Journal of Environmental Management 373 (2025) 123478

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Are more data always better? – Machine learning forecasting of algae based on long-term observations

D. Atton Beckmann [a,*], M. Werther [b], E.B. Mackay [c], E. Spyrakos [a], P. Hunter [a,d], I.D. Jones [a]

[a] Biological and Environmental Sciences, School of Natural Sciences, University of Stirling, Stirling, United Kingdom
[b] Swiss Federal Institute of Aquatic Science and Technology, Department of Surface Waters - Research and Management, Dübendorf, Switzerland
[c] UK Centre for Ecology and Hydrology, Lancaster Environment Centre, Lancaster, LA1 4AP, United Kingdom
[d] Scotland's International Environment Centre, School of Natural Sciences, University of Stirling, Stirling, United Kingdom

ARTICLE INFO

Handling editor: Lixiao Zhang

Keywords:
Algal blooms
Cyanobacteria
Forecasting
Freshwater
Early warning
Machine learning

ABSTRACT

Bloom-forming algae present a unique challenge to water managers as they can significantly impair provision of important ecosystem services and cause health risks to humans and animals. Consequently, effective short-term algae forecasts are important as they provide early warnings and enable implementation of mitigation strategies. In this context, machine learning (ML) emerges as a promising forecasting tool. However, the performance of ML models is heavily dependent on the availability of appropriate training data. Consequently, it is essential to determine the volume of data necessary to develop reliable ML forecasts. Understanding this will guide future monitoring strategies, optimize resource allocation, and set realistic expectations for management outcomes. In this study, we used 30 years of fortnightly measurements of 13 different parameters from a lake in the English Lake District (UK) to examine the impact of training data duration on the performance of ML models for forecasting chlorophyll-a two weeks in advance. Once training data availability exceeded four years, a Random Forest model was found to consistently outperform naive benchmarks (mean absolute percentage error 16.4 % lower than the best-performing benchmark). With more than 5 years of training data, model performance generally continued to improve, but with diminishing returns. Furthermore, it was found that equivalent and, in some cases, better performance could be achieved by only using a subset of the most important input features. Additionally, it was found that reducing the sampling frequency had negative impacts on performance, both due to the reduced number of training observations available, and increased forecast horizon. Our findings demonstrate that for lakes ecologically similar to the study site, a consistent and regular sampling programme focused on monitoring a limited number of key parameters can provide sufficient observations for generating short-term algae forecasts after approximately five years of data collection. Importantly, this result provides justification for the initiation of new monitoring programmes for sites where algal blooms are a concern, and suggests that there are likely many pre-existing monitoring datasets which would be suitable for training algae forecast models.

1. Introduction

Excessive growths of algae in freshwater, commonly referred to as algal blooms, compromise the safety of drinking water sources (Brooks et al., 2016; Igwaran et al., 2024), endanger recreational water activities (Carvalho et al., 2013; Wolf et al., 2017), and threaten the stability and diversity of aquatic ecosystems (Amorim and Moura, 2021; Dolah et al., 2001). Blooms can also impose significant economic impacts, affecting tourism, the fish industry, and even property prices (Hamilton et al., 2014; Hoagland and Scatasta, 2006). The far-reaching effects of this issue underscores the urgency for effective management and mitigation strategies to protect public health and preserve ecological integrity (Chorus and Welker, 2021; Codd et al., 2005; D'Anglada et al., 2016).

Measures such as drinking water supply switching, chemical treatment, hypolimnetic syphoning, ultrasonic control, and use of artificial mixing systems can be effective to mitigate some of the negative effects of harmful algal blooms (Stroom and Kardinaal, 2016). However, many of these short-term measures are best used for only a brief period, due to cost, efficacy, or water resource management implications. For example, a drinking water supply may be switched to an alternate reservoir whilst

* Corresponding author.
E-mail address: daniel.atton.beckmann@stir.ac.uk (D. Atton Beckmann).

https://doi.org/10.1016/j.jenvman.2024.123478
Received 17 July 2024; Received in revised form 24 October 2024; Accepted 24 November 2024
Available online 2 December 2024

there is an algal bloom in the main supply reservoir, but the secondary reservoir may only have limited water supply capacity. Therefore, the implementation of these management measures may benefit from early warning systems that enable implementation in a timely way. Short-term algae forecasting (days to weeks ahead) is therefore needed to enable effective management of algal blooms (Cruz et al., 2021; Hamilton et al., 2014; Ibelings et al., 2016; Rousso et al., 2020; Stroom and Kardinaal, 2016).

In recent years, machine learning (ML) approaches have become popular for algae forecasting due to their ability to model non-linear dynamics using complex, and large datasets (Cruz et al., 2021; Franks, 2018; Rousso et al., 2020; Xiao et al., 2024). ML models vary in complexity and form, and include linear regression models, decision-tree based approaches, and artificial neural networks (Sarker, 2021a). Appropriate training data are crucial for ML algae forecasts. Selection of input variables, data frequency, and data collection methods have an effect on performance and this has been explored to some extent for algae forecasts (Bai et al., 2017; Bertani et al., 2017; Lin et al., 2023; Muttil and Chau, 2006; Rousso et al., 2020; Thomas et al., 2018; Xiao et al., 2017). In contrast, there are very few studies which explicitly explore the effect of the length of training dataset on forecast performance. Understanding this would guide future monitoring strategies and set realistic expectations for management outcomes, and therefore be of great value to water managers considering developing algae forecast systems.

Climate change, variations in weather, and factors such as changes to land-use or management practices present a further challenge when forecasting water quality. That is, we cannot expect that water systems will behave in an unchanging way year after year: they exhibit non-stationarity (Milly et al., 2008, 2015). Therefore, whilst it might be assumed that longer training datasets will always give better performance, it may be that if older data represent conditions which are no longer relevant to the present, this could hinder forecasts rather than improve them.

The length of time required to collect sufficient samples for training ML forecasts is likely to be a more considerable constraint for manually collected data than for sub-daily data generated by automated monitoring platforms (Rousso et al., 2020). Consequently, it is particularly important to understand how the length of training data affects forecast performance for manually collected data. Importantly, sampling frequency influences forecasts in two ways. First, it limits how many observations can be collected over time, directly affecting how quickly sufficient training data is accumulated for accurate ML forecasts. Secondly, it establishes the lower boundary for the forecast horizon, how far ahead predictions can be made. The further ahead the prediction, the greater the uncertainty and errors (Derot et al., 2020).

In previous algae forecasting studies that used data sampled on weekly or longer intervals, the training dataset length has varied considerably, from less than four years (Teles et al., 2006; Torres et al., 2011), to more than ten years training data (Lin et al., 2023; Recknagel et al., 1997a; Talib et al., 2008; Welk et al., 2008). Whilst there is considerable variation in performance between studies using different length datasets (Rousso et al., 2020), attributing this specifically to the length of training dataset is difficult due to differences in the performance metrics, forecast horizons, models, and study sites used.

Understanding how many years of data are required for effective forecasting is only useful if coupled with an understanding of which input variables (features) should be measured. This is particularly important for managers in the early stages of initiating new water quality monitoring schemes for algal bloom management. There are some studies which found that blooms could be forecasted using only previous measurements of chlorophyll-a (chl-a), a common proxy for algal biomass (Muttil and Chau, 2006; Xiao et al., 2017). However, algal blooms are known to be caused by multiple drivers, for example water temperature, stratification, residence time and nutrient availability (Paerl and Otten, 2013; Sellner et al., 2003). Consequently, for some systems, these and perhaps other drivers will be necessary for training effective forecasts (Rousso et al., 2020). Therefore, understanding which features are most important, and how this couples with the length of training dataset available would be of great utility to water managers interested in forecasting algae.

Exploring how the performance of ML algae forecast models is influenced by the length of training data is crucial. This knowledge would enable evaluation of the applicability of ML models to existing data, thus guiding the development of algae forecasting systems for water bodies where monitoring is already being undertaken. Additionally, this understanding would inform the design of new monitoring programmes where forecasting algae is a specific objective. This would give an estimation of the number of years' data required for viable ML forecasts, enabling better planning. This is particularly pertinent as many water managers may soon wish to implement more regular and thorough monitoring programs given that harmful algal blooms may increase in intensity and frequency due to anthropogenic climate change (Ho et al., 2019; Paerl and Huisman, 2009; Paerl and Paul, 2012).

Evidently, there is a need to understand the data requirements for ML forecasts of algae. In particular, the question of how long data must be collected for before they become useful for forecasting has not been investigated explicitly. Therefore, we study the following questions using 30 years of fortnightly monitoring data from a small, meso-eutrophic lake.

1. What is the minimum number of years' training data required to surpass the performance of naive benchmark models in forecasting?
2. As we extend the dataset over more years, do we observe diminishing returns in forecast performance improvements?
3. How does the impact of dataset duration on forecast performance vary across different ML models?
4. What is the minimum number of features required to achieve satisfactory performance from a model?
5. How does a reduction in the sampling frequency affect the impact of dataset duration on performance?

2. Material and methods

2.1. Study site

Blelham Tarn is a small, sheltered lake in the English Lake District. Located at 54° 23′ 44″ N, 2° 58′ 41″ W, it has a mean depth of 6.8 m, maximum depth of 14.5 m, surface area of 0.102 km^2, and volume of 693000 m^3 (Ramsbottom, 1976). The lake has several significant external nutrient sources, including a sewage works for the nearby village of Outgate, and sheep grazing in the catchment. The lake is monomictic and is usually stratified from spring to late autumn (Atkinson, 1999). It is on the meso-eutrophic/eutrophic boundary and has diatom blooms (currently dominated by *Asterionella formosa*) in spring followed by algal blooms in the summer that presently include cyanobacteria (*Dolichospermum* sp.), green algae (*Paulschulzia* sp.), and cryptohytes (*Cryptomonas* sp.). The summer blooms in Blelham Tarn have been associated with hypolimnetic anoxia (Elliott and Thackeray, 2004; Foley et al., 2012).

Thirty years of data (1987–2017) were collated including measurements from in-lake biogeochemical monitoring, nearby meteorological stations and river flow stations. The year 2001 was excluded from all analyses as there was no in-lake monitoring for most of that year due to an outbreak of foot and mouth disease.

The in-situ lake data used were from the extensive long-term monitoring programme that was started by the Freshwater Biological Association in 1945 and continued by the UK Centre for Ecology and Hydrology (UKCEH) since 1989. Details of the data collection procedure can be found in the dataset documentation (Feuchtmayr et al., 2021; Maberly et al., 2017). The fortnightly monitoring data used includes a large variety of variables, such as surface temperature, surface oxygen,

several water chemistry parameters, and phytoplankton chl-a. Initially, all the available variables were used, excluding those with continuous data gaps of more than two months (56 days) during the study period (see Table 1).

Daily meteorological data from the UK Meteorological Office (Met Office, 2019) were obtained from four relatively nearby weather stations in the English Lake District (Walney Island, Shap, Newton Rigg, and Keswick). These data were down-sampled to fortnightly intervals by taking the mean from the two-week period of interest, and then the mean value from all four stations was used. These stations were chosen as they are all approximately equidistant (~30 km) from Blelham Tarn and therefore the mean should be representative of conditions at the lake itself. Additionally, this approach has the advantage of ensuring a weather data time series with very few missing values as it is rare for missing values to occur simultaneously at all four stations.

For catchment-specific rainfall measurements, catchment rainfall data from the two nearest stations to Blelham Tarn (Brathay and Rothay) were provided by the UK National River Flow Archive (NRFA) (NRFA, 2023a, 2023b; Tanguy et al., 2021). The average from these two stations was used for analysis.

All data were processed by first linearly interpolating to daily samples, and then down-sampling to generate data at regular fortnightly intervals with no missing values. All variables used are summarised in Table 1.

To explore any consistent changes that occurred in Blelham Tarn over the 30-year study period, we examined decadal trends in the input variables. This was done by linearly interpolating each variable to daily values and then averaging these over three ten-year periods (1987–1996, 1997–2007, 2008–2017), noting that 1997–2007 excluded the year 2001 because of the missing data then. This allowed for identification of variables that have changed either in magnitude or seasonal timing during the study period.

2.2. Algae forecasting

Chl-a was used as the forecast target, as it is commonly used as a proxy for phytoplankton biomass, and was the focus of many previous algae forecast studies (e.g. Lin et al., 2023; Luo et al., 2017; Mellios et al., 2020). Chl-a forecasts were made two weeks ahead to match the sampling frequency. The forecasting models developed in this study incorporate measurements from the four weeks leading up to the forecast date, equating to two sampling dates, or observations. This selection was strategic, prioritising recent changes in input features while avoiding an overly dimensional feature space. Such a constraint is crucial because a ML model's effectiveness hinges on training with more observations than features (Koutroumbas and Theodoridis, 2008). Considering the fortnightly frequency of chl-a data collection, one year yields 26 observations. Thus, employing 26 features, derived from 13 variables, across two separate observations, represents an upper limit to maintain model trainability with limited annual data.

Table 1
Summary of variables, and sources of data used in the study to train and test ML forecast models.

Variable	Units	Source
Ammonium	$\mu g\ L^{-1}$	UKCEH
Nitrate	$\mu g\ L^{-1}$	UKCEH
Surface oxygen saturation	% Air Saturation	UKCEH
Soluble reactive phosphate (SRP)	$\mu g\ L^{-1}$	UKCEH
Dissolved reactive silica	$\mu g\ L^{-1}$	UKCEH
Surface water temperature	°C	UKCEH
Phytoplankton chlorophyll-a	$\mu g\ L^{-1}$	UKCEH
Total phosphorus (TP)	$\mu g\ L^{-1}$	UKCEH
Mean daily wind speed	Knots	UK Met Office
Mean daily relative humidity	%	UK Met Office
Mean daily cloud amount	Oktas	UK Met Office
Mean daily air temperature	°C	UK Met Office
Mean daily rainfall (average from Brathay and Rothay catchments)	mm	NRFA

A simple approach was used to simulate the integration of forecasted weather observations into the models, as these data would likely be readily available for operational forecasting. Specifically, the five meteorological variables were adjusted to reflect a two-week forward shift, meaning that for a given observation date, the meteorological input features correspond to the average conditions of the subsequent two weeks. Forecasting models are usually more effective at making predictions of a stationary time series, one that does not exhibit a trend or seasonality (Granger and Newbold, 1974). Therefore, the forecast target should have a constant mean and variance over time, and low autocorrelation. To convert a non-stationary time series into a stationary one, a common technique is to calculate the differences between consecutive data points, a process known as 'differencing' (Hyndman and Athanasopoulos, 2021). Preliminary research demonstrated that applying first-order differencing - subtracting the previous observation from the current one in a time series - successfully eliminated both the seasonal patterns and the overall trend in chl-a. Consequently, we trained our forecasting models to predict the changes in chl-a levels, rather than the absolute values.

2.2.1. Machine learning models

It is anticipated that models with different complexities and architectures will require varying amounts of training data to perform optimally (Lones, 2024). Accordingly, several models were employed in this study to encompass a broad range of computational approaches. Specifically, we used a Random forest (RF), Neural Network (multilayer perceptron: MLP), recurrent neural network with gated units (GRU), and a Support Vector Machine (SVM), all implemented through the Python package *scikit-learn* (Pedregosa et al., 2011). These models were chosen as they all use different model architectures and have been shown in previous studies to be effective for algae forecasting (Aláez et al., 2021; González Vilas et al., 2014; Harris and Graham, 2017; Hill et al., 2020; Izadi et al., 2021; Kim et al., 2022; Mellios et al., 2020; Park et al., 2015; Recknagel et al., 1997b; Talib et al., 2008; Velo-Suárez and Gutiérrez-Estrada, 2007). Alongside these four non-linear machine learning models, a multiple linear regression model with L2 regularisation (ridge regression) was also used to provide a comparison against a more simplistic approach (Ahn et al., 2011; Fornarelli et al., 2013; Onderka, 2007; Peretyatko et al., 2010; Soranno, 1997). Given that these approaches have different underlying statistical models, they can be considered broadly representative of the classical ML learning landscape. The MLP and GRU served as an initial test to determine the potential benefits of exploring more sophisticated architectures, such as long-short-term-memory (LSTM) neural networks. Advanced architectures, as noted by LeCun et al. (2015) and Sarker (2021b), demand extensive datasets for effective training. Therefore, the decision to expand our research to include more advanced models was contingent on the MLP and GRU's performance. Specifically, we looked to determine whether the MLP and GRU's effectiveness was constrained by the size of the available data. If these models demonstrated robust performance despite potential data limitations, this would justify further exploration into more sophisticated network architectures. Our stepwise approach ensures that the extension of research into advanced models is predicated on empirical evidence of their potential efficacy under constrained, real-world conditions.

Recurrent neural networks are designed to take advantage of data with inherent order, such as time series (Schmidt, 2019). Therefore, to take advantage of this, the GRU model was configured to use a longer series of inputs than the other models. An input series of four observations (data from the last two months) was chosen to balance providing

sufficient data for the GRU to learn time series patterns whilst ensuring that the ratio of features to observations was not excessive.

Prior to training, input features to the GRU, MLP, SVM, and ridge regression models were scaled to improve convergence, and reduce bias towards larger magnitude variables. All data were standardized by removing the mean and scaling to unit variance using the following equation:

$$z = \frac{x - u}{s} \quad (1)$$

where x is the sample of interest, u the mean of the training sample, and s the standard deviation of the training sample. RF models generally do not require scaling of the input features as decision thresholds for each feature are learnt independently of other features (de Amorim et al., 2023; Werther et al., 2022). As there was potential for models to predict negative values of chl-a, for example if the test data contained values outside the typical ranges of the training data, any negative values were set to a very small positive number (1×10^{-12} mg/m^3). This is a reasonable theory-guided adjustment, as a negative value of chl-a is meaningless (Karpatne et al., 2017).

2.2.2. Hyperparameter optimization

The selected ML models expose several hyperparameters that were optimized using a randomized parameter search with 5-fold cross validation (Table 2). Optimization was undertaken once for each model using 20 years of training data from 1987 to 2007, excluding 2001 due to the lack of data from that year. The data from 2008 onwards were set aside as test data and therefore not used for hyperparameter optimization. In the random search, the layer sizes for both the MLP and GRU were constrained to a maximum of 100 to avoid excessive over-parameterisation.

2.2.3. Benchmark models

Naive models are important for benchmarking more complex forecast models, as they provide a minimum baseline level of performance that should be achieved (Hyndman and Athanasopoulos, 2021; McLaughlin, 1983). For this study, two different benchmarks were used, chosen for their ubiquity in the forecasting literature, and previous use in algae and other water management forecasting studies (Hyndman and Athanasopoulos, 2021; Jackson-Blake et al., 2022; Page et al., 2018; Thomas et al., 2020).

1. Persistence Forecast – chl-a persists over time and does not change since the last time it was measured
2. Seasonal Naive Forecast – chl-a measurements are linearly interpolated to daily values and the value from the same day and month, but previous year to the forecast date used as the prediction.

2.3. Statistical analysis of performance

For testing purposes, the dataset's most recent decade, spanning from 2008 to 2017, was selected. To assess how the duration of the training dataset impacts model performance, we conducted experiments varying the number of training years from 1 to 20 for each year under test. This setup was designed to mimic a real-world forecasting context, where training is performed using the most recent data available before the year being forecasted. For instance, in forecasting the year 2008, we developed 20 separate models using the RF algorithm. These models began with a minimal approach, utilising only 2007 as the single year of training data, and incrementally added more years, culminating in a model trained on a 20-year span from 1987 to 2007 (excluding 2001). For each subsequent test year, models were retrained using the years directly prior to the test year. For example, forecast models for 2009 were at first trained only with data from 2008, then 2008 and 2007, and so on. Considering each number of training years separately, every test year was used as a fold in a non-randomized ten-fold cross validation, ensuring a systematic evaluation of model performance. To identify significant performance variations across the models, we applied a Friedman test (Demšar, 2006; Derrac et al., 2011). Following the Friedman test, a post-hoc test using Holm's procedure for p-value adjustment was used to make multiple comparisons between the benchmark and machine learning models (Holm, 1979).

The criterion for determining the minimal requisite training data duration was established at the juncture where all model-specific p-values for a given performance metric fell below 0.1, with over half also under 0.05, indicating statistical significance. This procedure was repeated for two different performance metrics: the root mean square error (RMSE) and mean absolute percentage error (MAPE). These metrics were chosen for their prevalent use in forecasting tasks, and for the differences in their properties: RMSE penalises errors equally regardless of the magnitude of the true value, whereas MAPE scales errors according to the magnitude of the true value (Bowerman et al., 2005; Chai and Draxler, 2014). Performance metrics were calculated excluding data from the winter months (December, January, and February). This adjustment accounts for the natural absence of algal blooms during these periods in Blelham Tarn, documented by Atkinson (1999). This prevents over-ranking of any conservative models that tend to predict minimal changes in chl-a, something which is pertinent in the context of algal bloom forecasting, a key application of this work.

2.4. Importance of model features

Following the initial investigation, which used all available input features, the models were re-trained using the same train-test splits. However, this time, we systematically varied the number of input features. This approach allowed us to assess the impact of feature selection on model performance explicitly, providing insights into the optimal set of features for effective predictions (Breiman, 2001). To avoid the need for a fully exhaustive evaluation of all possible feature combinations, the relative importance of the input features was first established using the permutation feature importance. This method evaluates a trained model's dependency on specific features by randomly shuffling the values of each feature in isolation and observing the resultant decrease in model performance. To capture some of the variability in feature importance as the number of training years varied, feature importances from two scenarios were evaluated – "low-data availability" with five years training data, and "high-data availability" with fifteen years training data. For both scenarios, the permutation importance of all ML

Table 2
Hyperparameters obtained from randomized cross validation search for the models used in this study. Parameter naming follows the scikit learn v1.3.2 convention (Pedregosa et al., 2011).

Random Forest		Support Vector Machine		Multilayer Perceptron		Ridge Regression		Gated Recurrent Neural Network	
max depth	84	*C*	43	*activation*	relu	*alpha*	117	*units*	4
max features	0.43	*epsilon*	0.21	*alpha*	1.76e-3			*activation*	relu
min samples leaf	3	*kernel*	linear	*Hidden layer sizes*	[16, 36]			*dropout*	7.4e-5
min samples split	7			*Learning rate*	adaptive			*recurrent activation*	sigmoid
n estimators	1845			*solver*	sgd			*batch size*	100
oob score	True			*Early stopping*	True				
				validation fraction	0.10				

models was evaluated for the ten test years 2008–2017, using the years directly prior to these for training. Given that we were most interested in the importance of specific variables rather than the associated measurement lag of two or four weeks, the permutation importances were averaged for each variable to give a hierarchy of 13 permutation importances. This hierarchy was subsequently used to assess scenarios in which only the top *n* most critical variables were accessible for analysis. Therefore, the permutation importances for all models were averaged, and an overall order of feature importance was established by ranking the importances from the averages of the two scenarios. The RMSE metric was used for this ranking as it is more stable than MAPE for evaluating performance where the target is close to zero (Makridakis, 1993). This characteristic of RMSE minimises the likelihood of generating misleading feature importances.

Subsequently, the training and testing process, which varied the training duration from 1 to 20 years for the test period spanning 2008 to 2017, was applied to scenarios where the ML models had access to only the *n* most critical variables. Here, *n* was adjusted from one, representing a single variable, up to the full set of thirteen variables.

Additionally, the Pearson correlation between all input variables was calculated to identify any strongly correlated features which might be redundant. Then, pairs of highly correlated features (correlation coefficient >0.7) were selected for further testing. To check these for redundancy the train-test routine was run first with all features excluding one of the pair, and then ran again with the exclusion of the other.

2.5. Sampling frequency

Adequate machine learning performance is contingent on there being a sufficient number of training observations available. Therefore, the frequency at which data is collected can be expected to influence how long data would need to be collected for before the desired performance is reached. Additionally, sampling frequency dictates the minimum forecast horizon (number of days ahead) that a model can be trained to predict over, which is expected to have a strong effect on performance. For these reasons, the effect of reducing the sampling frequency from fortnightly to monthly (28 days) was investigated. Four different scenarios were compared. Firstly, the standard train test procedure was used as a baseline (fortnightly observations used to make predictions two weeks into the future). Then to isolate the effect of increasing the forecast horizon, the same fortnightly data was used to make predictions 28 days into the future. Following this, the data was down sampled to monthly (28 day) observations by removing every other observation. This monthly data was used to make forecasts 28 days into the future. Finally, 14 day ahead forecasts were trained using input data sampled at 28 day intervals. This isolates the effect of reducing the sampling frequency without changing the forecast horizon or number of observations available in a given training year. For each of these sampling scenarios, the number of training years made available to the models was varied from 1 to 20 years, and performance tested for each of the 10 training years (2008–2017).

3. Results

3.1. Historical patterns and non-stationarity

There have been some shifts in the timing and behaviour of chl-a blooms. In the two most recent decades, the summer chl-a bloom generally starts in July, but in the earliest decade (1987–1996) this occurs approximately one month later. Furthermore, this earliest decade shows a more distinct “clearwater” phase, where the chl-a levels decrease between the spring and summer blooms (Fig. 1a). Additionally, nutrient levels (nitrate, total phosphorus, ammonium) in the most recent decade are generally lower than in the two previous decades (Fig. 1b, c, d).

3.2. Effect of training data on model performance

With less than ten years training data, the RF and ridge regression models outperformed all other models. With just five years training data the RF had an RMSE of 7.9 μg L^{-1}, and MAPE of 39.4 %. This is 1.0 μg L^{-1} (11.4 %) and 16.4 % lower than the respective values for the better performing of the two benchmarks, the persistence forecast. With

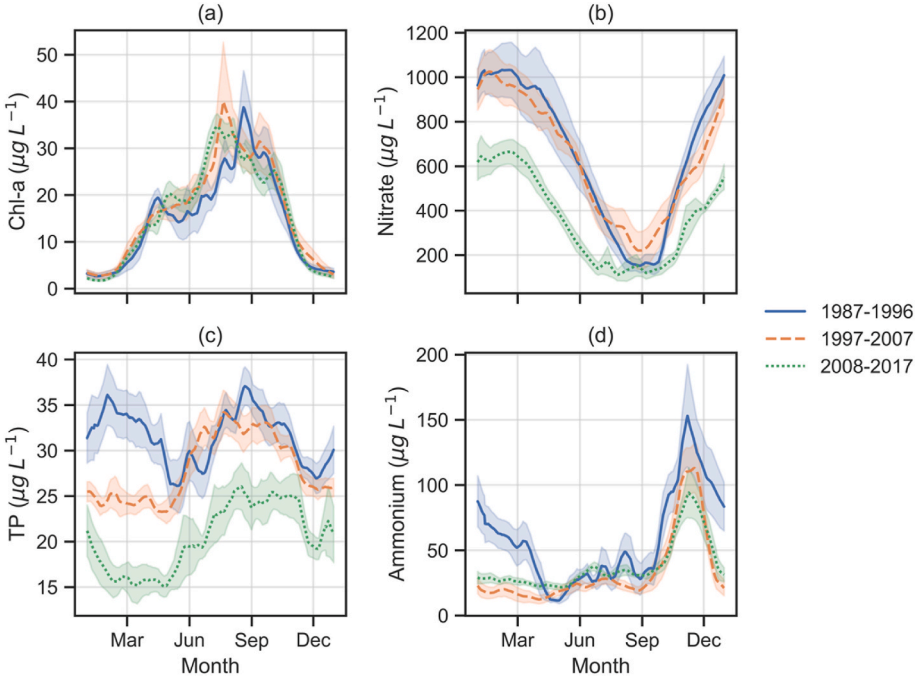

Fig. 1. Decadal means of (a) Chl-a, (b) nitrate, (c) total phosphorus, and (d) ammonium obtained by interpolating the time-series to daily values and averaging these over three ten-year periods: 1987–1996, 1997–2007 (excluding 2001), 2008–2017. The shaded margins indicate standard error.

sufficient training data, all ML models achieved similar performance, but the ridge regression achieved the lowest RMSE of 7.6 μg L^{-1} (which represents a 15.5 % improvement over the better of the benchmarking estimates) with fourteen years of training data, and the SVM achieved the lowest MAPE of 36.8 % (which represents a 22.0 % improvement over the better of the benchmarking estimates) with nineteen years training data. The MLP was generally not competitive with the best performing models.

The bulk of the improvement in the performance of the most competitive models was obtained with the first five years of training data, and whilst the addition of more years of training data did further improve performance, this was generally with diminishing returns (Fig. 2).

For a given number of training years, the Holm adjusted *p*-values from the post-hoc tests indicate if the differences in performance between the machine learning models and benchmarks are statistically significant. It was found that the RF and ridge regression models had much lower requirements in terms of the number of training years for statistical significance than the other models (Table 3, Fig. 3b and. c). The MLP never consistently outperformed the benchmarks in a statistically significant way.

With a low number of training years (<3), SVM performance is close to being statistically significantly worse than the persistence forecast, which is why the p-values are initially low (Fig. 3d). However, all other models achieved similar performance to the persistence forecast even with only one year of training data (Figs. 2 and 3).

The seasonal naive forecast had the poorest performance overall, and accordingly the number of years training data needed for statistically significant differences was lower for all models.

3.3. Feature importance

In both the five- and fifteen-year training scenarios, chl-a was the most important feature, followed by surface water temperature and air temperature (Fig. 4). Although there was some variation in the order of importances between the two scenarios, they follow very similar patterns. Notably, the most obvious difference is that the average change in RMSE from shuffling chl-a was 0.6 μg L^{-1} higher than the next most important feature, water temperature, in the higher data availability scenario. Comparing this with the lower data availability scenario, where the difference was 0.3 μg L^{-1}, suggests that on average, given more training data, models were less reliant on lower ranking features.

Models trained exclusively with the four most significant parameters or fewer exhibited reduced performance, despite being trained over the maximum duration of years available. Generally, using a larger number of input features had a positive effect on the performance of most models (Fig. 5), but there was little benefit to including more than six features (Fig. S1). For example, all ridge regression models trained with more than four input features had very similar RMSE curves, indicating that for this model, the addition of the least important features (total phosphorus, ammonium, SRP, cloud amount, nitrate) had little benefit to performance (Fig. 5e). The SVM models trained with just the six most important features reached lower RMSE values with fewer training years than the same model trained with the maximum number of features (Fig. 5g, Fig. S1). The benefit of including the five least important features is more apparent in the MAPE curves but is still minimal (Fig. 5b–d, f, h).

The only two input features with a high correlation (>0.7) were air temperature and surface water temperature, which had a correlation coefficient of 0.93. For the best performing model with five years of training data, the RF, the removal of air temperature reduced the RMSE by 0.05 μg L^{-1} (0.60 %) and increased the MAPE by 0.40 %. Similarly, the removal of surface water temperature increased the RMSE by 0.01 μg L^{-1} (0.14 %) and increased the MAPE by 0.33 %. For the best performing model with 20 years training data, the SVM, the removal of air temperature increased the RMSE by 0.05 μg L^{-1} (0.69 %) and increased the MAPE by 1.2 %. Likewise, the removal of surface water temperature increased the RMSE by 0.02 μg L^{-1} (0.2 %) and increased the MAPE by 0.28 %. Therefore, given the low performance changes observed, there is a degree of redundancy in these two variables.

Table 3
Minimum number of years training data needed to outperform the persistence forecast in a consistent and statistically significant way, defined as the number of years training data beyond which all p-values were <0.1, and over half of these <0.05.

	Minimum number of years for consistent performance difference	
Model	RMSE	MAPE
MLP	–	–
RF	4	3
Ridge	4	3
SVM	10	12
GRU	20	8

3.4. Sampling frequency

For all models, increasing the forecast horizon from 14 to 28 days incurred a larger performance reduction than reducing the sampling frequency alone (Fig. 6). The performance of 14 day ahead forecasts made using data sampled at 28 day intervals were largely very similar to those made with a 14 day sampling period. In general, 28 day ahead forecasts with 14 day samples had lower RMSE than those made with 28 day samples. However, there were only two models (SVM and GRU) for which the MAPE of 28 day ahead forecasts made with 14 day samples was generally lower than those made with 28 day samples (Fig. 6h–j). For the SVM model, the performance of 28 day ahead forecasts made

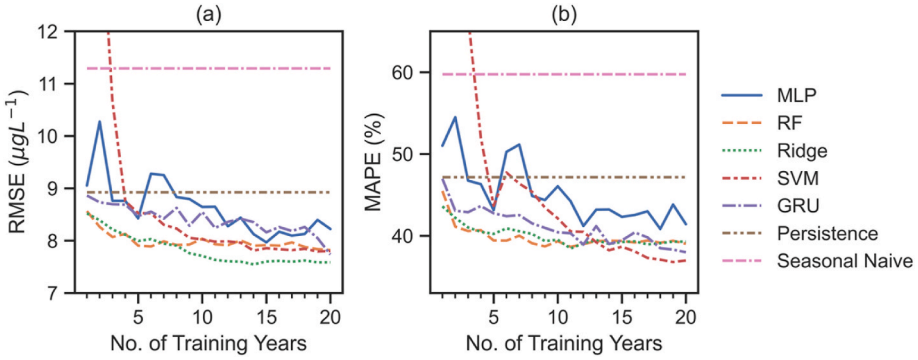

Fig. 2. Performance metrics, (a) RMSE, and (b) MAPE plotted against number of training years. Lines indicate mean values. Blue solid line is MLP, orange dashed line is RF, green dotted line is ridge regression, red dash-dot line is SVM, purple horizontal dash-dot line is persistence forecast, and brown horizontal dash-dot line is seasonal naive forecast. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

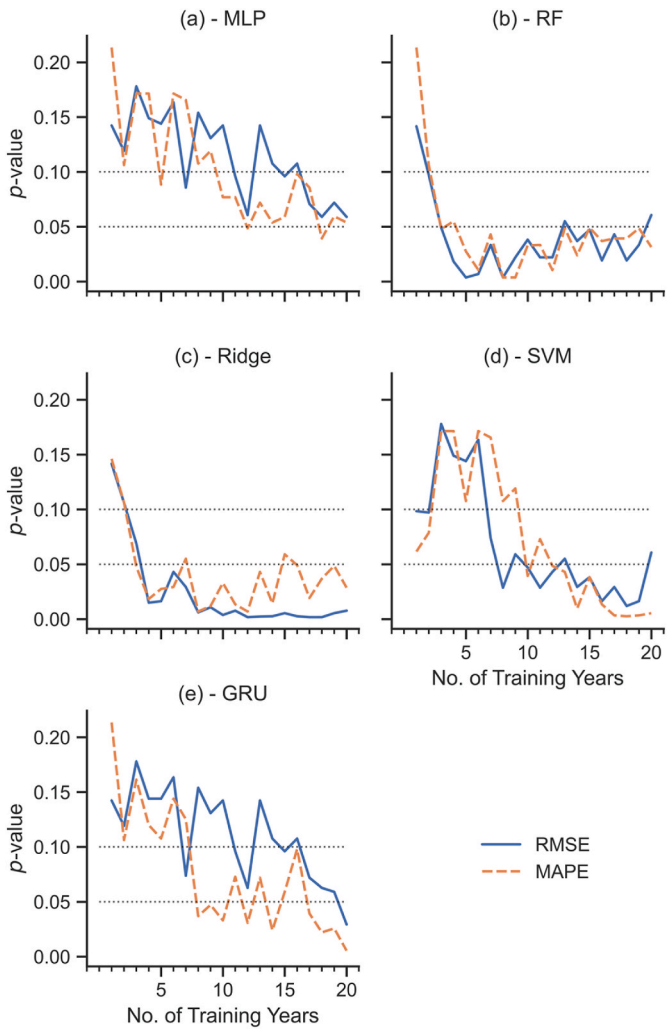

Fig. 3. Performance comparisons of persistence forecast against (a) MLP, (b) RF, (c) ridge regression, and (d) SVM. All plots show p-values from Holm multiple comparisons procedure for both RMSE (solid blue lines) and MAPE (dashed orange lines). Dotted horizontal lines indicate 0.05 and 0.1 significance levels. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

with more than 15 years' training data were approximately equivalent in performance regardless of the sampling frequency (Fig. 6g and h). In contrast, with less than 10 years' training data, 28 day ahead SVM forecasts made with 14 day samples performed better than those made with 28 day samples (Fig. 6g and h).

4. Discussion

4.1. How many years of data are needed?

With the full input parameter set and four or more years' training data, we found that the RF and ridge models were able to consistently outperform both benchmarks in terms of RMSE and MAPE. This suggests that ML can be effective for algae forecasting even when the number of years' training data available are limited, with the obvious caveat that some models may be more suitable than others. Furthermore, the RF, GRU and ridge regression models always had mean RMSE and MAPE lower than both benchmarks. Therefore, it can be argued that certain ML approaches would be effective, or at least no-less effective than naive models, even with only one year of training data. However, the best RMSE (ridge regression with fourteen years training data) was only 4.6 % lower than the RF with five years training data, and the best MAPE (SVM with nineteen years training data) was only 2.6 % lower than the RF with five years training data. Ultimately, we have shown that for this study site, only a low number of training years are required for ML to be useful for forecasting, but that there is not very much to be gained from the inclusion of additional years' training data. For the management of fortnightly sampled lakes which are ecologically-similar to Blelham Tarn, we would therefore encourage managers to implement ML forecasts even when the length of the available data is limited, and to continue data collection for at least five years before making judgements about the utility of the data for forecasting. For lakes with a significantly different sampling approach or ecology from Blelham Tarn, we would encourage repetition of the workflow that has been presented. In short, this would involve examining the forecast performance of multiple models over several test years whilst varying the number of training years.

In a review of forecasting models for cyanobacterial blooms in freshwater lakes, Rousso et al. (2020) found that of the reviewed 90 data-driven studies using similar input data to the present study (i.e. algal concentrations estimated through microscopy or pigment analysis), 90.0 % used at least one year of monitoring data to develop data-driven prediction models. Of the same group of studies, 55.6% used more than four years, 26.7 % used more than 10 years, and 8.9 % used 15 or more years of data. Importantly, this distribution is likely to reflect the general availability of long-term monitoring data: datasets of less than 10 years are much more common than those with longer time series than this. Therefore, placed alongside our findings, this suggests that numerous existing datasets are suitable for ML forecasts of algal concentrations, if a similar level of predictability and sampling frequency to

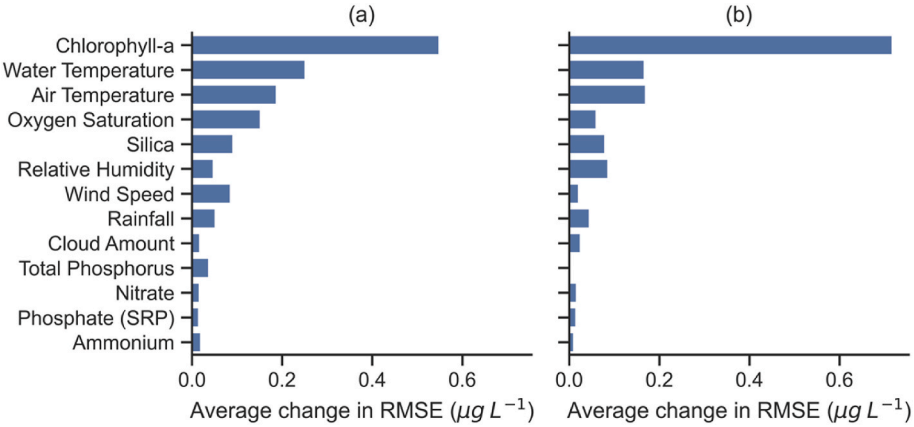

Fig. 4. Overall average permutation feature importances (evaluated through a performance change in RMSE) with (a) five years, and (b) fifteen years training data.

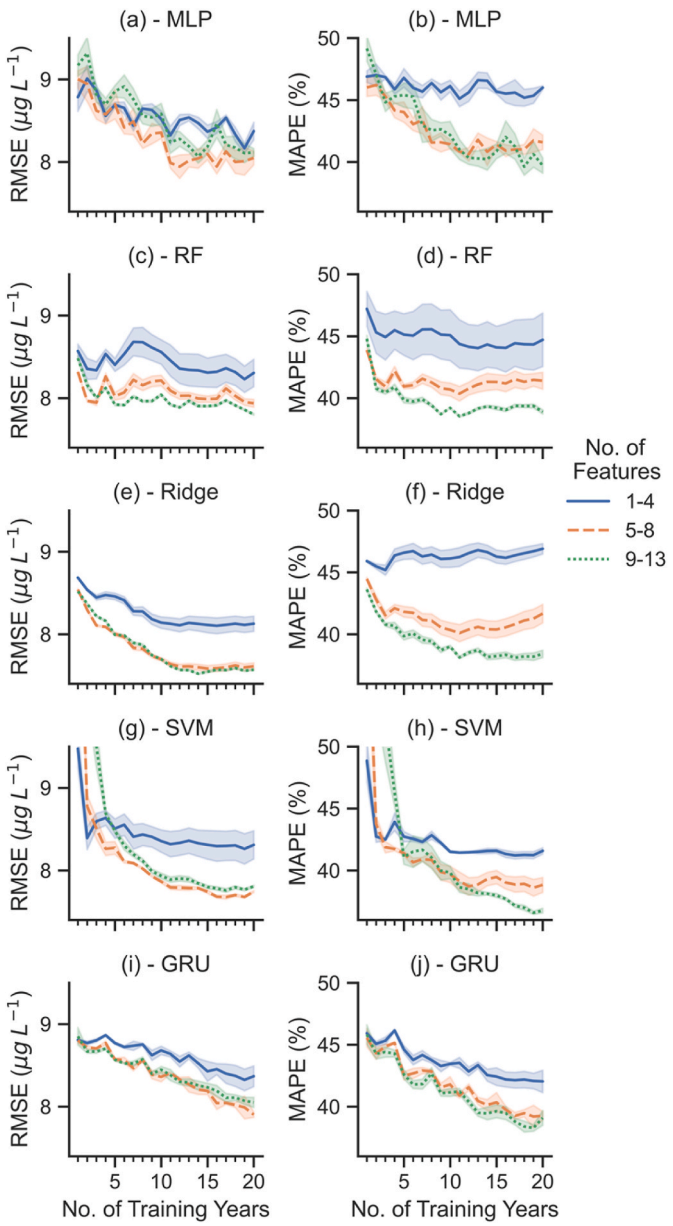

Fig. 5. RMSE (left column) and MAPE (right column) plotted against the number of training years with the number of input features varied for MLP: (a), (b); RF: (c), (d); ridge regression: (e), (f); SVM: (g), (h); and GRU: (i), (j). Blue lines indicate the mean performance of models trained with the 1–4 most important features (i.e. 1: only chl-a; 2: chl-a and water-temperature; 3: chl-a, water temperature and air temperature; 4: chl-a, water temperature, air temperature and oxygen saturation). Orange dashed lines indicate models trained with the 5–8 most important features, and green dotted lines indicate models trained with the 9–13 most important features. Shaded margins indicate standard error around the mean centre line. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

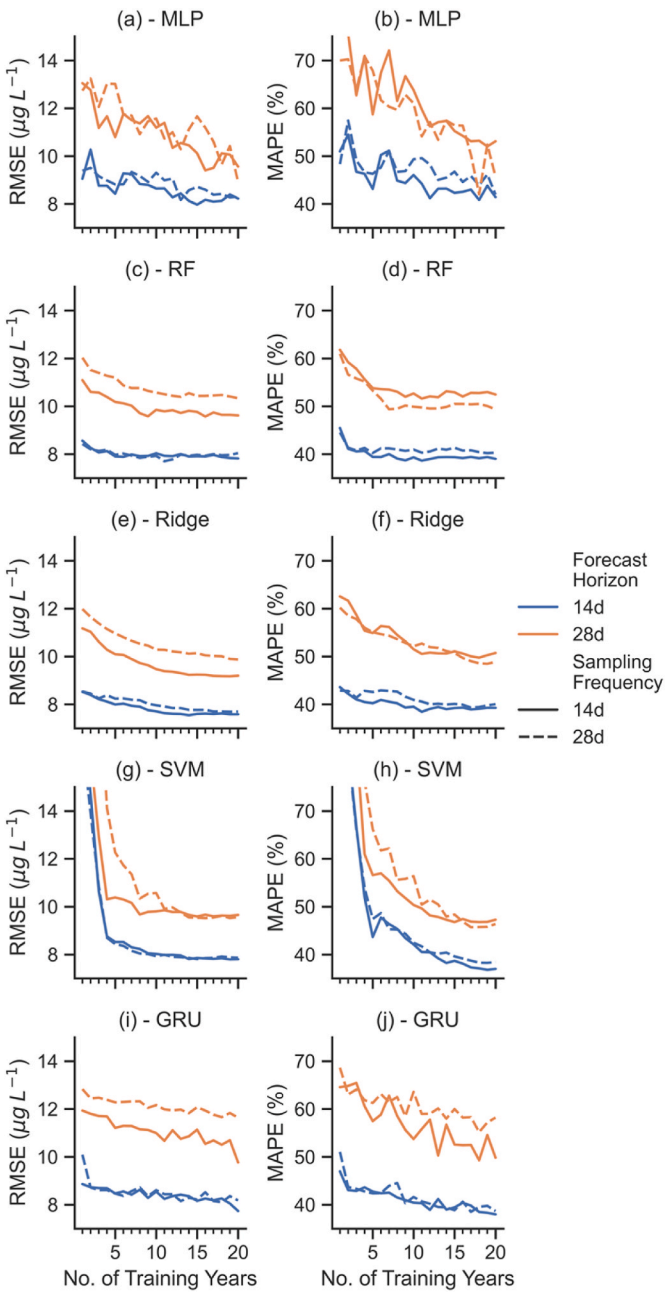

Fig. 6. RMSE (left column) and MAPE (right column) plotted against the number of training years for two different forecast horizons of 14 (blue lines) and 28 (orange lines) days. Sampling frequencies of 14 days are indicated by solid lines, and 28 days dashed lines. Results are shown for all five ML models: MLP: (a), (b); RF: (c), (d); ridge regression: (e), (f); SVM: (g), (h); and GRU: (i), (j). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Blelham Tarn is assumed. It is likely that some systems, for example those which do not experience algal blooms every year, would require more training years' training data than this (Shyalika et al., 2023). However, Blelham Tarn can be considered a typical example of a fertile temperate lake. It consistently has spring diatom blooms succeeded by cyanobacteria in the summer, and therefore our findings are likely widely applicable to the many lakes that share similar seasonal behaviour (e.g. Bailey-Watts, 1981; Rose et al., 2021; Wei et al., 2020). Furthermore, for monitoring routines where automated high frequency sensors are not used, fortnightly or monthly lake sampling is common (Dubelaar et al., 2004; Marcé et al., 2016; Spaulding et al., 2024), and so there are likely many existing datasets for which these findings are applicable.

Rousso et al. (2020) also found that almost half (44.4 %) of the reviewed studies with similar input data to the present study used less than five years data. This could be an indication that these studies were able to achieve good performance with fewer training years than in the present study, or it might indicate that some of these models were not performing significantly better than naive benchmarks would have. Whilst benchmarking against naive models is not a new concept in the

wider field of forecasting (McLaughlin, 1983), the authors of the present study are aware of only a handful of studies which use naive benchmarks for algae forecasts (Jackson-Blake et al., 2022; Matthews, 2023; Page et al., 2018). Our results demonstrate that some relatively sophisticated forecasts, such as the MLP, may not be guaranteed to be significantly better than a naive model, unless provided with much more training data than is available for most lakes. Therefore, we strongly suggest the use of naive benchmarking for algae forecasting.

Many studies have found that ML performance improves with diminishing returns as the training dataset size increases (Frey and Fisher, 1999; Last, 2007; Mahmood et al., 2022). In the present study, this was likely caused by several factors. Changes in the sampling approach and improvements to measurement methods over the duration of the data collection period may affect the utility of older training data. For example, there have been organizational changes, sampling frequency variations, and upgrades to the lab equipment used for analysis of samples (Maberly et al., 2017). Alongside this, nutrient loads and associated phytoplankton dynamics in Blelham Tarn have changed over the 30-year study period (Fig. 1). These have been accompanied by many other changes including an earlier onset and longer duration of stratification; increased mean surface water temperature; and increased occurrence of hypolimnetic anoxia (Foley et al., 2012). Evidently this is a non-stationary system, which likely means that older training years are less representative of the conditions being modelled in a given test year. Globally, lakes are experiencing rapid warming and changes to mixing patterns due to anthropogenic climate change (Maberly et al., 2020; O'Reilly et al., 2015; Woolway and Merchant, 2019). This is likely to have a profound effect on phytoplankton species composition and biomass, and is expected to lead to increased cyanobacterial dominance in many water bodies (Carey et al., 2012; Huisman et al., 2018). Therefore, if warming rates further increase, it may be that historic data becomes increasingly less relevant to modelling short-term fluctuations. Investigating this explicitly would be of great value in understanding how the performance of near-term ML algae forecasts could be affected by a rapidly changing climate.

The use of ML approaches does improve our ability for modelling forecasts over benchmarking forecasts, but the observational sampling approach is likely to have a limiting effect on performance. During the summer growing season, phytoplankton dynamics are complex and influenced by many factors, including grazing, effects from higher trophic levels, and physical factors such as flushing (Sommer et al., 2012; Stockwell et al., 2020). Whilst some of these drivers may be correlated with the input features, we cannot expect a monitoring program to fully capture all the complexities of a dynamic ecosystem. Additionally, the horizontal movement of algae, particularly cyanobacteria, will also have some influence on the observed chl-a, which is only measured at a single location at Blelham Tarn (Feuchtmayr et al., 2021; Maberly et al., 2017; Xue et al., 2023). Therefore, given this complexity, and potential for rapid changes, certain aspects of the summer bloom dynamics will not be sufficiently captured by the sampling efforts (Pobel et al., 2011). In other words, limitations of the monitoring approach contribute uncertainty to the forecasts, regardless of the length of data used to train the model.

4.2. Model selection

Appropriate model selection is critical, particularly in data-limited scenarios. With the full input parameter set, the GRU, MLP and SVM both required more than three times the number of years training data required by the RF to consistently outperform the benchmarks in a statistically significant way (Fig. 2, Table 3). The model dependence of this result likely has several reasons: firstly, the MLP and GRU are higher complexity models than the RF and ridge regression models as they have a larger number of parameters (weights) to learn. This may be the reason that several studies which used similar datasets also found decision tree models to perform better than neural networks at algae forecasting (Aláez et al., 2021; Fornarelli et al., 2013; Harris and Graham, 2017). The result that the more complex models performed poorly in data-limited scenarios is not surprising but highlights the importance of using naive forecasts for benchmarking, as some ML models may not be able to outperform these. Secondly, as the hyperparameters were tuned using 5-fold cross validation with twenty years training data, there was some inherent bias towards larger numbers of training years, which may have resulted in some overfitting in scenarios where a lower number of training years were available. This is a plausible explanation for the very poor performance of the SVM models trained with the full feature set and a low number of training years. This issue was alleviated by removing some of the less important input features which indicates that it was at least partly a result of overfitting to spurious correlations between chl-a and the less important input features. Often, it is argued that more complex ML models have the advantage of being able to effectively model non-linear dynamics (Cruz et al., 2021), however our findings suggest that multiple linear regression with regularisation may be all that is required to model algal concentrations using fortnightly data. This finding also suggests that more complex ML architectures which would be expected to outperform linear regression given sufficient training data are likely not appropriate unless datasets are available for several decades or at higher than fortnightly resolution.

4.3. Importance of input features

As well as only a few years' training data being sufficient to optimize model performance, only a handful of lake measurements were necessary. We found that excluding the seven least important predictor variables generally gave similar performance to using the full parameter set (Fig. 5, Fig. S1). Furthermore, the SVM model's performance was actually improved by removing these parameters (Fig. S1). Several of the more important parameters were meteorological variables taken from weather stations up to 37 km from the lake, data that are commonly available from regular meteorological stations or weather forecasts (Fig. 4). Furthermore, it was found that the air and surface water temperature variables were highly correlated, and that therefore either one of these could be excluded with a minimal performance impact. Therefore, for fortnightly sampled lakes ecologically similar to Blelham Tarn, only a relatively small amount of in-lake monitoring would be necessary to provide a basis for using ML approaches. Nevertheless, unlike Muttil and Chau (2006) and Xiao et al. (2017), we found that using chl-a alone was not effective, despite this variable standing out as the most important predictor. This difference in outcome is likely due to differences in sampling frequency and study site: Xiao et al. (2017) used daily monitoring data, and Muttil and Chau (2006) used data from a large estuarine harbour.

Several of the features which we found to be least important to ML forecasts are those often considered to be strongly linked with algal blooms: ammonium, SRP, and nitrate (O'Neil et al., 2012; Paerl and Otten, 2013). This indicates that for Blelham Tarn, the fortnightly fluctuations of these variables are not useful for predicting chl-a. This is likely because nutrient concentrations are often at limiting values during the summer, where most of the large chl-a fluctuations occur (Fig. 1). Page et al. (2018) identified that difficulties in modelling SRP fluxes in Esthwaite Water was a significant hindrance to chl-a forecast performance, and suggested that monitoring nutrients more regularly than fortnightly could ameliorate this issue. Given the geographic proximity, and similarity in trophic status between Blelham Tarn and Esthwaite Water, this could likely explain why we did not find fortnightly nutrient measurements to be particularly useful for forecasting. Whilst there are many other studies which found that water temperature and other physical parameters were generally stronger predictors of algal concentrations than nutrient measurements (Li et al., 2021; Rousso et al., 2020; Wang et al., 2019), there are also many studies which identify nutrients as key predictors. For example, Rousso et al. (2020) found that 30.5 % of the reviewed cyanobacterial bloom forecast studies identified

either phosphorus or nitrogen as the most important predictor in their models. This indicates that the optimal set of features for forecasts is likely to be highly variable across different study sites and monitoring schemes. Where possible, monitoring a broad suite of water quality parameters is advantageous because this will likely cover statutory monitoring requirements; provide data useful for other scientific and management investigations; and importantly, provide a degree of flexibility in forecast models should the system change in such a way that the feature importance hierarchy is modified. For example, forecasts of cyanobacterial blooms in a temperate lake might initially be highly reliant on water temperature measurements. However, if over a number of years, the lake was to warm to the point where summer temperatures were consistently optimal for cyanobacteria growth, then nutrient availability could become a more important predictor than temperature (Bonilla et al., 2023; Carey et al., 2012; Paerl and Huisman, 2009; Reynolds, 2006; Richardson et al., 2018). Critically, this highlights that whilst automated high frequency monitoring platforms are increasingly being adopted for water quality monitoring applications, the ability of manual sampling approaches to measure a wide range of parameters in a consistent and reliable way is a compelling argument for their continuation (Marcé et al., 2016; Park et al., 2020).

4.4. Sampling frequency

It has been previously demonstrated that increasing sampling frequency has a positive impact on the performance of phytoplankton forecasts (Derot et al., 2020; Lin et al., 2023), and our results corroborate this. Specifically, we found that the strongest way in which sampling frequency affected performance was through limiting the minimum forecast horizon possible. However, with this effect removed, comparisons with 28 day ahead forecasts still showed that higher frequency data generally gave lower errors and, for the SVM, converged to peak performance with fewer years' training data. That this effect was either not present, or much more subtle, for the comparisons made with 14 day ahead forecasts is a reflection of the number of observations available for training in each of these scenarios. Both 14 day ahead forecast scenarios were trained with the same number of observations for a given number of training years, whereas the 28 day ahead forecasts made with 14 day samples had approximately twice as many training observations per training year than those made with 28 day samples. Therefore, alongside making considerations concerning the forecast horizon, managers designing or assessing sampling routines for forecasting should understand that the total number of observations collected annually may affect how many years' training data are required to achieve adequate performance.

5. Conclusion

Whilst this study was only carried out for a single lake, there are many lakes and reservoirs that suffer algal blooms that are monitored at a similar frequency to Blelham Tarn. We have shown that with four or more years' data consisting of just a few parameters sampled on a fortnightly basis, ML can outperform standard naive benchmarks, and is therefore appropriate for generating two week ahead algae forecasts. For those already working with more than ten years' fortnightly data for lakes ecologically-similar to Blelham Tarn, our results suggest that with appropriate model selection, tuning, and feature selection, the performance reached at this point is likely very close to the maximal performance that can be expected, even if more years of training data are added as the sampling continues. Ultimately, our results suggest that there are likely numerous lakes and reservoirs with sufficient existing data for ML forecasting, and that users interested in this approach may see significant improvements in forecast ability just a few years after the initiation of a simple monitoring programme. However, it is still essential to better understand data requirements for ML algae forecasts across a diverse range of sampling approaches and lake ecologies. By adopting a similar workflow to that which has been presented in this study, managers and scientists can continue to advance our understanding of these requirements and therefore refine future algae forecasting efforts.

CRediT authorship contribution statement

D. Atton Beckmann: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **M. Werther:** Writing – review & editing, Supervision. **E.B. Mackay:** Writing – review & editing. **E. Spyrakos:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **P. Hunter:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **I.D. Jones:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Funding

This study was funded by the Scottish Government's Hydro Nation Scholars Programme, which funded Daniel Atton Beckmann's PhD. Blelham Tarn long-term monitoring is currently funded by the UK Natural Environment Research Council as part of the UK-SCAPE programme delivering National Capability (ref NE/R016429/1). Mortimer Werther was supported by the Swiss Science Foundation grant Lake3P, no. 204783.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to the Centre for Environmental Data Analysis and UK National River Flow Archive for providing historical UK Met Office weather station data and catchment rainfall data respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2024.123478.

Data availability

All data used in this study can be accessed from the UKCEH Environmental Information Data Centre, UK Met Office, and UNRFA (Feuchtmayr et al., 2021; Maberly et al., 2017; Met Office, 2019; NRFA, 2023a, 2023b). Model output data available on request.

References

Ahn, C.-Y., Oh, H.-M., Park, Y.-S., 2011. Evaluation of environmental factors on cyanobacterial bloom in eutrophic reservoir using artificial neural Networks1. J. Phycol. 47, 495–504. https://doi.org/10.1111/j.1529-8817.2011.00990.x.

Aláez, F.M.B., Palenzuela, J.M.T., Spyrakos, E., Vilas, L.G., 2021. Machine learning methods applied to the prediction of pseudo-nitzschia spp. blooms in the Galician rias baixas (NW Spain). ISPRS Int. J. Geo-Inf. 10, 199. https://doi.org/10.3390/ijgi10040199.

Amorim, C.A., Moura, A. do N., 2021. Ecological impacts of freshwater algal blooms on water quality, plankton biodiversity, structure, and ecosystem functioning. Sci. Total Environ. 758, 143605. https://doi.org/10.1016/j.scitotenv.2020.143605.

Atkinson, K.M., 1999. Some English lakes as diverse and active ecosystems: a factual summary and source book. Freshwater Biological Association 6 (80).

Bai, X., Zhang, H., Wang, X., Wang, L., Xu, J., Yu, J., 2017. The adaptive-clustering and error-correction method for forecasting cyanobacteria blooms in lakes and reservoirs. Adv. Math. Phys. 2017, e9037358. https://doi.org/10.1155/2017/9037358.

Bailey-Watts, A.E., 1981. Loch Leven Phytoplankton Succession. NERC/ITE, pp. 88–92.

Bertani, I., Steger, C.E., Obenour, D.R., Fahnenstiel, G.L., Bridgeman, T.B., Johengen, T. H., Sayers, M.J., Shuchman, R.A., Scavia, D., 2017. Tracking cyanobacteria blooms: do different monitoring approaches tell the same story? Sci. Total Environ. 575, 294–308. https://doi.org/10.1016/j.scitotenv.2016.10.023.

Bonilla, S., Aguilera, A., Aubriot, L., Huszar, V., Almanza, V., Haakonsson, S., Izaguirre, I., O'Farrell, I., Salazar, A., Becker, V., Cremella, B., Ferragut, C., Hernandez, E., Palacio, H., Rodrigues, L.C., Sampaio da Silva, L.H., Santana, L.M., Santos, J., Somma, A., Ortega, L., Antoniades, D., 2023. Nutrients and not temperature are the key drivers for cyanobacterial biomass in the Americas. Harmful Algae 121, 102367. https://doi.org/10.1016/j.hal.2022.102367.

Bowerman, B.L., O'Connell, R.T., Koehler, A.B., 2005. Forecasting, Time Series, and Regression: an Applied Approach. Thomson Brooks/Cole.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brooks, B.W., Lazorchak, J.M., Howard, M.D.A., Johnson, M.-V.V., Morton, S.L., Perkins, D.A.K., Reavie, E.D., Scott, G.I., Smith, S.A., Steevens, J.A., 2016. Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? Environ. Toxicol. Chem. 35, 6–13. https://doi.org/10.1002/etc.3220.

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 Res., Cyanobacteria: impacts of climate change on occurrence. toxicity and water quality management 46, 1394–1407. https://doi.org/10.1016/j.watres.2011.12.016.

Carvalho, L., McDonald, C., de Hoyos, C., Mischke, U., Phillips, G., Borics, G., Poikane, S., Skjelbred, B., Solheim, A.L., Van Wichelen, J., Cardoso, A.C., 2013. Sustaining recreational quality of European lakes: minimizing the health risks from algal blooms through phosphorus control. J. Appl. Ecol. 50, 315–323. https://doi.org/10.1111/1365-2664.12059.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chorus, I., Welker, M. (Eds.), 2021. Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management, second ed. CRC Press, London. https://doi.org/10.1201/9781003081449.

Codd, G.A., Lindsay, J., Young, F.M., Morrison, L.F., Metcalf, J.S., 2005. Harmful cyanobacteria. In: Huisman, J., Matthijs, H.C.P., Visser, P.M. (Eds.), Harmful Cyanobacteria, Aquatic Ecology Series. Springer, Netherlands, Dordrecht, pp. 1–23. https://doi.org/10.1007/1-4020-3022-3_1.

Cruz, R.C., Reis Costa, P., Vinga, S., Krippahl, L., Lopes, M.B., 2021. A review of recent machine learning advances for forecasting harmful algal blooms and shellfish contamination. J. Mar. Sci. Eng. 9, 283. https://doi.org/10.3390/jmse9030283.

D'Anglada, L., Hilborn, E.D., D'Anglada, L., Hilborn, E.D., Backer, L.C. (Eds.), 2016. Harmful Algal Blooms (HABs) and Public Health: Progress and Current Challenges. MDPI - Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/books978-3-03842-156-6.

de Amorim, L.B.V., Cavalcanti, G.D.C., Cruz, R.M.O., 2023. The choice of scaling technique matters for classification performance. Appl. Soft Comput. 133, 109924. https://doi.org/10.1016/j.asoc.2022.109924.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

Derot, J., Yajima, H., Schmitt, F.G., 2020. Benefits of machine learning and sampling frequency on phytoplankton bloom forecasts in coastal areas. Ecol. Inform. 60, 101174. https://doi.org/10.1016/j.ecoinf.2020.101174.

Derrac, J., García, S., Molina, D., Herrera, F., 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm Evol. Comput. 1, 3–18. https://doi.org/10.1016/j.swevo.2011.02.002.

Dolah, F.M.V., Roelke, D., Greene, R.M., 2001. Health and ecological impacts of harmful algal blooms: risk assessment needs. Hum. Ecol. Risk Assess. Int. J. 7, 1329–1345. https://doi.org/10.1080/20018091095032.

Dubelaar, G.B.J., Geerders, P.J.F., Jonker, R.R., 2004. High frequency monitoring reveals phytoplankton dynamics. J. Environ. Monit. 6, 946–952. https://doi.org/10.1039/B409350J.

Elliott, J.A., Thackeray, S.J., 2004. The simulation of phytoplankton in shallow and deep lakes using PROTECH. Ecol. Model. 178, 357–369. https://doi.org/10.1016/j.ecolmodel.2004.02.012.

Feuchtmayr, H., Clarke, M.A., De Ville, M.M., Dodd, B.A., Fletcher, J., Guyatt, H., Hunt, A.G., James, J.B., Mackay, E., Rhodes, G., Thackeray, S.J., Maberly, S.C., 2021. Surface temperature, surface oxygen, water clarity, water chemistry and phytoplankton chlorophyll a data from Blelham Tarn, 2014 to 2018 (Dataset) [WWW Document]. EIDC. URL. https://catalogue.ceh.ac.uk/id/ae8c850d-211e-4560-8b37-437b6e0e2a16. (Accessed 9 December 2023).

Foley, B., Jones, I.D., Maberly, S.C., Rippey, B., 2012. Long-term changes in oxygen depletion in a small temperate lake: effects of climate change and eutrophication. Freshw. Biol. 57, 278–289. https://doi.org/10.1111/j.1365-2427.2011.02662.x.

Fornarelli, R., Galelli, S., Castelletti, A., Antenucci, J.P., Marti, C.L., 2013. An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers. Water Resour. Res. 49, 3626–3641. https://doi.org/10.1002/wrcr.20268.

Franks, P.J.S., 2018. Recent advances in modelling of harmful algal blooms. In: Glibert, P.M., Berdalet, E., Burford, M.A., Pitcher, G.C., Zhou, M. (Eds.), Global Ecology and Oceanography of Harmful Algal Blooms, Ecological Studies. Springer International Publishing, Cham, pp. 359–377. https://doi.org/10.1007/978-3-319-70069-4_19.

Frey, L.J., Fisher, D.H., 1999. Modeling decision tree performance with the power law. In: Seventh International Workshop on Artificial Intelligence and Statistics. Presented at the Seventh International Workshop on Artificial Intelligence and Statistics. PMLR.

González Vilas, L., Spyrakos, E., Torres Palenzuela, J.M., Pazos, Y., 2014. Support Vector Machine-based method for predicting Pseudo-nitzschia spp. blooms in coastal waters (Galician rias, NW Spain). Prog. Oceanogr. 124, 66–77. https://doi.org/10.1016/j.pocean.2014.03.003.

Granger, C.W.J., Newbold, P., 1974. Spurious regressions in econometrics. J. Econom. 2, 111–120. https://doi.org/10.1016/0304-4076(74)90034-7.

Hamilton, D.P., Wood, S.A., Dietrich, D.R., Puddick, J., 2014. Costs of harmful blooms of freshwater cyanobacteria. In: Cyanobacteria. John Wiley & Sons, Ltd, pp. 245–256. https://doi.org/10.1002/9781118402238.ch15.

Harris, T.D., Graham, J.L., 2017. Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. Lake Reserv. Manag. 33, 32–48. https://doi.org/10.1080/10402381.2016.1263694.

Hill, P.R., Kumar, A., Temimi, M., Bull, D.R., 2020. HABNet: machine learning, remote sensing-based detection of harmful algal blooms. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3229–3239. https://doi.org/10.1109/JSTARS.2020.3001445.

Ho, J.C., Michalak, A.M., Pahlevan, N., 2019. Widespread global increase in intense lake phytoplankton blooms since the 1980s. Nature 574, 667–670. https://doi.org/10.1038/s41586-019-1648-7.

Hoagland, P., Scatasta, S., 2006. The economic effects of harmful algal blooms. In: Granéli, E., Turner, J.T. (Eds.), Ecology of Harmful Algae, Ecological Studies. Springer, Berlin Heidelberg, pp. 391–402. https://doi.org/10.1007/978-3-540-32210-8_30.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M.H., Visser, P.M., 2018. Cyanobacterial blooms. Nat. Rev. Microbiol. 16, 471–483. https://doi.org/10.1038/s41579-018-0040-1.

Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: Principles and Practice, third ed.

Ibelings, B.W., Fastner, J., Bormans, M., Visser, P.M., 2016. Cyanobacterial blooms. Ecology, prevention, mitigation and control: editorial to a CYANOCOST Special Issue. Aquat. Ecol. 50, 327–331. https://doi.org/10.1007/s10452-016-9595-y.

Igwaran, A., Kayode, A.J., Moloantoa, K.M., Khetsha, Z.P., Unuofin, J.O., 2024. Cyanobacteria harmful algae blooms: causes, impacts, and risk management. Water. Air. Soil Pollut. 235, 71. https://doi.org/10.1007/s11270-023-06782-y.

Izadi, M., Sultan, M., Kadiri, R.E., Ghannadi, A., Abdelmohsen, K., 2021. A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom. Remote Sens 13, 3863. https://doi.org/10.3390/rs13193863.

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. Hydrol. Earth Syst. Sci. 26, 3103–3124. https://doi.org/10.5194/hess-26-3103-2022.

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168.

Kim, T., Shin, J., Lee, D., Kim, Y., Na, E., Park, J., Lim, C., Cha, Y., 2022. Simultaneous feature engineering and interpretation: forecasting harmful algal blooms using a deep learning approach. Water Res. 215, 118289. https://doi.org/10.1016/j.watres.2022.118289.

Koutroumbas, K., Theodoridis, S., 2008. Pattern Recognition. Elsevier Science & Technology. UNITED STATES, San Diego.

Last, M., 2007. Predicting and optimizing classifier utility with the power law. In: Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). Presented at the Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007), pp. 219–224. https://doi.org/10.1109/ICDMW.2007.31.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Li, K., Xu, T., Xi, J., Jia, H., Gao, Z., Sun, Z., Yin, D., Leng, L., 2021. Multi-factor analysis of algal blooms in gate-controlled urban water bodies by data mining. Sci. Total Environ. 753, 141821. https://doi.org/10.1016/j.scitotenv.2020.141821.

Lin, S., Pierson, D.C., Mesman, J.P., 2023. Prediction of algal blooms via data-driven machine learning models: an evaluation using data from a well-monitored mesotrophic lake. Geosci. Model Dev. 16, 35–46. https://doi.org/10.5194/gmd-16-35-2023.

Lones, M.A., 2024. How to avoid machine learning pitfalls: a guide for academic researchers. https://doi.org/10.48550/arXiv.2108.02497.

Luo, Y., Yang, K., Yu, Z., Chen, J., Xu, Y., Zhou, X., Yang, Y., 2017. Dynamic monitoring and prediction of Dianchi Lake cyanobacteria outbreaks in the context of rapid urbanization. Environ. Sci. Pollut. Res. 24, 5335–5348. https://doi.org/10.1007/s11356-016-8155-2.

Maberly, S.C., Brierley, B., Carter, H.T., Clarke, M.A., De Ville, M.M., Fletcher, J., James, J.B., Keenan, P., Kelly, J.L., Mackay, E., Parker, J.E., Patel, M., Pereira, M.G., Rhodes, G., Tanna, B., Thackeray, S.J., Vincent, C., Feuchtmayr, H., 2017. In: Surface Temperature, Surface Oxygen, Water Clarity, Water Chemistry and Phytoplankton Chlorophyll a Data from Blelham Tarn, 1945 to 2013 (Dataset) [WWW Document]. EIDC. URL. https://catalogue.ceh.ac.uk/id/393a5946-8a22-4350-80f3-a60d753beb00.

Maberly, S.C., O'Donnell, R.A., Woolway, R.I., Cutler, M.E.J., Gong, M., Jones, I.D., Merchant, C.J., Miller, C.A., Politi, E., Scott, E.M., Thackeray, S.J., Tyler, A.N., 2020. Global lake thermal regions shift under climate change. Nat. Commun. 11, 1232. https://doi.org/10.1038/s41467-020-15108-z.

Mahmood, R., Lucas, J., Acuna, D., Li, D., Philion, J., Alvarez, J.M., Yu, Z., Fidler, S., Law, M.T., 2022. How much more data do I need? Estimating requirements for downstream tasks. https://doi.org/10.48550/arXiv.2207.01725.
Makridakis, S., 1993. Accuracy measures: theoretical and practical concerns. Int. J. Forecast. 9, 527–529. https://doi.org/10.1016/0169-2070(93)90079-3.
Marcé, R., George, G., Buscarinu, P., Deidda, M., Dunalska, J., de Eyto, E., Flaim, G., Grossart, H.-P., Istvanovics, V., Lenhardt, M., Moreno-Ostos, E., Obrador, B., Ostrovsky, I., Pierson, D.C., Potužák, J., Poikane, S., Rinke, K., Rodríguez-Mozaz, S., Staehr, P.A., Šumberová, K., Waajen, G., Weyhenmeyer, G.A., Weathers, K.C., Zion, M., Ibelings, B.W., Jennings, E., 2016. Automatic high frequency monitoring for improved lake and reservoir management. Environ. Sci. Technol. 50, 10780–10794. https://doi.org/10.1021/acs.est.6b01604.
Matthews, M.W., 2023. Near-term forecasting of cyanobacteria and harmful algal blooms in lakes using simple univariate methods with satellite remote sensing data. Inland Waters 13, 62–73. https://doi.org/10.1080/20442041.2022.2145839.
McLaughlin, R.L., 1983. Forecasting models: sophisticated or naive? J. Forecast. 2.
Mellios, N., Moe, S.J., Laspidou, C., 2020. Machine learning approaches for predicting health risk of cyanobacterial blooms in northern European lakes. Water 12, 1191. https://doi.org/10.3390/w12041191.
Met Office, 2019. Dataset collection record: Met Office MIDAS open: UK land surface stations data, 1853-current: Dataset) [WWW Document]. URL. https://catalogue.ceda.ac.uk/uuid/dbd451271eb04662beade68da43546e1. (Accessed 9 December 2023).
Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., 2008. Stationarity is dead: whither water management? Science 319, 573–574. https://doi.org/10.1126/science.1151915.
Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., Stouffer, R.J., Dettinger, M.D., Krysanova, V., 2015. On critiques of "stationarity is dead: whither water management?". Water Resour. Res. 51, 7785–7789. https://doi.org/10.1002/2015WR017408.
Muttil, N., Chau, K.-W., 2006. Neural network and genetic programming for modelling coastal algal blooms. Int. J. Environ. 28, 223–238.
NRFA, 2023a. NRFA Station Data for 73014 - Brathay at Jeffy Knotts [WWW Document]. Natl. River Flow Arch. URL. https://nrfa.ceh.ac.uk/data/station/info/73014.
NRFA, 2023b. NRFA Station Data for 73013 - Rothay at Miller Bridge House [WWW Document]. URL. https://nrfa.ceh.ac.uk/data/station/info/73013. (Accessed 9 December 2023).
Onderka, M., 2007. Correlations between several environmental factors affecting the bloom events of cyanobacteria in Liptovska Mara reservoir (Slovakia)—a simple regression model. Ecol. Model. 209, 412–416. https://doi.org/10.1016/j.ecolmodel.2007.07.028.
O'Neil, J., Davis, T., Burford, M., Gobler, C., 2012. The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. Harmful Algae 14, 313–334. https://doi.org/10.1016/j.hal.2011.10.027.
O'Reilly, C.M., Sharma, S., Gray, D.K., Hampton, S.E., Read, J.S., Rowley, R.J., Schneider, P., Lenters, J.D., McIntyre, P.B., Kraemer, B.M., Weyhenmeyer, G.A., Straile, D., Dong, B., Adrian, R., Allan, M.G., Anneville, O., Arvola, L., Austin, J., Bailey, J.L., Baron, J.S., Brookes, J.D., de Eyto, E., Dokulil, M.T., Hamilton, D.P., Havens, K., Hetherington, A.L., Higgins, S.N., Hook, S., Izmest'eva, L.R., Joehnk, K. D., Kangur, K., Kasprzak, P., Kumagai, M., Kuusisto, E., Leshkevich, G., Livingstone, D.M., MacIntyre, S., May, L., Melack, J.M., Mueller-Navarra, D.C., Naumenko, M., Noges, P., Noges, T., North, R.P., Plisnier, P.-D., Rigosi, A., Rimmer, A., Rogora, M., Rudstam, L.G., Rusak, J.A., Salmaso, N., Samal, N.R., Schindler, D.E., Schladow, S.G., Schmid, M., Schmidt, S.R., Silow, E., Soylu, M.E., Teubner, K., Verburg, P., Voutilainen, A., Watkinson, A., Williamson, C.E., Zhang, G., 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophys. Res. Lett. 42 (10). https://doi.org/10.1002/2015GL066235, 773-10,781.
Paerl, H., Huisman, J., 2009. Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. Environ. Microbiol. Rep. 1, 27–37. https://doi.org/10.1111/j.1758-2229.2008.00004.x.
Paerl, H., Otten, T., 2013. Harmful cyanobacterial blooms: causes, consequences, and controls. Microb. Ecol. 65, 995–1010. https://doi.org/10.1007/s00248-012-0159-y.
Paerl, H., Paul, V.J., 2012. Climate change: links to global expansion of harmful cyanobacteria. Water Res. 46, 1349–1363. https://doi.org/10.1016/j.watres.2011.08.002.
Page, T., Smith, P.J., Beven, K.J., Jones, I.D., Elliott, J.A., Maberly, S.C., Mackay, E.B., De Ville, M., Feuchtmayr, H., 2018. Adaptive forecasting of phytoplankton communities. Water Res. 134, 74–85. https://doi.org/10.1016/j.watres.2018.01.046.
Park, J., Kim, K.T., Lee, W.H., 2020. Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality. Water 12, 510. https://doi.org/10.3390/w12020510.
Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41. https://doi.org/10.1016/j.scitotenv.2014.09.005.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Peretyatko, A., Teissier, S., Backer, S.D., Triest, L., 2010. Assessment of the risk of cyanobacterial bloom occurrence in urban ponds: probabilistic approach. Ann. Limnol. - Int. J. Limnol. 46, 121–133. https://doi.org/10.1051/limn/2010009.
Pobel, D., Robin, J., Humbert, J.-F., 2011. Influence of sampling strategies on the monitoring of cyanobacteria in shallow lakes: lessons from a case study in France. Water Res. 45, 1005–1014. https://doi.org/10.1016/j.watres.2010.10.011.
Ramsbottom, A.E., 1976. Depth Charts of the Cumbrian Lakes. Freshwater Biological Association.
Recknagel, F., French, M., Harkonen, P., Yabunaka, K.-I., 1997a. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28. https://doi.org/10.1016/S0304-3800(96)00049-X.
Recknagel, F., French, M., Harkonen, P., Yabunaka, K.-I., 1997b. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28. https://doi.org/10.1016/S0304-3800(96)00049-X.
Reynolds, C.S., 2006. The Ecology of Phytoplankton, Ecology, Biodiversity and Conservation. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511542145.
Richardson, J., Miller, C., Maberly, S.C., Taylor, P., Globevnik, L., Hunter, P., Jeppesen, E., Mischke, U., Moe, S.J., Pasztaleniec, A., Søndergaard, M., Carvalho, L., 2018. Effects of multiple stressors on cyanobacteria abundance vary with lake type. Glob. Change Biol. 24, 5044–5055. https://doi.org/10.1111/gcb.14396.
Rose, V., Rollwagen-Bollens, G., Bollens, S.M., Zimmerman, J., 2021. Effects of grazing and nutrients on phytoplankton blooms and microplankton assemblage structure in four temperate lakes spanning a eutrophication gradient. Water 13, 1085. https://doi.org/10.3390/w13081085.
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 Res. 182. https://doi.org/10.1016/j.watres.2020.115959.
Sarker, I.H., 2021a. Machine learning: algorithms, real-world applications and research directions. SN Comput. Sci. 2, 160. https://doi.org/10.1007/s42979-021-00592-x.
Sarker, I.H., 2021b. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput. Sci. 2, 420. https://doi.org/10.1007/s42979-021-00815-1.
Schmidt, R.M., 2019. Recurrent neural networks (RNNs): a gentle introduction and overview. https://doi.org/10.48550/arXiv.1912.05911.
Sellner, K.G., Doucette, G.J., Kirkpatrick, G.J., 2003. Harmful algal blooms: causes, impacts and detection. J. Ind. Microbiol. Biotechnol. 30, 383–406. https://doi.org/10.1007/s10295-003-0074-9.
Shyalika, C., Wickramarachchi, R., Sheth, A., 2023. A comprehensive survey on rare event prediction. https://doi.org/10.48550/arXiv.2309.11356.
Sommer, U., Adrian, R., De Senerpont Domis, L., Elser, J.J., Gaedke, U., Ibelings, B., Jeppesen, E., Lürling, M., Molinero, J.C., Mooij, W.M., van Donk, E., Winder, M., 2012. Beyond the plankton ecology group (PEG) model: mechanisms driving plankton succession. Annu. Rev. Ecol. Evol. Syst. 43, 429–448. https://doi.org/10.1146/annurev-ecolsys-110411-160251.
Soranno, P.A., 1997. Factors affecting the timing of surface scums and epilimnetic blooms of blue-green algae in a eutrophic lake. Can. J. Fish. Aquat. Sci. 54, 1965–1975. https://doi.org/10.1139/f97-104.
Spaulding, S.A., Platt, L.R.C., Murphy, J.C., Covert, A., Harvey, J.W., 2024. Chlorophyll a in lakes and streams of the United States (2005–2022). Sci. Data 11, 611. https://doi.org/10.1038/s41597-024-03453-3.
Stockwell, J.D., Doubek, J.P., Adrian, R., Anneville, O., Carey, C.C., Carvalho, L., De Senerpont Domis, L.N., Dur, G., Frassl, M.A., Grossart, H., Ibelings, B.W., Lajeunesse, M.J., Lewandowska, A.M., Llames, M.E., Matsuzaki, S.S., Nodine, E.R., Nõges, P., Patil, V.P., Pomati, F., Rinke, K., Rudstam, L.G., Rusak, J.A., Salmaso, N., Seltmann, C.T., Straile, D., Thackeray, S.J., Thiery, W., Urrutia-Cordero, P., Venail, P., Verburg, P., Woolway, R.I., Zohary, T., Andersen, M.R., Bhattacharya, R., Hejzlar, J., Janatian, N., Kpodonu, A.T.N.K., Williamson, T.J., Wilson, H.L., 2020. Storm impacts on phytoplankton community dynamics in lakes. Glob. Change Biol. 26, 2756–2784. https://doi.org/10.1111/gcb.15033.
Stroom, J.M., Kardinaal, W.E.A., 2016. How to combat cyanobacterial blooms: strategy toward preventive lake restoration and reactive control measures. Aquat. Ecol. 50, 541–576. https://doi.org/10.1007/s10452-016-9593-0.
Talib, A., Recknagel, F., Cao, H., van der Molen, D.T., 2008. Forecasting and explanation of algal dynamics in two shallow lakes by recurrent artificial neural network and hybrid evolutionary algorithm. Math. Comput. Simul. 78, 424–434. https://doi.org/10.1016/j.matcom.2008.01.037. Special Issue: Selected Papers of the MSSANZ/IMACS 16th Biennial Conference on Modelling and Simulation, Melbourne, Australia, 12-15 December 2005.
Tanguy, M., Dixon, H., Prosdocimi, I., Morris, D.G., Keller, V.D.J., 2021. Gridded estimates of daily and monthly areal rainfall for the United Kingdom (1890-2019) [CEH-GEAR] (Dataset) [WWW Document]. EIDC. URL. https://catalogue.ceh.ac.uk/id/dbf13dd5-90cd-457a-a986-f2f9dd97e93c. (Accessed 9 December 2023).
Teles, L.O., Vasconcelos, V., Pereira, E., Saker, M., 2006. Time series forecasting of cyanobacteria blooms in the Crestuma Reservoir (Douro River, Portugal) using artificial neural networks. Environ. Manage. 38, 227–237. https://doi.org/10.1007/s00267-005-0074-9.
Thomas, M.K., Fontana, S., Reyes, M., Kehoe, M., Pomati, F., 2018. The predictability of a lake phytoplankton community, over time-scales of hours to years. Ecol. Lett. 21, 619–628. https://doi.org/10.1111/ele.12927.
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 Resour. Res. 56, e2019WR026138. https://doi.org/10.1029/2019WR026138.
Torres, R., Pereira, E., Vasconcelos, V., Teles, L.O., 2011. Forecasting of cyanobacterial density in Torrão reservoir using artificial neural networks. J. Environ. Monit. 13, 1761–1767. https://doi.org/10.1039/C1EM10127G.
Velo-Suárez, L., Gutiérrez-Estrada, J.C., 2007. Artificial neural network approaches to one-step weekly prediction of Dinophysis acuminata blooms in Huelva (Western Andalucía, Spain). Harmful Algae 6, 361–371. https://doi.org/10.1016/j.hal.2006.11.002.

Wang, J.-H., Yang, C., He, L.-Q.-S., Dao, G.-H., Du, J.-S., Han, Y.-P., Wu, G.-X., Wu, Q.-Y., Hu, H.-Y., 2019. Meteorological factors and water quality changes of Plateau Lake Dianchi in China (1990–2015) and their joint influences on cyanobacterial blooms. Sci. Total Environ. 665, 406–418. https://doi.org/10.1016/j.scitotenv.2019.02.010.

Wei, J., Wang, M., Chen, C., Wu, H., Lin, L., Li, M., 2020. Seasonal succession of phytoplankton in two temperate artificial lakes with different water sources. Environ. Sci. Pollut. Res. 27, 42324–42334. https://doi.org/10.1007/s11356-020-10387-x.

Welk, A., Recknagel, F., Cao, H., Chan, W.-S., Talib, A., 2008. Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms. Ecol. Inform. 3, 46–54. https://doi.org/10.1016/j.ecoinf.2007.12.002.

Werther, M., Odermatt, D., Simis, S.G.H., Gurlin, D., Jorge, D.S.F., Loisel, H., Hunter, P. D., Tyler, A.N., Spyrakos, E., 2022. Characterising retrieval uncertainty of chlorophyll-*a* algorithms in oligotrophic and mesotrophic lakes and reservoirs. ISPRS J. Photogramm. Remote Sens. 190, 279–300. https://doi.org/10.1016/j.isprsjprs.2022.06.015.

Wolf, D., Georgic, W., Klaiber, H.A., 2017. Reeling in the damages: harmful algal blooms' impact on Lake Erie's recreational fishing industry. J. Environ. Manage. 199, 148–157. https://doi.org/10.1016/j.jenvman.2017.05.031.

Woolway, R.I., Merchant, C.J., 2019. Worldwide alteration of lake mixing regimes in response to climate change. Nat. Geosci. 12, 271–276. https://doi.org/10.1038/s41561-019-0322-x.

Xiao, X., He, J., Huang, H., Miller, T.R., Christakos, G., Reichwaldt, E.S., Ghadouani, A., Lin, S., Xu, X., Shi, J., 2017. A novel single-parameter approach for forecasting algal blooms. Water Res. 108, 222–231. https://doi.org/10.1016/j.watres.2016.10.076.

Xiao, X., Peng, Y., Zhang, W., Yang, X., Zhang, Z., ren, B., Zhu, G., Zhou, S., 2024. Current status and prospects of algal bloom early warning technologies: a Review. J. Environ. Manage. 349, 119510. https://doi.org/10.1016/j.jenvman.2023.119510.

Xue, K., Ma, R., Shen, M., Wu, J., Hu, M., Guo, Y., Cao, Z., Xiong, J., 2023. Horizontal and vertical migration of cyanobacterial blooms in two eutrophic lakes observed from the GOCI satellite. Water Res. 240, 120099. https://doi.org/10.1016/j.watres.2023.120099.