD) than any of the individual models, especially at longer horizons (Figure 4b). SD increased over the forecast horizon at a faster rate for the PM MME than any individual PM (Figure 4b). At 1 day-ahead, the SD was similar for all PM forecasts (1.1°C), but by 14 days-ahead, the PM MME had 0.2°C higher SD than the best individual PM forecast (Fig. 4b).

Table 1. The mean ignorance score (IGN), absolute bias, standard deviation (SD), and shadowing time aggregated for all forecasts across all depths, times, and horizons for each forecast model, individual and multi-model ensemble (MME) across the two-year forecasting period. Models are sorted by most to least skillful, based on IGN, with the "best" forecast based on each metric in **bold**.

Forecast model	IGN	Absolute Bias (°C)	SD (°C)	Shadowing time (days)
Full MME	**1.52**	**0.69**	1.66	7.3
Climatology	1.63	0.98	1.45	3.2
PM1	1.67	0.95	1.47	4.5
Process model MME	1.68	0.94	1.62	4.2
PM2	1.80	1.16	1.49	3.7
PM3	1.81	1.09	1.53	4.0
Persistence	1.89	0.98	**1.26**	**7.9**

When using the IGN metric to evaluate performance, which combines accuracy and precision, the PM MME performance was similar but slightly lower than the performance of PM1 (Table 1). This result highlights the penalty given by the IGN score to the higher standard deviation in the PM MME. The best performing PM only placed 9% more probability on the observed outcome than the worst PM forecast on average, and 1% more probability than the PM MME (Equation 2). IGN increased over the forecast horizon at a similar rate for both the PM MME and most skillful individual forecast (PM1, Figure 4). At 14 days-ahead, the PM MME placed 13% more probability in the observed outcome and PM1 16% more probability than the two other individual forecast models. This change in probability demonstrates that the MME is not penalised strongly for including the "worse" models overall (Figure 4).

Using the shadowing time metric, the PM MME did not show increased ability to replicate observed water temperature dynamics relative to the individual models. The mean shadowing time for the PM MME (4.2 days) was slightly shorter than the best PM (4.5 days; Table 1). Shadowing time for the other PMs (PM2 and PM3) were shorter than the PM MME but all were between 3.7 and 4.5 days, less than half of the total forecast horizon.

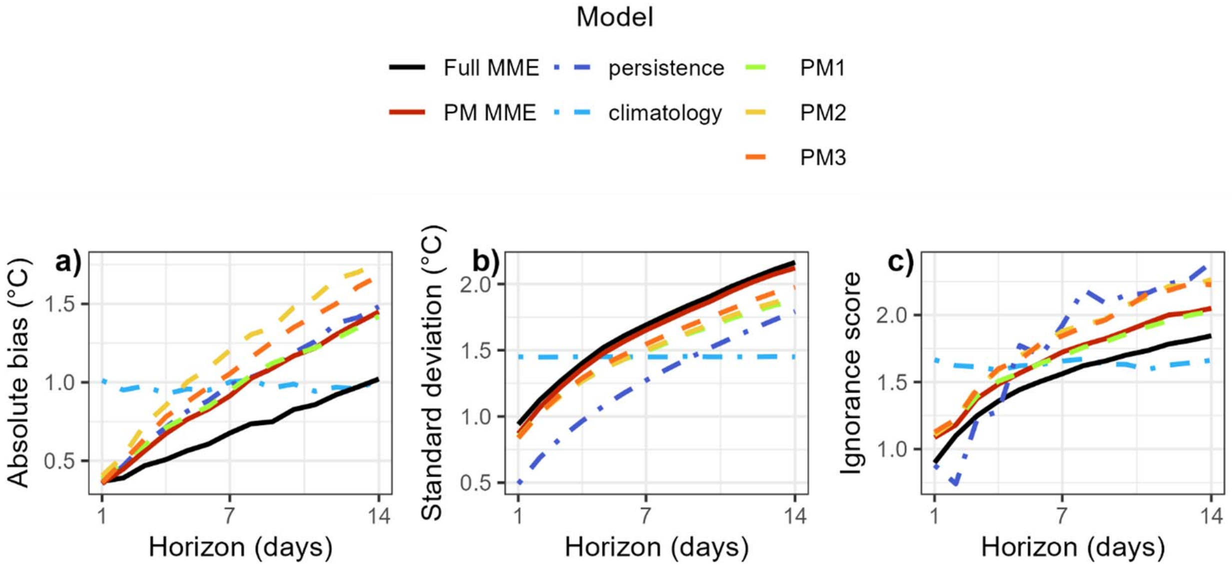

Figure 4. a) Mean absolute bias, b) standard deviation, and c) ignorance score across the 14-day forecast horizon for the three individual process models (PM), two baseline models (climatology and persistence), and the process model multi-model ensemble (PM MME) and full multi model ensemble (full MME).

3.3 Question 2: How does the addition of the baseline models into the full MEE affect forecast performance?

Altogether, the full MME had the lowest IGN, lowest bias, and highest standard deviation of any individual model or MME (Table 1). Aggregated across all depths, times of year, and horizons, the inclusion of the two baseline models into the full MME decreased bias by 26% but only increased the standard deviation by 2% (Table 1), relative to the PM MME. This large reduction in bias led to a lower IGN for the full MME (vs. the PM MME) despite the slight increase in uncertainty. Using the difference in the IGN metric, the full MME placed 12% more probability on the observed outcome than the PM MME. Overall, the improvement in performance of the full MME relative to the PM MME increased throughout the forecast horizon (Figure 4a), a 6% improvement at 2 days-ahead compared to 15% at 14 days-ahead.

The shadowing time of the full MME (7.3 days) was longer than the PM MME (4.2 days; Table 1). This improvement in shadowing time was due to the inclusion of the persistence model in the full MME. The persistence model had the longest shadowing time of any individual model or MME (7.9 days).

The individual PM forecasts exhibited high covariance with other PM forecasts and low covariance with the baseline model forecasts (Figure 5). At 1 m, the PM models exhibited strong positive correlations at 1, 7, and 14 day-ahead horizons (r = 0.73 to 0.96), with PM1 and PM2 being most correlated at these three horizons. In contrast to the individual PM models, the individual baseline models generally showed low covariance between each other and with the PM models (Figure 5). A few exceptions to this pattern were at 1 day-ahead, when the persistence model showed a moderate correlation with PM3 (r = 0.45) at 1 m. Similarly, at 7 and 14 days-ahead, the persistence and climatology showed a positive correlation (r = 0.58 and r = 0.60, respectively), and at 14 days-ahead the climatology model was positively correlated with all other models. The correlations among the PMs were always higher than any correlation involving a baseline model (Figure 5).

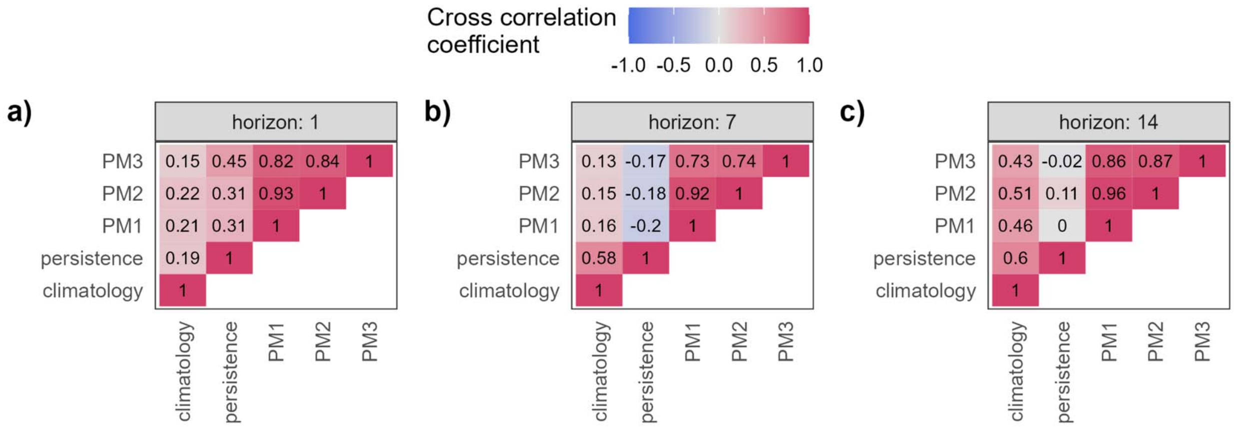

Figure 5. Correlation (Pearson r) of bias among individual model forecasts. The correlation coefficient between models was calculated for the mean forecast bias (mean - observations) at 1, 7, and 14 day-ahead horizons for 1 m. Red indicates a strong positive correlation and blue indicates a strong negative correlation.

3.4 Question 3: How does the forecast performance of the individual models and MMEs vary across horizons and depths?

The ranking of models demonstrates the hedging that occurs when using MMEs to forecast at different depths and horizons. The individual model forecasts were more likely to be ranked the ‘worst’ of the seven forecasts (Figure 6, Figure S4) than the two MME forecasts. Out of all n=104 forecasts generated, the full MME had <1% of rank 7 (worst) forecasts across 1 and 8 m (n=1 forecast) and 10% of rank 1 (best) forecasts (n=11). At 1 m, the full MME was most often ranked in the middle (65-95% of forecasts ranked 3-5, respectively; Figure 6a). At 8 m, the full MME was more often ranked the second-best forecast, especially at shorter horizons (Figure

6b), with more than 50% of forecasts at ranks 1 or 2 up to 4 days-ahead. Despite the decrease in high-ranking forecasts (ranks 1-2) at longer horizons, there was no appreciable increase in the proportion of worst-ranking forecasts (ranks 6-7), remaining between 2-6% of forecasts at most horizons.

The individual PM forecasts were dominated by rankings of either the best (1) or worst (7) performance, whereas the PM MME had fewer of these extreme ranks. At 1 m, the individual PM models had almost equal proportions of rank 1 and rank 7 forecasts across the full horizon (Figure 6i,k,m), with over 40% of forecasts ranked at one of these extremes, compared to only 2% of the PM MME forecasts exhibiting one of these extreme ranks. At 8 m, the individual PM models were more often at an intermediate rank than at 1 m (Figure 6j,l,n), although PM2 and PM3 had more than 40% of the worst forecast, whereas PM1 had up to 56% of forecasts with an intermediate rank and fewer very poor forecasts (rank=7).

The ranks of the baseline models varied substantially at different depths and horizons. At 1 m, the persistence model had more than 50% of forecasts in rank 1 for 1 day-ahead forecasts, which declined steeply to only 10% at horizons >5 days-ahead (Figure 6e). Concurrently, the proportion of forecasts for which persistence was the worst forecast also increased across the forecast horizon, with more than 50% of the forecasts having persistence at ranks 6 or 7 forecast at 13-14 days ahead (Figure 6e). At 8 m, the persistence forecasts dominated the best performing rank across the whole horizon (Figure 6f), only decreasing marginally from around 80% to 65% of total forecasts by 14 days-ahead (Figure S4). The climatology model demonstrated strengths at longer horizons at both 1 and 8 m. The proportion of climatology forecasts at 1 m with a rank 1 increased across the forecast horizon, from <5% at 1 day-ahead to 26% of forecasts at 14 days-ahead (Figure 6g). However, climatology was frequently the least skillful forecast at 1 m, especially at 1 day-ahead (Figure S4; 65% of forecasts). At horizons between 3 and 10 days-ahead, 40% of the climatology forecasts were either the first or second ranked forecast, which increased to 80% at horizons >10 days-ahead.

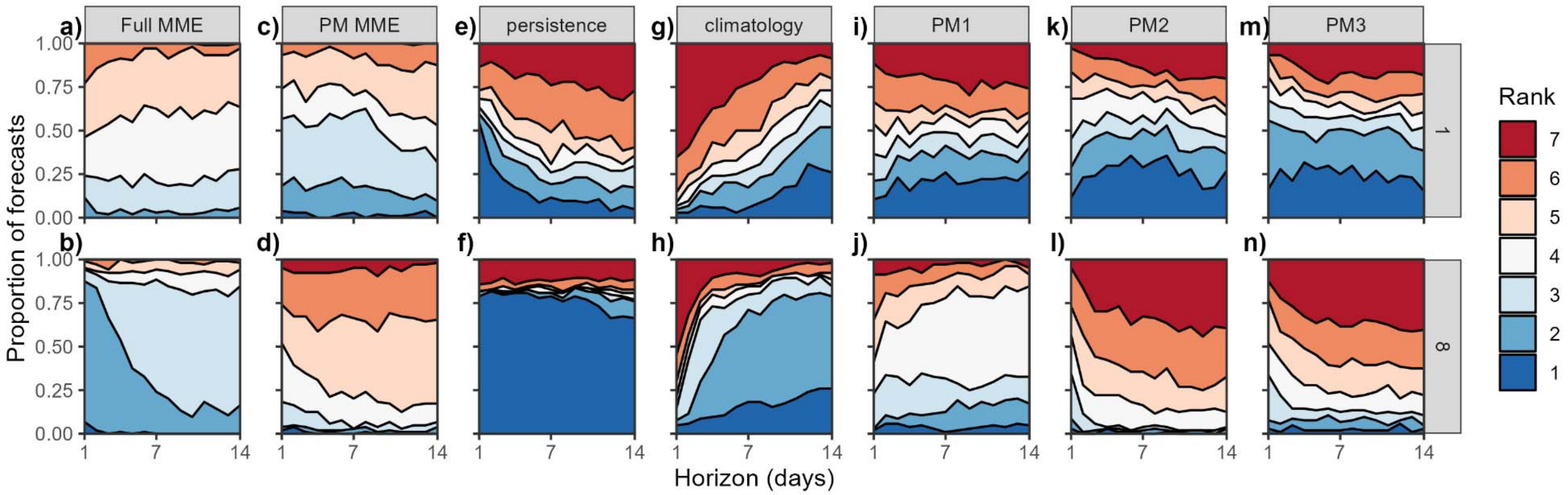

Figure 6. Proportion of total forecasts (n = 104) with each rank, from 1 (best) to 7 (worst), out of the five individual models and two multi-model ensembles (process model (PM) and full MMEs). Ranks were calculated for each individual forecast (n = 104) and each horizon (1 to 14 day-ahead) based on the ignorance forecast metric at 1 m (top row) and 8 m (bottom row) depths.

Inspection of the disaggregated forecast scores further demonstrates that there was no one consistently best-performing model or MME at all horizons and all depths, as determined by IGN scores and shadowing times (Figure 7a,b). At 1 m, the two MME forecasts had the highest skill across the total horizon (Table S2), although they were outperformed at certain horizons by the individual PM2 model and, beyond 10 days-ahead, climatology (Figure 7a). Conversely, at 8 m, the persistence model had the best performance for 2 days-ahead, the full MME exhibited the best performance 1 and 3-5 days-ahead, and then the climatology model had the highest skill up to 14 days-ahead (Figure 7b).

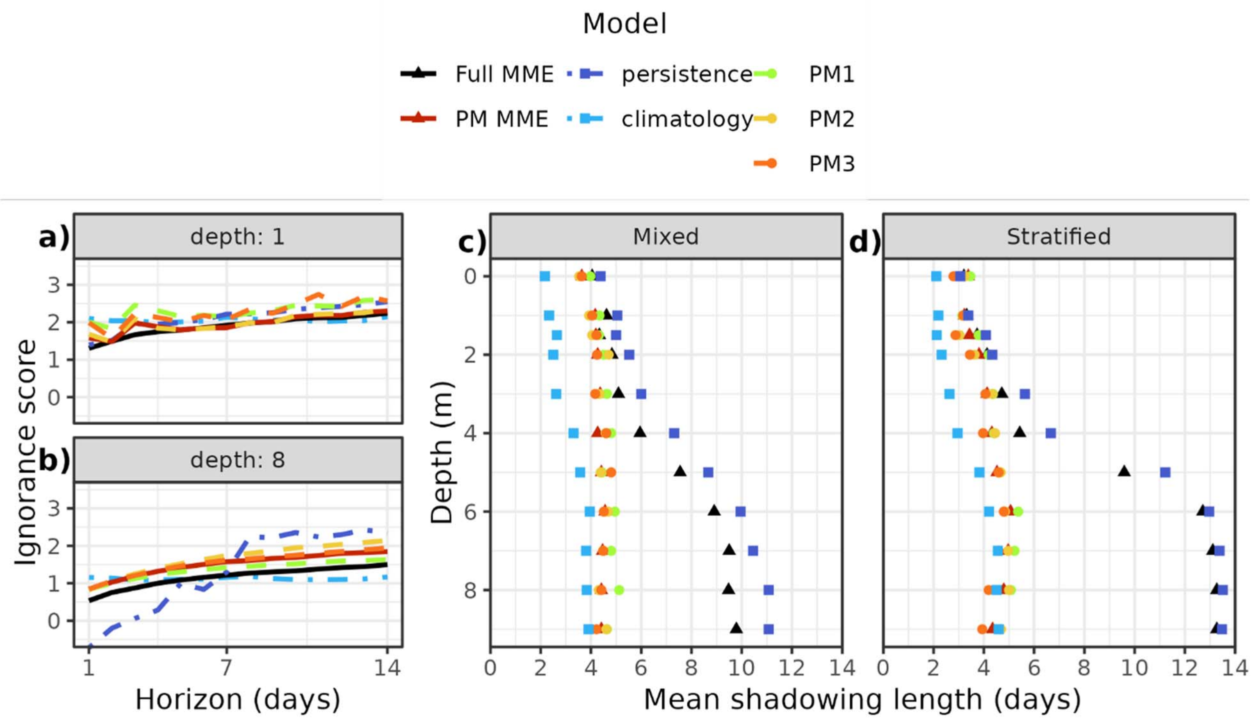

Figure 7. Disaggregated forecast performance (ignorance score) at 1 (a) and 8 m (b) for each and mean shadowing time at each observed depth in the water column in the mixed (c) and stratified periods (d) for the three individual process models (PM), two baseline models (climatology and persistence), and the process model (PM) MME and full MME.

As with the aggregated shadowing times, including the baseline models in the full MME extended the shadowing time compared to the PM MME at almost all depths during both the stratified and mixed periods (Figure 7c, d). The persistence model had the longest shadowing time across all forecasts (mean = 7.9 days, Table 1), which was consistent across depths, except for forecasts at the surface (0 m) during the stratified period (Figure 7d). The persistence model showed significantly better shadowing ability than the other individual model forecasts, especially at depths deeper than 4 m, which corresponded to depths below the thermocline, calculated at a depth of 2.7-3.1 m during the forecast period. For example, at 5 m, the shadowing time of the persistence forecast during the stratified period was 2.5 times longer than the next best individual model (PM1). The shadowing time of the PM MME did not improve on the best individual model (PM1), although all PM showed low shadowing ability (<6 days at all depths) relative to the persistence and full MME. At 8 m, both the persistence model and the full MME were able to almost shadow the full horizon (Figure 7d; mean shadowing times = 13.5 and 13.2 days, respectively).

4 Discussion

Reservoir water temperature forecasts generated using a multi-model ensemble (MME) consisting of process and baseline models performed better overall than using individual models or a process model (PM) MME. Our results support previous research that shows that MME methods often outperform individual models (Atiya, 2020; Johansson et al., 2019; Viboud et al., 2018). For example, in a large diverse forecasting competition of multiple finance and demography variables, 70% of the most accurate forecasts were MMEs (Atiya, 2020). Our results showed that no individual model performed best at all depths and horizons, as the best models at 1 m (the individual process models) were the worst performers at 8 m. In contrast to this finding, the full MME was rarely the worst-performing forecast, highlighting the hedging ability of MMEs to prevent very poor forecast performance (Atiya, 2020). MMEs incorporate the strengths of multiple models given that all models are likely imperfect representations of reality (Atiya, 2020) as well as acknowledging the between-model uncertainty (Humphries et al., 2018). Below, we examine some of the implications for using MME forecasts and highlight ways to further improve MME forecasts for other applications.

4.1 No one individual model is optimal for all forecast horizons or depths

For individual 1-14 day-ahead forecasts at specific horizons and depths, individual models outperformed the MMEs (Figure 7), accounting for >96% of the best forecasts at 1 m and >91% at 8 m (Figure S4). Each model captures slightly different dynamics of the mechanistic processes controlling reservoir water temperature and therefore performed optimally under different conditions (Lapeyrolerie & Boettiger, 2023). This was also observed in a multi-model river forecasting study in which individual models alternately performed best in predicting different stages, phases, or mechanisms of rainfall-runoff (Abrahart & See, 2002) and a penguin population forecasting study in which a range of models differentially captured inter-annual and inter-species variability (Humphries et al., 2018). Altogether, our study contributes to the evidence that combining forecasts from different models provides a more comprehensive and accurate representation of the forecasted system than one model alone.

In our analysis, the optimal model varied by depth and horizon, demonstrating the individual strengths of each model. The persistence forecast was significantly better across all horizons at 8 m than other models (ranked best in 66 - 82% of all forecasts, Figure 6e), but generally performed poorly at 1 m at horizons beyond 1 to 2 days-ahead (Figure 7). This finding is in agreement with a previous water temperature forecast study at the same reservoir, which found high forecast skill from a persistence model deeper in the lake and higher skill from a PM at the surface (Thomas et al., 2020). Individual PMs have been shown to be successful at forecasting water temperature dynamics at the lake surface at short horizons (Thomas et al., 2020; Wander et al., 2023). As weather forecast skill degrades further into the future, there is a subsequent reduction in water temperature forecasting skill at these shallower depths at longer horizons (Carey, Woelmer, et al., 2022; Thomas et al., 2020). This pattern is likely because meteorological driver data uncertainty has been shown to be the primary source of uncertainty in surface water temperature forecasts, due to the sensitivity of surface water temperatures to atmospheric forcing (Thomas et al., 2020).

One promising approach for better utilising the strengths of the individual models is to weight the individual models within the MME based on their historical forecast performance. Weighting the individual models may further increase MME skill (reviewed by Wang et al., 2022), as these methods seek to exploit the inherent benefits of each individual model represented in the MME (Abrahart & See, 2002). MME blending methods that weight accurate models more highly and adjust weights dynamically may leverage the strengths of the models whilst minimising their weaknesses (Chandler, 2013; Spence et al., 2018). For example, Abrahart & See (2002) used a fuzzy logic approach to use the previous forecast performance to weight the models used in the next forecast MME when forecasting river flow. However, selecting the optimal MME blending method was dependent on the dynamics of the flow conditions (Abrahart & See, 2002). Wang et al. (2022) note in their review that simple combination methods, such as the linear pooling with equal weights (as done here) or simple averaging, are some of the most robust approaches for model blending and that improvements from optimised weights can be outweighed by the error added by estimating these parameter values (Dormann et al., 2018). In short, estimating the weighting parameters adds another source of uncertainty to the forecasts whereas simple averaging is robust and easier to implement (Barrow & Kourentzes, 2016). A potential alternative to applying a weighting method would be to identify a suitable pool of models to use in the MME and omit the worst performing ones, thus diminishing the worst predictions within each individual model forecast (Abrahart & See, 2002; Dormann et al., 2018), unless it is very diverse from the other models (Atiya, 2020).

4.2 The PM MME did not significantly improve on the best individual process model

When aggregating the ignorance score across all forecasts, the PM MME performed slightly worse than the best individual PM model. However, the PM MME had many fewer individual forecasts when it was ranked as the least skillful model (Figure 6). This result demonstrates the value of hedging through MMEs. Even when the aggregate forecast skill of the PM MME is not significantly improved compared to its individual models, the process MME still provides value by preventing the generation of poorly-performing forecasts that can occur from individual models (Doblas-Reyes et al., 2005; Hagedorn et al., 2005).

Overall, the performance of the individual PMs was highly positively correlated (Figure 5), limiting the amount of unique information provided by individual models to the MME. Others have found that MME forecasts were most skillful when the covariance among models was low (Dormann et al., 2018; Renwick et al., 2018), as well as when models exhibit diverging bias in their mean predictions (Dormann et al., 2018; Petropoulos et al., 2022). This finding supports the need for more diverse model structures to fully optimise the MME forecasts. In this study, high covariance among PMs was likely caused by three key drivers. First, the three process models were all 1-D hydrodynamic models. Examining whether adding more complex process models (e.g., 3-D models) or simpler process models (e.g., Hanson et al., 2023) could help reduce inter-model covariance is another opportunity for further research. Second, the three PMs all used the same forecasted weather from the NOAA Global Ensemble Forecasting System as driver data. Future work could include models that use alternative weather drivers, such as different weather forecast products (e.g. Buizza & Richardson, 2017) or historical weather climatology. Third, all three PMs applied the same data assimilation algorithm (an ensemble Kalman filter). Future work could explore the influence of the diversity of data assimilation approaches on MME forecasts by including alternative data assimilation approaches, such as a particle filter (Fearnhead & Künsch, 2018).

4.3 Including baseline models in the MME improved forecast skill

Results from the full MME demonstrate that more model diversity within an MME increases forecast skill (Figure 4c; Table 1). The most model diversity was added to the MME by including the two baseline models that represent end members of empirical models. Specifically, the persistence model represents the most recent data and climatology represents the long-term historical average for the forecasted system. By including these baseline empirical models, water temperature forecast performance was substantially increased compared to the PM MME (Table 1).

Including baseline models in an MME presents a relatively easy approach with low computational costs to improve forecast performance if data are readily available for constructing the baseline models. While many forecasting studies use baseline models as null models to evaluate forecasts, here we show their value as a component of the forecast themselves. These baseline models, despite their simplicity, provide additional forecast information that the complex process models do not, and highlight that model complexity does not necessarily translate to forecast skill (Viboud et al., 2018; Ward et al., 2014). Even simple models, lacking any domain expertise, can provide useful information to an MME (Wang et al., 2022). For example, forecasting of penguin populations showed that simpler domain-agnostic time series models produced better forecasts than complex domain-specific population models (Humphries et al., 2018).

4.4 Recommendations and next steps

Identifying a set of models with low covariance is likely to increase aggregated forecast skill from an MME relative to its individual models. In advance of producing an MME forecast, a model selection process would help ensure that the MME will improve skill relative to individual models, based on among-model covariance and individual model variance and bias (Dormann et al., 2018; Hagedorn et al., 2005). It is likely that the optimal set of models to include in the MME will be specific to individual sites, given how individual models perform differently among lakes (e.g., Bruce et al., 2018). For example, the same forecast model performed better at some lakes than others in a multi-site comparison (Thomas et al., 2023), with similar differences in model performance found among sites when forecasting phytoplankton (Page et al., 2018; Rousso et al., 2020).

Further ways to improve forecast skill should also focus on constraining uncertainty. The full MME had the highest variance of any of the forecast models, which undermines some of the improvement in bias from the model averaging and leaves a forecast that is likely underconfident (Wang et al., 2022). Methods such as boosting, dimensionality reduction, and trimming can optimise bias-variance trade-offs (Wang et al., 2022). For example, trimming the tails (exterior) of the individual forecast distributions has been shown to increase confidence in the MME by reducing the variance of the individual model forecasts before being combined into an MME (Howerton et al., 2023; Zhao et al., 2022). Previous results showed that MMEs were more successful when their component model forecasts were overconfident (low variance) (Hagedorn et al., 2005; Wang et al., 2022; Weigel et al., 2008).

Finally, our results demonstrate the value of calculating multiple evaluation metrics when assessing the skill of forecasting methods, as each metric highlights potential areas to improve overall skill. For example, the forecast standard deviation evaluation showed that uncertainty was much larger for the MMEs than any individual model (Figure 7c, d). Simultaneously, the MMEs had the lowest bias (Figure 7a, b). The IGN score was able to combine these two evaluation components into a single metric of statistical performance, highlighting that improvements in overall performance would likely come from reducing forecast uncertainty. Although shadowing time is a metric infrequently used in freshwater forecast evaluation (Lofton et al., 2023), it is potentially valuable, given its focus on the model's ability to replicate actual dynamics, rather than just the statistics of the forecast (Gilmour & Smith, 1997; Petropoulos et al., 2022) or the shape of the distribution (Smith et al., 2015), providing information on likely lead times at which a forecast will have utility (Smith et al., 2010). Improving the capacity of the PMs to have longer shadowing times may help improve their overall representation of lake and reservoir dynamics.

5 Conclusions

This work has demonstrated the usefulness of multi-model ensembles in improving water temperature forecasts. A five-model MME had the highest forecast skill among all of the forecasts generated by individual models or a three-model MME, which is likely due to hedging: the five-model MME was able to avoid generating very bad forecasts despite being unable to provide the most skillful forecast at many individual horizons or depths. The addition of two baseline models, which had low covariance with the PM models, into the MME provided useful shadowing ability and complementary forecast information. Our results present an example of how existing models can be combined to improve water temperature forecasting in lakes and reservoirs. Future work could focus on including additional forecasting model structures to further increase the diversity of predictions included in the MME and investigate optimal methods to blend predictions and constrain model variance. Altogether, we highlight the value of including simple baseline models (which may in some cases be already calculated as null models for forecast evaluation) into multi-model ensembles for forecasting to improve forecasting skill effectively and efficiently with little additional effort.

Acknowledgments

This work was supported by US National Science Foundation grants DEB-1926388, DBI-1933016, DBI-1933102, and DEB-2320730.

Data Availability Statement

All data and code to produce the forecasts, scores, and figures presented in this manuscript are available in in the Zenodo repositories (Olsson et al., 2023a, 2023b) or the Environmental Data Initiative repositories (Carey et al., 2023; Carey & Breef-Pilz, 2023).

Author contribution statement

RQT and CCC developed the FLARE forecasting framework. RQT and TNM developed the FLARE-LER methodology and adapted the lake models for the FLARE framework. FO led the baseline model forecast generation, ran the forecasts, updated individual process-model parameterization, developed the forecast evaluation framework, and analysed the forecasts with RQT. ABP oversaw sensor data collection and field sampling. FO led manuscript writing with RQT and CCC; all authors reviewed and approved the final version.

References

Abrahart, R. J., & See, L. (2002). Multi-model data fusion for river flow forecasting: An evaluation of six alternative methods based on two contrasting catchments. *Hydrology and Earth System Sciences*, *6*(4), 655–670. https://doi.org/10.5194/hess-6-655-2002

Almeida, M. C., Shevchuk, Y., Kirillin, G., Soares, P. M., Cardoso, R. M. A. de P., Matos, J. P., et al. (2022). Modeling reservoir surface temperatures for regional and global climate models: a multi-model study on the inflow and level variation effects. *Geoscientific Model Development*, (15), 137–197. https://doi.org/10.5194/gmd-2021-64

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, *36*(1), 197–200. https://doi.org/10.1016/j.ijforecast.2019.03.010

Baracchini, T., Wüest, A., & Bouffard, D. (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

Barrow, D. K., & Kourentzes, N. (2016). Distributions of forecasting errors of forecast combinations: Implications for inventory management. *International Journal of Production Economics*, *177*, 24–33. https://doi.org/10.1016/j.ijpe.2016.03.017

Boettiger, C. (2022). The forecast trap. *Ecology Letters*, *25*(7), 1655–1664. https://doi.org/10.1111/ele.14024

Bradford, J. B., Weltzin, J. F., McCormick, M., Baron, J., Bowen, Z., Bristol, S., et al. (2020). *Ecological Forecasting—21st Century Science for 21st Century Management*. *U.S. Geological Survey Open-File Report 2020-1073*. https://doi.org/https://doi.org/10.3133/ofr20201073

Bruce, L. C., Frassl, M. A., Arhonditsis, G. B., Gal, G., Hamilton, D. P., Hanson, P. C., et al. (2018). A multi-lake comparative analysis of the General Lake Model (GLM): Stress-testing across a global observatory network. *Environmental Modelling & Software*, *102*, 274–291. https://doi.org/10.1016/J.ENVSOFT.2017.11.016

Buizza, R., & Richardson, D. (2017). 25 Years of ensemble prediction. *ECMWF Newsletter*, *153*, 20–31. https://doi.org/10.21957/bv418o

Carey, C. C., & Breef-Pilz, A. (2023). Ice cover data for Falling Creek Reservoir and Beaverdam Reservoir, Vinton, Virginia, USA for 2013-2023 ver. 1. Environmental Data Initiative. Retrieved from https://portal-s.edirepository.org/nis/mapbrowse?scope=edi&identifier=1076&revision=1

Carey, C. C., Ibelings, B. W., Hoffmann, E. P., Hamilton, D. P., & Brookes, J. D. (2012). Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate. *Water Research*, *46*(5), 1394–1407. https://doi.org/10.1016/j.watres.2011.12.016

Carey, C. C., Woelmer, W. M., Lofton, M. E., Figueiredo, R. J., Bookout, B. J., Corrigan, R. S., et al. (2022). Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. *Inland Waters*, *12*(1), 107–120. https://doi.org/10.1080/20442041.2020.1816421

Carey, C. C., Lewis, A. S. L., Howard, D. W., Woelmer, W. M., Gantzer, P. A., Bierlein, K. A., et al. (2022). Bathymetry and watershed area for Falling Creek Reservoir, Beaverdam Reservoir, and Carvins Cove Reservoir. Environmental Data Initiative. https://doi.org/https://doi.org/10.6073/pasta/352735344150f7e77d2bc18b69a22412

Carey, C. C., Breef-Pilz, A., & Woelmer, W. M. (2023). Time series of high-frequency sensor data measuring water temperature, dissolved oxygen, pressure, conductivity, specific conductance, total dissolved solids, chlorophyll a, phycocyanin, fluorescent dissolved organic matter, and turbidity at discrete dept. Environmental Data Initiative. https://doi.org/https://doi.org/10.6073/pasta/f6bb4f5f602060dec6652ff8eb555082

Chandler, R. E. (2013). Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1991), 20120388. https://doi.org/10.1098/rsta.2012.0388

Clark, N. J., Proboste, T., Weerasinghe, G., & Magalhães, R. J. S. (2022). Near-term forecasting of companion animal tick paralysis incidence: An iterative ensemble model. *PLoS Computational Biology*, *18*(2), e1009874. https://doi.org/10.1371/journal.pcbi.1009874

Clayer, F., Jackson-Blake, L., Mercado-Bettín, D., Shikhani, M., French, A., Moore, T., et al.

(2023). Sources of skill in lake temperature, discharge and ice-off seasonal forecasting tools. *Hydrology and Earth System Sciences*, *27*(6), 1361–1381. https://doi.org/10.5194/hess-27-1361-2023

Daneshmand, V., Breef-Pilz, A., Carey, C. C., Jin, Y., Ku, Y. J., Subratie, K. C., et al. (2021). Edge-to-cloud virtualized cyberinfrastructure for near real-time water quality forecasting in lakes and reservoirs. *Proceedings - IEEE 17th International Conference on EScience, EScience 2021*, 138–148. https://doi.org/10.1109/ESCIENCE51609.2021.00024

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., et al. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, *115*(7), 1424–1432. https://doi.org/10.1073/pnas.1710231115

Doblas-Reyes, F. J., Hagedorn, R., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 234–252. https://doi.org/10.3402/tellusa.v57i3.14658

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., et al. (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*(4), 485–504. https://doi.org/10.1002/ecm.1309

Evensen, G. (2003). The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. https://doi.org/10.1007/s10236-003-0036-9

Fearnhead, P., & Künsch, H. R. (2018). Particle filters and data assimilation. *Annual Review of Statistics and Its Application*, *5*, 421–449. https://doi.org/10.1146/annurev-statistics-031017-100232

Feldbauer, J., Ladwig, R., Mesman, J. P., Moore, T. N., Zündorf, H., Berendonk, T. U., & Petzoldt, T. (2022). Ensemble of models shows coherent response of a reservoir's stratification and ice cover to climate warming. *Aquatic Sciences*, *84*, 50. https://doi.org/10.1007/s00027-022-00883-2

La Fuente, S., Jennings, E., Gal, G., Kirillin, G., Shatwell, T., Ladwig, R., et al. (2022). Multi-model projections of future evaporation in a sub-tropical lake. *Journal of Hydrology*, *615*, 128729. https://doi.org/10.1016/j.jhydrol.2022.128729

Gilmour, I., & Smith, L. A. (1997). Enlightenment in shadows. In *Nonlinear Dynamics and Stochastic Systems Near the Millenium* (pp. 335–340). AIP. https://doi.org/1-56396-736-7/97/

Golub, M., Thiery, W., Marcé, R., Pierson, D., Vanderkelen, I., Mercado-Bettin, D., et al. (2022). A framework for ensemble modelling of climate change impacts on lakes worldwide: the ISIMIP Lake Sector. *Geoscientific Model Development*, *15*(11), 4597–4623. https://doi.org/10.5194/gmd-15-4597-2022

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *14*(1), 107–114.

Goudsmit, G.-H., Burchard, H., Peeters, F., & Wüest, A. (2002). Application of k-ϵ turbulence models to enclosed basins: The role of internal seiches. *Journal of Geophysical Research: Oceans*, *107*(C12), 3230. https://doi.org/10.1029/2001JC000954

Graf, R., Zhu, S., & Sivakumar, B. (2019). Forecasting river water temperature time series using a wavelet–neural network hybrid modelling approach. *Journal of Hydrology*, *578*, 124115. https://doi.org/10.1016/j.jhydrol.2019.124115

Hagedorn, R., Doblas-Reyes, F. J., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 219. https://doi.org/10.3402/tellusa.v57i3.14657

Hamill, T. M., Whitaker, J. S., Shlyaeva, A., Bates, G., Fredrick, S., Pegion, P., et al. (2022). The Reanalysis for the Global Ensemble Forecast System, Version 12. *Monthly Weather Review*, *150*(1), 59–79. https://doi.org/10.1175/MWR-D-21-0023.1

Hanson, P. C., Ladwig, R., Buelo, C., Albright, E. A., Delany, A. D., Carey, C., et al. (2023). Legacy phosphorus and ecosystem memory control future water quality in a eutrophic lake. *Preprint*. https://doi.org/10.22541/essoar.168677211.11983579/v1

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., 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*(1), 473–523. https://doi.org/10.5194/gmd-12-473-2019

Howerton, E., Runge, M. C., Bogich, T. L., Borchering, R. K., Inamine, H., Lessler, J., 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. https://doi.org/10.1098/rsif.2022.0659

Huang, B., Langpap, C., & Adams, R. M. (2011). Using instream water temperature forecasts for fisheries management: An application in the pacific northwest. *Journal of the American Water Resources Association*, *47*(4), 861–876. https://doi.org/10.1111/j.1752-1688.2011.00562.x

Humphries, G. R. W., Che-Castaldo, C., Bull, P. J., Lipstein, G., Ravia, A., Carrión, B., et al. (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., & Athanasopoulos, G. (2021). Forecasting: principles and practice. Retrieved June 28, 2023, from OTexts.com/fpp3

IPCC (Intergovernmental Panel on Climate Change). (2023). *Technical Summary*. *Climate Change 2021 – The Physical Science Basis*. Cambridge University Press. https://doi.org/10.1017/9781009157896.002

Jackson-Blake, L. A., Clayer, F., Haande, S., Sample, J. E., & Moe, S. J. (2022). Seasonal forecasting of lake water quality and algal bloom risk using a continuous Gaussian Bayesian network. *Hydrology and Earth System Sciences*, *26*(12), 3103–3124. https://doi.org/10.5194/hess-26-3103-2022

Johansson, M. A., Apfeldorf, K. M., Dobson, S., Devita, J., Buczak, A. L., Baugher, B., 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*(48), 24268–24274. https://doi.org/10.1073/pnas.1909865116

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: a practitioner’s guide in* *atmospheric science*. (I. T. Jolliffe & D. B. Stephenson, Eds.) (2nd ed.). Oxford: Wiley Blackwell.

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, *90*(12), 1–37. https://doi.org/10.18637/jss.v090.i12

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., et al. (2014). The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Kraemer, B. M., Chandra, S., Dell, A. I., Dix, M., Kuusisto, E., Livingstone, D. M., et al. (2017). Global patterns in lake ecosystem responses to warming based on the temperature dependence of metabolism. *Global Change Biology*, *23*(5), 1881–1890. https://doi.org/10.1111/gcb.13459

Lapeyrolerie, M., & Boettiger, C. (2023). Limits to ecological forecasting: Estimating uncertainty for critical transitions with deep learning. *Methods in Ecology and Evolution*, *14*(3), 785–798. https://doi.org/10.1111/2041-210X.14013

Lewis, A. S. L. L., Woelmer, W. M., Wander, H. L., Howard, D. W., Smith, J. W., McClure, R. P., et al. (2022). Increased adoption of best practices in ecological forecasting enables comparisons of forecastability. *Ecological Applications*, *32*(2), e02500. https://doi.org/10.1002/eap.2500

Lofton, M. E., Howard, D. W., Thomas, R. Q., & Carey, C. C. (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

Long, X., Widlansky, M. J., Spillman, C. M., Kumar, A., Balmaseda, M., Thompson, P. R., et al. (2021). Seasonal Forecasting Skill of Sea-Level Anomalies in a Multi-Model Prediction Framework. *Journal of Geophysical Research: Oceans*, *126*, e2020JC017060. https://doi.org/10.1029/2020JC017060

Machete, R. L., & Smith, L. A. (2016). Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus A: Dynamic Meteorology and Oceanography*, *68*(1), 28393. https://doi.org/10.3402/tellusa.v68.28393

Mercado-Bettín, D., Clayer, F., Shikhani, M., Moore, T. N., Frías, M. D., Jackson-Blake, L., et al. (2021). Forecasting water temperature in lakes and reservoirs using seasonal climate prediction. *Water Research*, *201*, 117286. https://doi.org/10.1016/j.watres.2021.117286

Mironov, D. V. (2021). *Parameterization of Lakes in Numerical Weather Prediction Description of a Lake Model*. *COSMO Technical Report*. Offenbach am Main, Germany. https://doi.org/10.1007/978-3-030-58292-0_60177

Moore, T. N., Mesman, J. P., Ladwig, R., Feldbauer, J., Olsson, F., Pilla, R. M., et al. (2021). LakeEnsemblR: An R package that facilitates ensemble modelling of lakes. *Environmental Modelling & Software*, *143*, 105101. https://doi.org/10.1016/j.envsoft.2021.105101

O'Hara-Wild, M., Hyndman, R., & Wang, E. (2022). fable: Forecasting Models for Tidy Time Series. R package version 0.3.2. Retrieved from https://cran.r-project.org/package=fable

Olsson, F., Moore, T. N., Carey, C. C., Breef-Pilz, A., & Thomas, R. Q. (2023a). A multi-model ensemble of baseline and process-based models improves the predictive skill of near-term lake forecasts: data, forecasts, and scores [dataset]. Zenodo. https://doi.org/10.5281/zenodo.8136960

Olsson, F., Moore, T., Carey, C. C., Breef-Pilz, A., & Thomas, R. Q. (2023b). OlssonF/FCRE-forecast-code: A multi-model ensemble of baseline and process-based models improves the predictive skill of near-term lake forecasts: code (v1.0.0). Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.8172783

Page, T., Smith, P. J., Beven, K. J., Jones, I. D., Elliott, J. A., Maberly, S. C., et al. (2018). Adaptive forecasting of phytoplankton communities. *Water Research*, *134*, 74–85. https://doi.org/10.1016/j.watres.2018.01.046

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., et al. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, *522*, 697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

R Core Team. (2021). R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-Guided Deep Learning predictions of lake water temperature. *Water Resources Research*, *55*(11), 9173–9190. https://doi.org/10.1029/2019WR024922

Renwick, K. M., Curtis, C., Kleinhesselink, A. R., Schlaepfer, D., Bradley, B. A., Aldridge, C. L., et al. (2018). Multi-model comparison highlights consistency in predicted effect of warming on a semi-arid shrub. *Global Change Biology*, *24*(1), 424–438. https://doi.org/10.1111/gcb.13900

Rousso, B. Z., Bertone, E., Stewart, R., & Hamilton, D. P. (2020). A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Research*, *182*, 115959. https://doi.org/10.1016/j.watres.2020.115959

Saloranta, T. M., & Andersen, T. (2007). MyLake—A multi-year lake simulation model code suitable for uncertainty and sensitivity analysis simulations. *Ecological Modelling*, *207*(1), 45–60. https://doi.org/10.1016/j.ecolmodel.2007.03.018

Smith, L. A., Cuéllar, M. C., Du, H., & Judd, K. (2010). Exploiting dynamical coherence: A geometric approach to parameter estimation in nonlinear models. *Physics Letters, Section A: General, Atomic and Solid State Physics*, *374*(26), 2618–2623. https://doi.org/10.1016/j.physleta.2010.04.032

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T., & Du, H. (2015). Towards improving the framework for probabilistic forecast evaluation. *Climatic Change*, *132*(1), 31–45. https://doi.org/10.1007/s10584-015-1430-2

Spence, M. A., Blanchard, J. L., Rossberg, A. G., Heath, M. R., Heymans, J. J., Mackinson, S., et al. (2018). A general framework for combining ecosystem models. *Fish and Fisheries*, *19*(6), 1031–1042. https://doi.org/10.1111/faf.12310

Thomas, R. Q., Figueiredo, R. J., Daneshmand, V., Bookout, B. J., Puckett, L. K., & Carey, C. C. (2020). A Near-Term Iterative Forecasting System Successfully Predicts Reservoir Hydrodynamics and Partitions Uncertainty in Real Time. *Water Resources Research*, *56*, e2019WR026138. https://doi.org/10.1029/2019WR026138

Thomas, R. Q., McClure, R. P., Moore, T. N., Woelmer, W. M., Boettiger, C., Figueiredo, R. J., et al. (2023). Near-term forecasts of NEON lakes reveal gradients of environmental predictability across the US. *Frontiers in Ecology and the Environment*, *21*(5), 220–226. https://doi.org/10.1002/fee.2623

Umlauf, L., Burchard, H., & Bolding, K. (2005). GOTM - Sourcecode and Test Case Documentation. Retrieved from http://www.gotm.net/pages/documentation/manual/stable/pdf/a4.pdf

Velázquez, J. A., Anctil, F., Ramos, M. H., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. *Advances in Geosciences*, *29*, 33–42. https://doi.org/10.5194/adgeo-29-33-2011

Viboud, C., Sun, K., Gaffey, R., Ajelli, M., Fumanelli, L., Merler, S., et al. (2018). The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, *22*, 13–21. https://doi.org/10.1016/j.epidem.2017.08.002

Wander, H. L., Thomas, R. Q., Moore, T. N., Lofton, M. E., & Carey, C. (2023). Data assimilation experiments inform monitoring needs for near-term ecological forecasts in a eutrophic reservoir Data assimilation experiments inform monitoring needs for near-term ecological forecasts in a eutrophic reservoir. *Preprint*. https://doi.org/10.22541/essoar.168500255.59108131/v1

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2022). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, 1–56. https://doi.org/10.1016/j.ijforecast.2022.11.005

Ward, E. J., Holmes, E. E., Thorson, J. T., & Collen, B. (2014). Complexity is costly: a meta-analysis of parametric and non-parametric methods for short-term population forecasting. *Oikos*, *123*(6), 652–661. https://doi.org/10.1111/J.1600-0706.2014.00916.X

Weber, M., Rinke, K., Hipsey, M. R., & Boehrer, B. (2017). Optimizing withdrawal from drinking water reservoirs to reduce downstream temperature pollution and reservoir hypoxia. *Journal of Environmental Management*, *197*, 96–105. https://doi.org/10.1016/j.jenvman.2017.03.020

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model conbination really enhance the preidction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, (134), 241–260. https://doi.org/10.1002/qj

Wilson, H. L., Ayala, A. I., Jones, I. D., Rolston, A., Pierson, D., de Eyto, E., et al. (2020). Variability in epilimnion depth estimations in lakes. *Hydrology and Earth System Sciences*, *24*(11), 5559–5577. https://doi.org/10.5194/hess-24-5559-2020

Wynne, J. H., Woelmer, W., Moore, T. N., Thomas, R. Q., Weathers, K. C., & Carey, C. C. (2023). Uncertainty in projections of future lake thermal dynamics is differentially driven by lake and global climate models. *PeerJ*, *11*, e15445. https://doi.org/10.7717/peerj.15445

Yvon-Durocher, G., Allen, A. P., Cellamare, M., Dossena, M., Gaston, K. J., Leitao, M., et al. (2015). Five Years of Experimental Warming Increases the Biodiversity and Productivity of Phytoplankton. *PLOS Biology*, *13*(12), e1002324. https://doi.org/10.1371/journal.pbio.1002324

Zhao, F., Zhan, X., Xu, H., Zhu, G., Zou, W., Zhu, M., et al. (2022). New insights into eutrophication management: Importance of temperature and water residence time. *Journal of Environmental Sciences*, *111*, 229–239. https://doi.org/10.1016/j.jes.2021.02.033

Zhu, S., Ptak, M., Yaseen, Z. M., Dai, J., & Sivakumar, B. (2020). Forecasting surface water temperature in lakes: A comparison of approaches. *Journal of Hydrology*, *585*, 124809. https://doi.org/10.1016/j.jhydrol.2020.124809

Zwart, J. A., Oliver, S. K., Watkins, W. D., Sadler, J. M., Appling, A. P., Corson-Dosch, H. R., et al. (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*(2), 317–337. https://doi.org/10.1111/1752-1688.13093