Near-term forecasting of cyanobacterial and harmful algal blooms in lakes using simple univariate methods with satellite remote sensing data

Mark William Matthews[a,b]*

[a]*CyanoLakes (Pty) Ltd, Cape Town, South Africa;* [b]*CyanoLakes Australia, Sydney, Australia*

**mark@cyanolakes.com*

Near-term forecasting of cyanobacteria and harmful algal blooms (HABs) in lakes is essential to reduce risks to human and animal health and water treatment. Cyanobacteria forecasting models are typically complex, requiring input of biophysical and chemical measurements or DNA sequencing in situ. Satellite imagery presents a unique opportunity to estimate cyanobacteria concentration directly at low cost and over wide spatial and long timescales. This study explores the hypothesis that simple univariate forecasting methods can reliably forecast cyanobacterial blooms in the near-term (one week ahead) detected using satellite remote sensing. A simple univariate model based on logical decomposition with a moving average and seasonal component was developed to forecast chlorophyll-*a* concentrations from cyanobacteria and algal blooms in lakes using spatially-aggregated satellite remotely sensed data. A small test set of fifteen spatially distributed waterbodies was used to assess forecast performance on 1-week, 2-week and 4-week forecast horizons using a year-long hold-out timeseries. For a 1-week time horizon, cyanobacterial blooms posing a high health risk could be forecast with 80% accuracy. The 2-week and 4-week forecast accuracy dropped off to 71% and 69%, respectively. Forecast performance was only weakly influenced by lake size, suggesting that the spatial-aggregation approach may be valid even for large lakes. Additionally, longer time series reduced the observed forecast error presumably due to better seasonal characterization. The study is the first to demonstrate that simple univariate models with remotely sensed timeseries can forecast cyanobacteria and HABs with almost the same reliability as complex models.

Keywords: forecasting; cyanobacterial blooms; algal blooms; satellite remote sensing; HABs

1. Introduction

The CyanoLakes web and mobile applications (“Your Weather App for Lakes” – the App) for iOS and Android provide weather-like information for cyanobacterial and algal blooms from satellite remote sensing for estimating human health risks and eutrophication in the world’s lakes (www.cyanolakes.com). The satellite imagery is processed in near real-time (two to three hours after acquisition) providing same-day updates with notifications. The Apps are useful to water utility companies (water managers) that experience taste, odor, and toxin problems because of cyanobacterial or algal blooms in their source waters; and to recreational water users who benefit from health risk advisories. Although not all cyanobacteria produce toxins, they are considered a nuisance for water treatment due to the production of compounds causing taste and odor problems.

There is a need to provide forecasts through the App to water managers to enable early implementation of in-lake and in-plant remedial actions and for recreational water users to prevent exposure to cyanobacterial blooms or poor water quality (Recknagel et al. 2017; Carey et al. 2020; Rousso et al. 2020). The forecast horizon that is most used by water managers is 7-days and a near-term forecast horizon of 1-week or less is useful for recreational users (Rousso et al. 2020).

Methods for forecasting cyanobacterial blooms incorporating satellite remote sensing data often combine it with meteorological models and hydrodynamic models to create lake-specific forecasts on a weekly or seasonal basis (e.g., Recknagel et al. 2018 and references therein). This approach, while comprehensive, is impractical for global implementation as it cannot easily be generalized and requires often-complex models to be run simultaneously. Further, these models often rely on in situ datasets which negates the benefits of remote sensing. One study that does not rely on in situ data is Myer et al (2020) who derived a model for high-risk cyanobacterial blooms in 102 lakes in Florida. Their approach used spatially-aggregated estimates of cyanobacteria cell counts from the Ocean and Land Color Instrument (OLCI) from the Copernicus Sentinel-3 satellite, with additional forecast variables including surface water temperature, precipitation, air temperature and lake geomorphology. Their model could predict the probability of high-risk cyanobacterial blooms with an 82% accuracy for a 1-week forecast horizon. However, they did not forecast the actual concentration of cyanobacteria cells, nor did they test forecast performance for longer time-horizons.

Non-remote sensing methods of forecasting cyanobacterial blooms utilize timeseries of in situ datasets with variables such as chlorophyll-*a*, phytoplankton cell counts, algal density, nutrient concentrations, and meteorological data. Forecasts are produced using complex models or statistical methods such as multi-variate regression analysis and neural networks (see Rousso et al. 2020, Echard 2021 and references therein). Hydrodynamic forecasting tools use water temperature and meteorology sensors to forecast water temperature profiles to predict the onset of autumn turnover, which is a significant trigger for bloom formation (e.g., Thomas et al. 2019; Carey et al. 2022). Three-dimensional coupled hydro-algal biomass models also exist that use autonomous measurements of algal density and meteorological variables (e.g., Li et al., 2013; Qin et al., 2015). These models are largely site-specific and require continuous in situ monitoring (usually via autonomous sensors) that may be time-consuming and costly. Alternative approaches based on sequencing the DNA of the microbial community have a high degree of predictive success, but again rely on complex measurements (e.g., Tromas et al. 2017).

These approaches neglect the possibility of utilizing simple univariate forecasting methods from remote sensing data which are often successfully used for meteorological or commercial applications (e.g., Ji and Peters 2004; White and Nemani 2006; Peng and Chu 2009; Caillault and Bigand 2018). A growing number of recent publications have demonstrated the utility of single variable models, primarily using deep learning approaches, for forecasting algal blooms on hourly, daily or monthly time-scales using in situ measurements of algal density (e.g., Xiao et al., 2017; Liu et al., 2022; Shan et al., 2022). Forecasting from satellite data using pixel values aggregated by area (White and Nemani 2006) that has a consistent measurement basis, long time-series and near real-time availability seems a favorable approach for near-term forecasting of cyanobacterial blooms.

Further, given that cyanobacterial blooms exhibit strong seasonality in response to turnover, light availability and temperature, a phenology-based approach shows promise (e.g., White and Nemani 2006; García-Mozo et al., 2008, Peng and Chu 2009). Bloom phenology in lakes is strongly controlled by the timing of the spring or autumn turnover which releases nutrients from the lake bottom that are mixed into the photic zone (Nürnberg 1988). The bacterial community also exhibits predictable cyclical changes which further drives seasonality and predictability of blooms (Tromas et al. 2017). Logical forecasting approaches known as classical seasonal decomposition, decompose univariable timeseries into error, trend and seasonal components (for detailed methods see Hyndman and Athanasopoulos 2018; Svetunkov 2021). Simple decomposition and moving-average techniques are likely to have some success for near-term forecasts given that the current inoculation (i.e., today's value) of cyanobacteria cells is the strongest predictor of tomorrow's value (Swanepoel et al. 2016).

This paper aims to develop a simple univariate approach for forecasting cyanobacteria and algal blooms from satellite remote sensing timeseries that could be generalized to thousands of lakes worldwide. This necessitates an approach that is:

- computationally simple and light-weight (fast)
- applicable across very diverse lake systems (generalized)
- does not require model parameterization (logical)
- does not require training of complex algorithms using external data (independent)

Further two hypotheses are tested, firstly that forecast performance is sensitive to lake size (larger lakes are more challenging to forecast owing to increased heterogeneity), and secondly that a longer time-series of historical data to determine seasonal variability leads to improved forecasts (more data leads to better forecasts). This study is the first of its kind to test the feasibility for using simple univariate methods based exclusively on satellite remote sensing data for forecasting cyanobacterial and algal blooms in the world's lakes.

2. Methods

2.1 Satellite-derived data

The App uses data from OLCI onboard the Sentinel-3A and Sentinel-3B satellites. OLCI provides synoptic coverage of the globe with up to 6 updates per week for any site on earth, at a spatial resolution of 300 m, with 21 spectral bands with high signal-to-noise ratios that are a requirement for water color related applications and cyanobacteria quantification and discrimination. The spatial resolution allows for lakes larger than approx. 1 km^2 to be resolved, which accounts for roughly 60% of the total surface area of the world's lakes or more than 350 000 lakes worldwide (calculated from data in Verpoorter et al. 2014).

Estimation of chlorophyll-a concentration and the detection of cyanobacteria are performed using the maximum peak-height (MPH) algorithm (Matthews and Odermatt 2015). The accuracy of cyanobacteria detection has been determined from *in situ* datasets in Australia and South Africa at greater than 70% (unpublished), and the error for chlorophyll-*a* is around 50%, which is acceptable for a log-scaled variable that routinely ranges over four orders of magnitude from 0.1 to 1000 µg/L in lakes. The algorithm provides chlorophyll-*a* values for cyanobacterial and algal blooms, as well as a binary mask for the presence of cyanobacteria. Cyanobacteria are detectable at concentrations exceeding approx. 5 μg/L given the signal-to-noise sensitivity of the satellite instrument.

The image data are spatially-aggregated using mean or median values for the lake surface area to produce two variables that are forecasted:

(1) Chla_cyano – the mean whole-lake value of Chl-*a* for pixels identified as cyanobacteria
(2) Chla_med – the median whole-lake value of Chl-*a* for both cyanobacteria and non-cyanobacteria pixels

Chla_cyano represents the average biomass of cyanobacteria detected in the lake for pixels identified as cyanobacteria. Chla_med represents the average algal and cyanobacteria biomass for the lake at a moment in time. In addition to these variables, the App provides two main indices for lake management which are derived based on thresholds from the forecast variables:

(1) the cyanobacteria risk level based on World Health Organization guidelines (WHO 2003); and,
(2) the trophic status levels based on OECD guidelines

The cyanobacteria risk level (CRL) thresholds are defined as:

- Low risk: Chla_cyano < 10 µg/L, up to 20 000 cells/mL or 4 µg/L microcystin[1]
- Medium risk: Chla_cyano between 10 and 50 µg/L, up to 100 000 cells/mL or 20 µg/L microcystin
- High risk: Chla_cyano between 50 and 100 µg/L, up to 200 000 cells/mL or 40 µg/L microcystin
- Very high risk: Chla_cyano > 100 µg/L, more than 200 000 cells/mL or 40 µg/L microcystin

The trophic state (TS) class thresholds are defined as:

- Low: Chla_med < 10 µg/L
- Medium: Chla_med between 10 and 20 µg/L
- High: Chla_med between 20 and 50 µg/L
- Very high: Chla_med larger than 50 µg/L

The forecast accuracy was calculated with regards to the two indices by calculating the percentage agreement between observed and forecast values (i.e. the agreement between the forecast and observed values in the test dataset).

2.2 Study areas

This study uses timeseries from a cross-section of fifteen lakes mainly from mid-latitude regions and known to be affected by seasonal or permanent cyanobacterial blooms (Figure 1, Table 1). The lakes (and/or reservoirs) are mostly from South Africa and Australia where the satellite data have been corroborated with *in situ* datasets (Shah

[1] 1 µg/L chlorophyll-a equivalent to 2 000 cells/mL or 0.4 µg/L microcystin (World Health Organization 2003)

et al. 2021; Lacey et al. 2021; Kravitz et. al. 2020; Lawrence et al. 2020), as well as Italy (Trasimeno), the USA (Clear Lake) and China (Taihu). Most of the lakes have been chosen as they are known to be affected by seasonal or permanent cyanobacterial blooms, evident by Chla_cyano values (see Table 1). This includes lakes such as Clear Lake, Copeton, Hartbeespoort (also affected by floating vegetation), Roodeplaat, Trasimeno, Taihu and Vaal. Several lakes exhibiting infrequent cyanobacterial blooms function as controls (e.g., Burragorang, Grahamstown, Midmar and Zeekoevlei). The lakes also vary according to trophic class: oligotrophic lakes include Burragorang, Grahamstown, Midmar and Trasimeno, while hypertrophic lakes include Zeekoevlei, Roodeplaat and Vaal.

The number of lakes selected is somewhat random but represents a large-enough sample to draw conclusions using regression statistics. The lakes vary in size from just 2,6 km^2 (Zeekoevlei) to a massive 3000 km^2 (Taihu). The timeseries differ in length and frequency, with start dates ranging from October 2016 to July 2019 (Clear Lake), and number of acquisitions from 435 (Taihu) to 949 (Vaal). Data frequencies differ mainly due to regional cloud cover differences. These differences are intentionally chosen to assess relationships between forecast performance and data quantity and target size. All timeseries data were available until 15 November 2021.

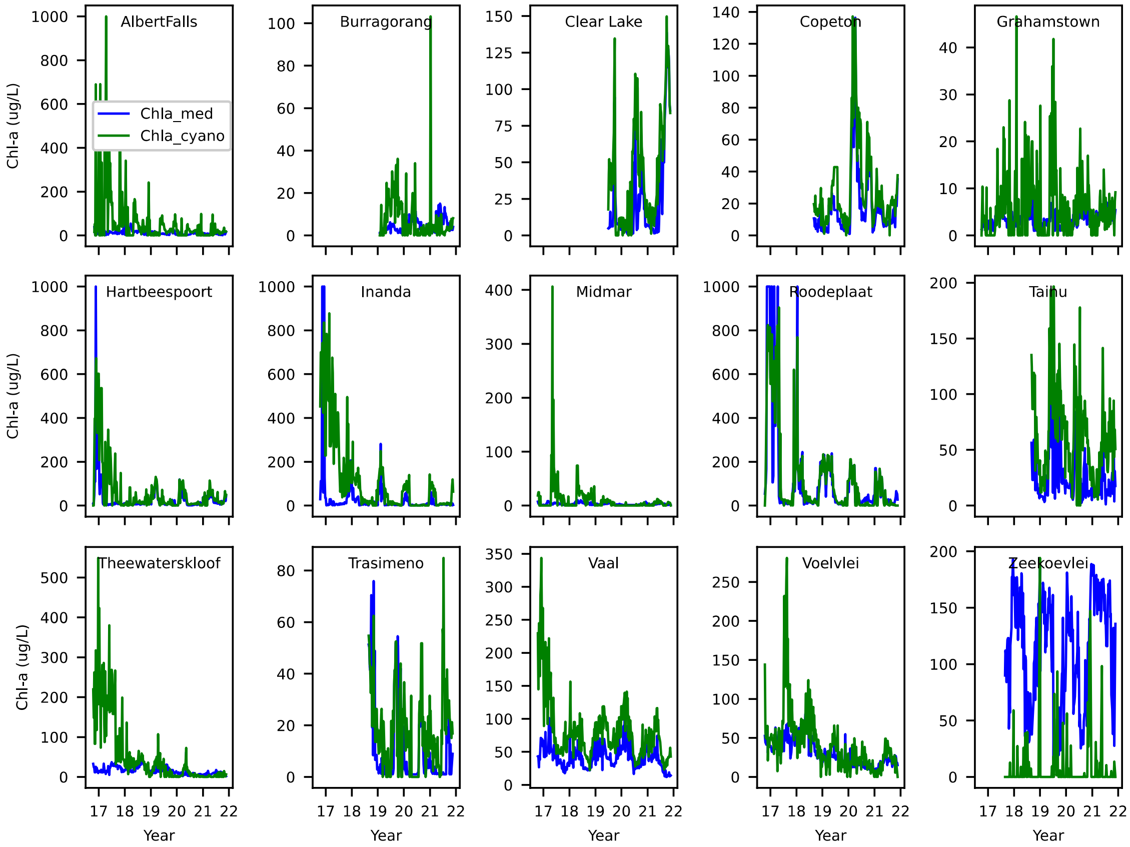

Figure 1. Timeseries of weekly averages for Chla_med and Chla_cyano between 2017 to 2022. Timeseries vary intentionally in duration. Note differences on y-axis.

2.3 Forecasting approach

Forecasts were derived at weekly (7-day), 2-weekly (14-day) and 4 weekly (28-day) time-horizons. These intervals are likely the most useful horizons for near-term forecasting for management and/or recreational purposes when combined with updated near real-time data (Rousso et al. 2020, Carey et al. 2022). The satellite timeseries are sporadic (not regular interval) due to the presence of cloud and the nature of the satellite acquisitions. The timeseries were aggregated (downscaled) from sporadic daily updates to weekly averages (mean) and backfilled (using prior values, not interpolation) where data were incomplete.

Table 1. Characteristics of lakes chosen in this study, showing mean values for Chla_med and Chla_cyano and the start and end dates and number of acquisitions of the satellite timeseries.

Name	Lat	Lon	Area (km sq.)	Chla_med (µg/L)	Chla_cyano (µg/L)	Start	End	No.
AlbertFalls	-29,44	30,40	31	10,1	50,0	2016-10-26	2021-11-15	749
Burragorang	-34,00	150,40	68	4,8	8,3	2019-01-25	2021-11-15	534
Clear Lake	39,04	-122,81	265,6	26,4	46,8	2019-06-27	2021-11-14	585
Copeton	-29,91	150,99	19,9	19,2	28,1	2018-08-27	2021-11-15	623
Grahamstown	-32,74	151,81	27	3,5	6,0	2016-10-02	2021-11-15	722
Hartbeespoort	-25,75	27,86	21,4	26,4	49,1	2016-10-14	2021-11-15	861
Inanda	-29,68	30,85	15,9	19,5	75,1	2016-10-15	2021-11-15	683
Midmar	-29,51	30,18	20,4	2,8	9,8	2016-10-15	2021-11-15	748
Roodeplaat	-25,63	28,36	4,6	92,5	92,9	2016-10-15	2021-11-15	778
Taihu	31,27	120,18	3033,7	25,8	62,4	2018-09-02	2021-11-15	435
Theewaterskloof	-34,03	19,20	33,3	15,1	44,3	2016-10-14	2021-11-15	922
Trasimeno	43,14	12,10	222,3	9,8	18,0	2018-08-23	2021-11-11	577
Vaal	-26,97	28,25	349,6	40,8	75,7	2016-10-15	2021-11-15	949
Voelvlei	-33,37	19,04	21,5	28,0	37,0	2016-10-14	2021-11-15	913
Zeekoevlei	-34,06	18,51	2,6	115,7	9,0	2017-08-26	2021-11-15	675

The timeseries were split into training and test (hold out) datasets. The test dataset consisted of a 1-year timeseries from November 15, 2020 to November 15, 2021, while the training dataset consisted of all data prior to November 15, 2020. Forecasts were calculated using a rolling-origin technique for a period of one year using the training set to produce seasonal information and the test dataset to test forecast performance. In this manner, only past data were used to determine the seasonal components (lake ‘climatology’). Forecast performance was assessed using the root mean square error (RMSE) which was determined as the optimal measure of performance for the type of data under consideration (other metrics such as mean average error and mean average percent error are not shown). R-squared values (obtained by linear correlation), despite commonly being used to test the performance of lake-related forecasts, should not be used because they do not account for bias in the forecast result.

Common univariate forecasting methods were used as a benchmark against which other models were assessed. The benchmarks included naïve, moving average, seasonal naïve, exponential smoothing, and trend adjusted exponential smoothing techniques (see Svetunkov 2021 and Hyndman and Athanasopoulos 2018). There is little justification for using complex models if they are outperformed using these simple methods, provided the performance of the simple models is satisfactory for the application in question (e.g., Peng and Chu 2009)

The naïve forecast (denoted NA) was determined by:

$$\widehat{y_t} = y_{t-1} \tag{1}$$

where $\widehat{y_t}$ is the forecasted value at time t and y_{t-1} is the observed value of the previous period. The simple moving average (MA) forecast was determined by:

$$\widehat{y_t} = \frac{1}{m}\sum_{j=1}^{m} \hat{y}_{t-j} \tag{2}$$

where *j* is the period over which the moving average is calculated, and m is the window of previous values (a value of two was used, meaning that the average of the two previous observations was determined as the forecast value).

The seasonal naïve (SN) forecast was determined by the average value of all previous observations in the training dataset for the period, *m* (in this case the period was a week) in question:

$$\widehat{y_t} = y_{t-m} \tag{3}$$

The simple exponential smoothing (ES) model was:

$$\hat{y}_{t+1} = \hat{\alpha} y_t + (1-\hat{\alpha})\widehat{y_t} \tag{4}$$

where $\hat{\alpha}$ is the smoothing parameter between 0 and 1, and $\hat{y}_t$ is the forecast value for the previous period. The value of $\hat{\alpha}$ was determined as 0.7 by minimizing the RMSE for the entire set of timeseries.

The trend-adjusted exponential smoothing (TAES) model was:

$$\hat{y}_{t+1} = \hat{\alpha} y_t + (1-\hat{\alpha})\widehat{y_t} \tag{5}$$

$$T_{t+1} = \beta(\hat{y}_{t+1} - \widehat{y_t}) + (1-\beta)T_t \tag{6}$$

$$TAF_{t+1} = \hat{y}_{t+1} + T_{t+1} \tag{7}$$

where T is the trend, and β is the trend adjustment factor usually ranging from 0 to 1, and TAF means trend-adjusted forecast. A value of zero was used for the initial trend T_t. Varying the values of the smoothing parameters between 0.3 and 0.7 produced only slight changes in the overall RMSE, therefore values of 0.5 were used.

It was essential that any advanced model outperformed or matched the performance of the above baseline models, while also being able to be used for forecasts more than one period ahead (i.e., 2-week and 4-week forecasts). A model based on classical decomposition and on a combination of the methods above was conceived by combining the following logic:

(1) The best predictor of future cyanobacteria biomass is the current inoculation of cells which is equal to the current value or moving average
(2) Cyanobacteria exhibit strong annual seasonality that varies on a roughly weekly timescale, therefore, the value for the current week of year is likely to be similar to the value of the average of all previous years in that week (i.e., the 'lake climatology')
(3) There should be some accounting for the anomaly from the seasonal norm (i.e., deviation of the current value from the expected value based on the seasonal signal)

Using the above logic, a so-called moving average seasonal error-adjusted (MASEA) model was developed based on a combination of logical decomposition (seasonality) and a moving average approach:

$$\hat{y}_t = \hat{\alpha} * MA + (1 - \hat{\alpha})(s_t + e) \qquad (8)$$

$$e = MA - s_{t-1} \qquad (9)$$

where s_t is the seasonal average, calculated as the mean of all previous observations for period *t*, and *e* is the seasonal anomaly calculated as the difference between the moving average and the seasonal average in the previous period, and $\hat{\alpha}$ is the weighting factor adjusting the contribution of the anomaly-adjusted seasonal value and the moving average.

A centralized moving average was used to compute the seasonal weekly component. The MASEA model is based on a combined moving average and seasonal approach and enables forecasting at 2- and 4-week horizons. The weighting factor was adjusted to determine the optimal values for 1-week, 2-week and 4-week forecast horizons, and determined as 0.8, 0.7 and 0.6, respectively, with as expected, the longer forecast-horizons being more heavily weighted towards the value of the seasonal average.

3. Results

3.1 Performance of baseline models

All the baseline models, except for the seasonal naïve model, performed similarly for forecasting Chla_cyano and Chla_med on a weekly forecast horizon, with the exponential smoothing model performing the best overall (

Table 2, Table 3). The SN model was significantly worse than models using the observed value of the previous period, giving weight to the theory that the current cyanobacteria inoculation is the primary forecast variable, with annual seasonality, in general, playing a lesser, but sometimes important role (SN performed best for 2 lakes).

Table 2 RMSE (µg/L) of 1-we ek Chla_cyano forecasts for baseline models. Color scale indicates best (green) and worst (red) forecasts for each lake.

Waterbody	NA	MA	SN	ES	TAES	MASEA
AlbertFalls	30.6	24.9	54.4	26.1	26.8	25.4
Burragorang	20.3	17.5	12.0	17.8	18.5	16.8
Clear Lake	14.5	14.6	23.7	14.2	14.0	14.4
Copeton	6.7	6.7	17.2	6.4	6.6	7.0
Grahamstown	3.7	3.8	4.0	3.4	3.7	3.7
Hartbeespoort	22.8	22.7	51.8	21.6	22.2	22.0
Inanda	26.0	27.1	100.5	25.8	26.4	29.3
Midmar	3.0	2.6	12.9	2.7	2.7	2.8
Roodeplaat	40.1	44.0	117.5	40.6	42.4	39.6
Taihu	26.6	25.3	24.4	24.2	25.3	24.4
Theewaterskloof	5.6	3.9	51.8	4.6	4.6	5.6
Trasimeno	13.2	12.0	13.8	12.3	12.5	11.8
Vaal	16.6	15.2	27.3	15.0	15.2	14.8
Voelvlei	9.2	9.2	26.4	8.7	8.9	9.7
Zeekoevlei	24.2	26.4	23.6	24.5	25.0	26.3
Mean RMSE	17.5	17.1	37.4	16.5	17.0	16.9

Table 3 RMSE (µg/L) of 1-week Chla_med forecasts for baseline models. Color scale indicates best (green) and worst (red) forecasts for each lake.

Waterbody	NA	MA	SN	ES	TAES	MASEA
AlbertFalls	5.4	5.7	6.2	5.4	5.7	5.5
Burragorang	3.0	3.0	2.5	2.8	2.9	2.9
Clear Lake	13.1	15.1	28.0	13.9	14.0	15.2
Copeton	5.5	6.2	13.7	5.6	5.8	6.5
Grahamstown	1.6	1.4	1.4	1.4	1.5	1.4
Hartbeespoort	14.3	14.1	31.8	13.4	13.9	14.2
Inanda	11.0	11.2	45.4	10.4	10.8	12.3
Midmar	1.5	1.4	1.1	1.3	1.4	1.3
Roodeplaat	32.1	35.3	112.8	32.2	34.2	30.9
Taihu	18.7	18.1	18.0	17.3	18.0	17.8
Theewaterskloof	2.7	2.6	7.8	2.6	2.7	2.7
Trasimeno	4.7	4.4	7.9	4.4	4.4	4.1
Vaal	10.6	9.4	12.0	9.3	9.4	9.4
Voelvlei	4.3	4.7	14.2	4.3	4.5	4.8
Zeekoevlei	28.7	28.4	45.9	27.3	27.8	28.0
Mean RMSE	10.5	10.7	23.3	10.1	10.5	10.5

3.2 Performance of MASEA model

When compared to the baseline models, the MASEA model performs slightly more poorly than the ES model, but no worse than the other baseline models. Although it would be ideal to have a model that significantly outperforms the baselines, it is not always feasible to achieve this even with very complex and advanced models (e.g., Peng and Chu 2009). Given that more complex models were not evaluated in this study (owing to the application) a comparison with more complex models was not required.

The 1-week, 2-week and 4-week MASEA forecasts generally correlate closely with observed values (see Table 4, , Figure 3), although there is a noticeable lag in some cases between the observed values and those of the 4-week forecast. The 1-week forecast, with few exceptions, consistently has the lowest RMSE values, which become larger for the 2- and 4-week forecasts, respectively.

Table 4. Performance of MASEA forecasts for 1-week, 2-week and 4-week horizons for Chla_cyano and Chla_med (RMSE, units µg/L). Normalized 1-week RMSE values are shown on right. Color scale indicates best (green) and worst (red) forecasts for each lake.

	Chla_cyano			Chla_med				chla_cyano	chla_med
Waterbody	1wk	2wk	4wk	1wk	2wk	4wk		1wk norm.	1wk norm.
AlbertFalls	25.4	21.1	27.0	5.5	6.5	6.2		0.51	0.55
Burragorang	16.8	16.9	16.8	2.9	3.1	3.2		2.03	0.60
Clear Lake	14.4	17.5	23.1	15.2	19.6	24.2		0.31	0.57
Copeton	7.0	8.8	11.8	6.5	8.2	10.9		0.25	0.34
Grahamstown	3.7	3.8	3.4	1.4	1.5	1.4		0.62	0.40
Hartbeespoort	22.0	26.7	34.5	14.2	16.3	17.9		0.45	0.54
Inanda	29.3	37.5	41.0	12.3	14.4	19.0		0.39	0.63
Midmar	2.8	3.7	6.1	1.3	1.0	1.4		0.28	0.48
Roodeplaat	39.6	50.3	56.2	30.9	37.0	35.4		0.43	0.33
Taihu	24.4	24.9	24.5	17.8	18.8	19.6		0.39	0.69
Theewaterskloof	5.6	7.6	8.9	2.7	3.3	3.4		0.13	0.18
Trasimeno	11.8	15.2	19.7	4.1	4.0	4.3		0.66	0.42
Vaal	14.8	15.7	17.9	9.4	8.3	9.9		0.20	0.23
Voelvlei	9.7	10.9	12.4	4.8	6.1	7.5		0.26	0.17
Zeekoevlei	26.3	33.7	37.7	28.0	31.1	38.1		2.92	0.24
AVERAGE	16.9	19.6	22.7	10.5	11.9	13.5			

The disadvantage with using a 4-week forecast that is heavily weighted on a moving average, is that the forecasted values may significantly underestimate larger changes (, Figure 3). This is particularly evident for targets where there is less historical data for computing seasonal averages (e.g., Clear Lake, Copeton and Burragorang). In these cases, the model produces a more conservative 4-week estimate than the actual observed increase or decrease. However, even a conservative 4-week forecast may prove valuable, assuming that there is enough historical data to characterize the seasonal changes.

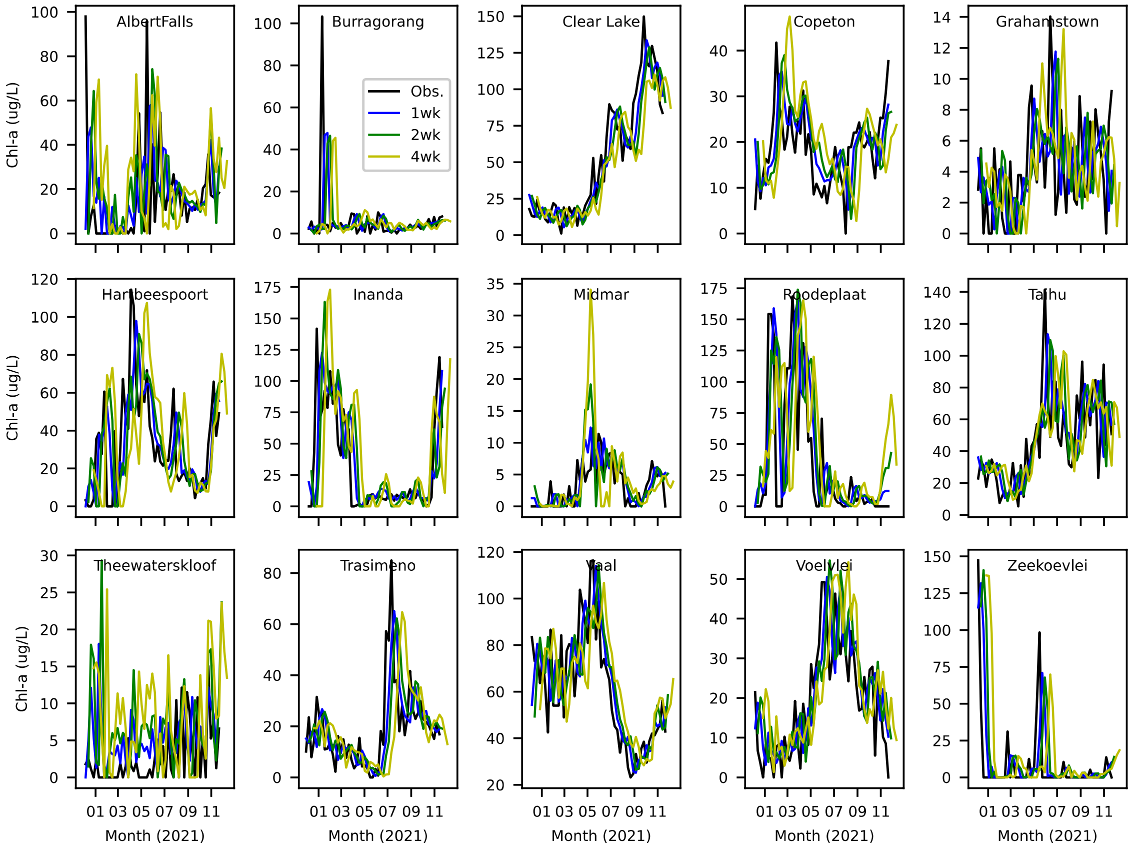

Figure 2. Rolling-origin Chla_cyano forecasts at 1-week, 2-week and 4-week horizons.

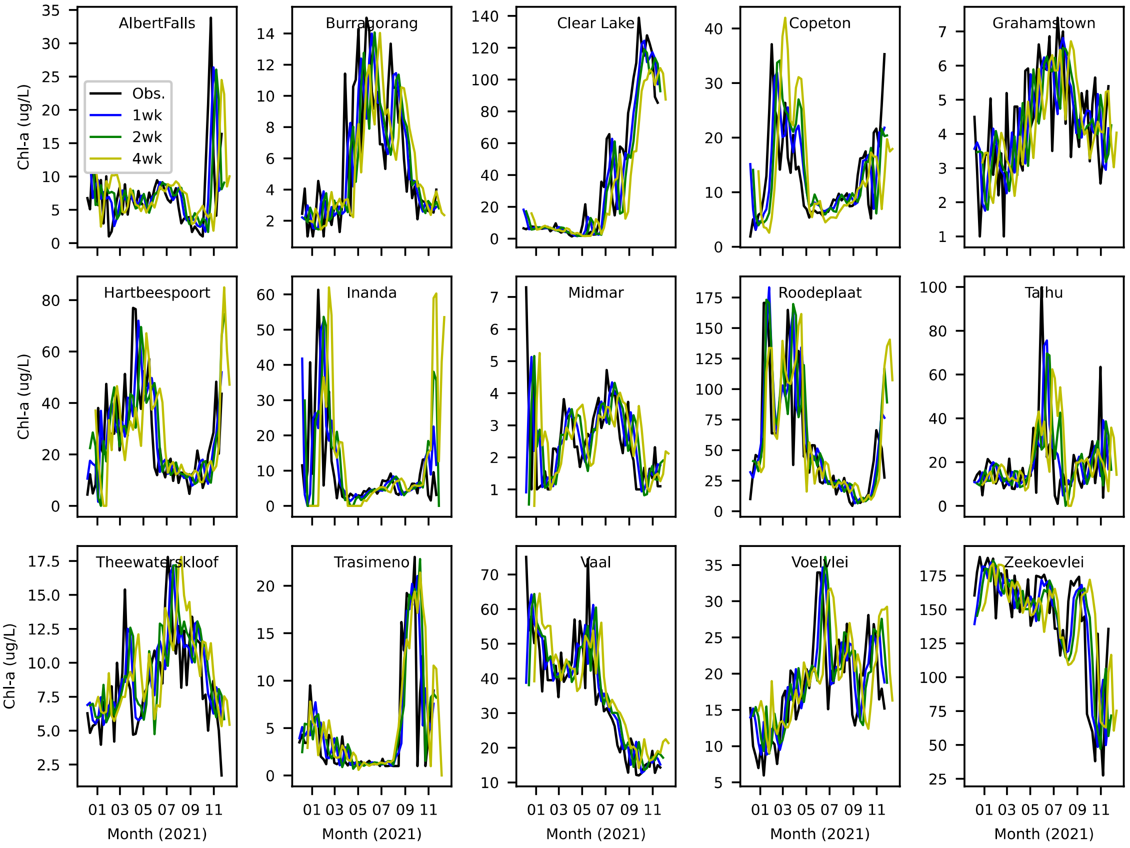

Figure 3. Rolling-origin Chla_mean forecasts at 1-week, 2-week and 4-week horizons.

With respect to Chla_cyano, the 4-week forecast was overall 5 µg/L less accurate than a 1-week forecast (RMSE of 22.7 vs 16.9 µg/L). Similarly, for Chla_med, the 4-week forecast accuracy was overall 3 µg/L less accurate than the 1-week forecast (RMSE of 13.5 vs 10.5 µg/L, respectively). This indicates that the 4-week forecast, although somewhat insensitive to sudden or larger changes, does account for seasonal changes enough for forecast performance not to drop off significantly from a 1-week horizon.

3.3 Performance for forecasting indices

The 1-week horizon forecast ability for cyanobacteria risk level and trophic state had a 74% and 75% accuracy, respectively (Table 5). This result is encouraging when considering the wide range of Chl-*a* concentrations of the lakes under consideration (Table 1). For the 4-week forecast horizon, the forecasting ability drops off to 66% for CRL and 71% for TS, respectively. The finding that a simple univariate forecasting technique can forecast CRL and TS with more than 50% accuracy 4-weeks in advance lends weight to the approach presented in this study.

Table 5. Performance of MASEA forecasts (in percentage agreement between observed and forecasted value) for 1-week, 2-week and 4-week horizons for CRL, TS and high CRL.

	CRL				TS				High CRL			
	1wk	2wk	4wk		1wk	2wk	4wk		N	1wk	2wk	4wk
AlbertFalls	66.7	50.0	50.0		78.4	84	79.2		4	0	0	25
Burragorang	94.1	94.0	89.6		80.4	78	70.8		1	0	0	0
Clear Lake	72.5	72.0	75.0		72.5	76	72.9		24	87.5	91.7	83.3
Copeton	82.4	84.0	75.0		62.7	62	56.2		0			
Grahamstown	94.1	94.0	93.8		100	100	100		0			
Hartbeespoort	54.9	50.0	39.6		54.9	56	52.1		10	60	50	40
Inanda	51.0	48.0	56.2		72.5	74	64.6		15	73.3	60	66.7
Midmar	90.2	90.0	89.6		100	100	100		0			
Roodeplaat	49.0	40.0	37.5		62.7	66	52.1		14	85.7	64.3	71.4
Taihu	70.6	68.0	60.4		51	46	45.8		19	89.5	78.9	73.7
Theewaterskloof	82.4	74.0	58.3		80.4	78	72.9		0			
Trasimeno	72.5	70.0	68.8		88.2	90	87.5		3	33.3	0	0
Vaal	82.4	78.0	70.8		62.7	74	64.6		30	96.7	90	86.7
Voelvlei	72.5	70.0	64.6		64.7	68	52.1		0			
Zeekoevlei	74.5	72.0	66.7		92.2	94	95.8		3	33.3	0	0
AVERAGE	74.0	70.3	66.4		74.9	76.4	71.1					

3.4 Impact of lake size and timeseries length

There was only weak evidence for the hypothesis that forecast accuracy for Chla_med is inversely related to lake area given that larger lakes exhibit more heterogeneity (Figure 4). In general, the 1-week forecast accuracy for Chla_med and Chla_cyano was not significantly affected by lake area, as indicated by the RMSE normalized by the average value of each variable for each lake (Table 4). In contrast, it was evident that the length of historical time-series, or number of images, had a significant effect on Chla_med forecasting skill ($R^2 = 0.45$, Figure 4), with longer timeseries exhibiting lower normalized RMSE values.

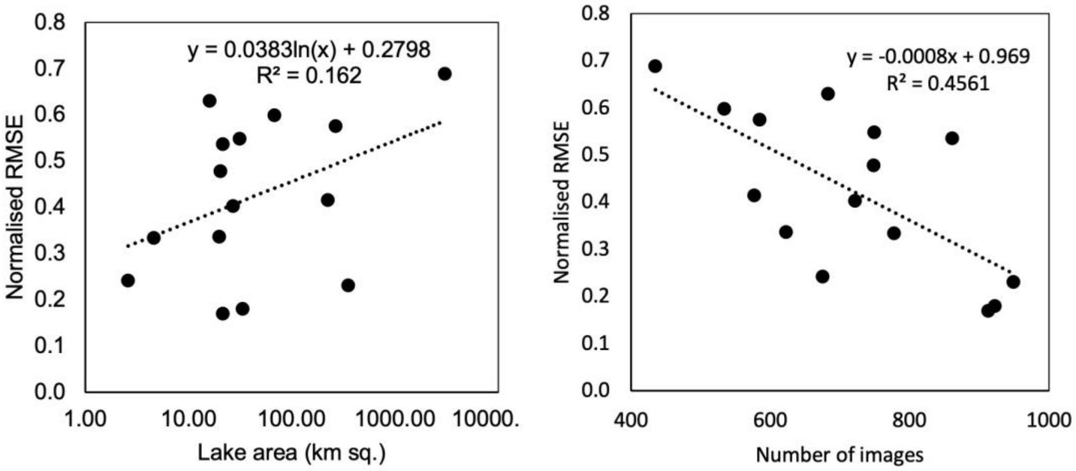

Figure 4. Relationship between 1-week forecast accuracy for Chla_med and log-scaled lake size (left) and time-series duration represented by the number of images (right).

A similar relationship, although weaker, was also observed for Chla_cyano. This confirms the hypothesis that longer time-series leads to increased forecasting skill owing to improved characterization of the seasonal component used by the MASEA model.

3.5 Forecast Accuracy for median whole-lake chlorophyll-a

On average, the 1-week forecast for Chla_med can be estimated with an error of 10.5 μg/L, although the forecast error must be contextualized by the trophic state (the typical range of chlorophyll-*a* values) for each waterbody. For an oligotrophic system, such as Grahamstown, with typical Chl-*a* values between 1 and 10 μg/L that are rarely exceeded, the forecast error was only 1.4 μg/L. For a hypertrophic system, such as Roodeplaat, which regularly exhibits chlorophyll-a concentrations exceeding 100 μg/L, the forecast accuracy was 30.9 μg/L, which also appears adequate given the context. As demonstrated by these examples, the forecast performance is lake-specific, and the values must be considered with reference to the range of values typical for the waterbody under examination – e.g., much larger RMSE values are present for waterbodies that exhibit very large ranges such as Roodeplaat and Hartbeespoort. Examining the RMSE of each waterbody normalized by its mean Chl-*a* concentration, the overall 1-week forecast error was 42%. This value is more representative of typical forecast performance. Examining the above two examples, errors were 40% for Grahamstown and 33% for Roodeplaat, respectively. This value can compare to the sampling error for *in situ* Chl-*a* measurements, which is typically between 10 and 30%. It is unclear whether forecasting chlorophyll-*a* concentration 1-week ahead with a roughly 40% error would meet the requirements for applications such as water treatment, the issuing of recreational advisories or ecological modelling.

3.6 Forecast accuracy for cyanobacteria

Chl-*a* for cyanobacteria can be forecast with an average error of 16.9 μg/L with a 1-week forecast horizon, which increases to 22.7 μg/L for a 4-week horizon. Contextualizing this value by normalizing the result with the average Chla_cyano value for each of the waterbodies, the normalized error can be expressed as 65% for a 1-week forecast. The prevalence of cyanobacterial blooms relative to algal blooms within each of the waterbodies can be determined by examining the ratio of Chla_med to Chla_cyano. Waterbodies that exhibited the lowest Chla_med:Chla_cyano ratios (e.g., Zeekoevlei, Burragorang, and Grahamstown) had significantly larger 1-week forecast errors (292%, 203% and 62%, respectively). By contrast, the lakes with higher concentrations and prevalence of cyanobacteria, such as Vaal, Copeton and Theewaterskloof, had significantly smaller 1-week normalized forecast errors (20%, 25%, 13%). This suggests that forecasting of cyanobacteria is more successful in lakes where cyanobacterial blooms occur more often and at higher concentrations. Forecasting in lakes with few cyanobacteria detections and low Chla_cyano concentration is understandably associated with larger errors.

4. Discussion

The results demonstrate the feasibility of using simple univariate forecasting methods with spatially aggregated statistical variables derived from satellite remote sensing to forecast near-term chlorophyll-*a* and cyanobacteria concentrations across a variety of lakes of different size and type, with various lengths of historical data. Here, the accuracy of the forecasts, the factors affecting forecast performance, variations on the MASEA model, and how the approach may be implemented in practice is discussed.

4.1 Forecast Accuracy

How do these forecasting performances compare with alternative forecasting approaches? Based on a literature review of forecasting models, the primary measure of performance is the correlation coefficient, or R^2, which is unable to measure bias or provide a quantitative measure of model performance (Rousso et al. 2020). The use of R^2 as the primary measure of model performance seems to be a major oversight with publications in the domain as it precludes quantitative assessment of forecast errors in cells/mL or µg/L for cell counts or toxin concentrations. Nevertheless, based on the data presented in Rousso et al., the R^2 values for a 7-day time horizon for various advanced algorithms reached a maximum value of ± 0.875 with typical values near 0.6 (see Fig 6). The models therefore account for a maximum of 87%, but usually around 60%, of the variability in cyanobacteria cell counts or microcystin toxin concentrations. The only results from a 30-day (4-week) horizon had R^2 near 0.4. It is not immediately clear how the MASEA model compares to these models, however, the 65% mean forecast error for Chla_cyano appears comparable to the 60% of variability accounted for in these complex models.

The multivariate model presented by Myer et al. (2021) based on remotely sensed data had a 1-week forecast accuracy of 82% (n = 103 lakes) when forecasting whole-lake spatially aggregated high-risk cyanobacteria bloom probabilities (cells > 100 000 cells/mL). They found that there was a high degree of autocorrelation (90%), indicating that the largest factor in predictive success is the previous week's value (which is one of the primary underpinnings of the present study). They also expected their model to remain accurate for a time-horizon of no more than 2 weeks, but performance was not assessed for this or longer time horizons. By comparison, the current approach, which is vastly simpler than the one presented by Myer et al. (2021), the 1-week CRL forecast accuracy across all cyanobacteria risk level scenarios (low to very high) over a test dataset of one year (52 weeks) was 74% (n = 780). When assessing the 1-week forecast accuracy for high or very high-risk scenarios (Chl-a > 50 µg/L or 100 000 cells/mL), the overall accuracy was 80% (98 of 123 events detected across all lakes), with variability between lakes as described above (Table 5). The 2-week and 4-week accuracy dropped off to 71% and 69%, respectively. Using this metric, the MASEA 1-week forecast performance for high-risk cyanobacterial blooms is comparable to that of Myer et al. (2021) (80% versus 82%, respectively).

4.2 Factors affecting forecast performance

Interestingly, longer forecast horizons (4-week) that arguably depend more on seasonality, are often only slightly less accurate than a 1-week forecast horizon, again with variability between lakes. There was only weak evidence to suggest that a spatially-aggregated statistical approach may be less suitable for forecasting larger waterbodies. This lends some confidence that the approach may be valid for both large and small waterbodies. There was stronger evidence that longer time-series used to characterize the underlying seasonal component led to increased forecast accuracy. Those lakes for which there were more than 900 historical images (around 4 years of data) had the lowest normalized RMSE. By contrast, those lakes less than 600 images, had the worst or near-worst forecast performance.

4.3 Variations and improvements on the MASEA model

The MASEA model is a simple moving average model that adjusts for the seasonal average and the observed anomaly from seasonal norms. Alternatives to the selected model were tested, including a simpler model that included no error adjustment, and a variation using the seasonal probability of a cyanobacteria bloom. However, not all combinations of model possibilities were tested, and it is likely that it could be improved upon by adjusting the model's configuration, or by selecting more advanced approaches that are also computationally simple. As the use of more advanced models is excluded in this study, it would be interesting to assess the comparative performance of machine learning models in future. Further improvement could be made by incorporating an estimate of the turnover date, which can be forecast with some certainty using relatively simple models using the average hypolimnion temperature, mean depth and latitude adjusted for altitude effects (e.g., Nürmburg 1988). Studies have simulated depth-resolved water temperature using mean lake depth and surface area in a 0.5° grid (e.g., Woolway et al. 2021). Such an approach could be implemented and incorporated to increase forecast robustness, although the effects of the turnover will necessarily be captured in long timeseries.

4.5 Operational considerations for forecasts

Practically applying a model for forecasting blooms without prior knowledge of the system in question is feasible, however there are some operational limitations. Firstly, a forecast for which there is no recent data due to prolonged cloud cover would effectively revert to a seasonal naïve forecast resulting in considerably less precision. Further, forecasting for waterbodies where there is less or no historical data would revert to a moving average model, also with significantly worse predictive ability, especially for 2 and 4-week horizons. It would appear from the analysis above, that 900 images (or 5 years of data) provided more adequate seasonal characterization. This implies that long timeseries need to be analyzed (now easily performed using cloud-based computation) for a lake before forecasts become more accurate. One additional challenge is that many lakes are frozen for up to six months or more in a calendar year. Whilst this is not a problem *per se* given that seasonal characterization is on a week-by-week basis, some account has to be taken not to provide forecasts while lakes remain frozen, as the freeze date may vary significantly from year to year. Lastly, an important limitation is that the forecast is only as accurate as the remotely sensed data, which may be prone to errors and anomalies.

4.6 Outlook

The present study has demonstrated the significant value that remotely sensed data can add to near-term forecasting for lake management of cyanobacteria and HABs. The simple model forecast high-risk cyanobacterial blooms 1-week ahead with comparable accuracy with the complex model used by Myer et al. (2021). Forecasting performance showed that the seasonal component should be considered when forecasting and that longer historical time-series result in improved forecast accuracy.

The optimal forecasting models of the future will necessarily be comprised of diverse data from autonomous sensors (including hydrodynamic variables) *in situ*, metrological data, long-term sampling records and remotely sensed time-series. This study has demonstrated how simple univariate forecasting procedures can play a role alongside these more complex models in operational weather-like applications such as the CyanoLakes mobile application.

Acknowledgements. This work was supported by the Water Research Commission under Grant C2019_2020-00198. I would like to acknowledge the Project Manager Dr. Eunice Ubomba-Jaswa and the reference group members, in particular Dr. Annelie Swanepoel, for their insights and reviews. The code and data for this study can be found in the CyanoLakes public repository on Github: https://github.com/CyanoLakes/forecasting. I am thankful for the valuable contributions of two reviewers and the editors.

Declaration of Interest. This research is sponsored by CyanoLakes (Pty) Ltd and may lead to the development of products which may be licensed to CyanoLakes (Pty) Ltd, in which I have a business and/or financial interest. I have disclosed those interests fully to Taylor & Francis and have in place an approved plan for managing any potential conflicts arising from this arrangement.

References

Beal MR, O'Reilly B, Hietpas KR, Block P. 2021. Development of a sub-seasonal cyanobacteria prediction model by leveraging local and global scale predictors. Harmful Algae. 108:102100.

Caillault ÉP, Bigand A. Comparative study on univariate forecasting methods for meteorological time series. In: 26th European Signal Processing Conference (EUSIPCO); Sep 3-7; Rome, Italy: IEEE. p. 2380-2384.

Carey CC, Woelmer WM, Lofton ME, Figueiredo RJ, Bookout BJ, Corrigan RS, Daneshmand V, Hounshell AG, Howard DW, Lewis AS, McClure RP. 2022. Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. Inland Waters. 12(1):107-20.

Echard JS. 2021. A Review of Harmful Algal Bloom Prediction Models for Lakes and Reservoirs [master’s thesis]. Utah: Utah State University.

García-Mozo H, Chuine I, Aira MJ, Belmonte J, Bermejo D, de la Guardia CD, Elvira B, Gutiérrez M, Rodríguez-Rajo J, Ruiz L, Trigo MM. 2008. Regional phenological models for forecasting the start and peak of the Quercus pollen season in Spain. agricultural and forest meteorology. 148(3):372-80.

Hyndman RJ, Athanasopoulos G. 2018. Forecasting: principles and practice, 2nd edition. Melbourne, Australia: OTexts.

Ji L, Peters AJ. Forecasting vegetation greenness with satellite and climate data. 2004. IEEE Geoscience and Remote Sensing Letters. 1(1):3-6.

Kravitz J, Matthews M, Bernard S, Griffith D. 2020. Application of Sentinel 3 OLCI for chl-a retrieval over small inland water targets: Successes and challenges. Remote Sensing of Environment. 237:111562.

Lacey H, Mundava C, Hamilton L. 2021. Watch this Space: Algal Monitoring using Satellite Remote Sensing. Proceedings of the New South Wales Australian Water Association Conference; Nov; Tamworth; AWA.

Lawrence C, Morrow A, Sneddon A, Stanmore J, Platell M, Evans C, Lundmark A. 2020. The View from Up There – Monitoring A Severe Algal Bloom from Space. Proceedings of Australian Water Association OzWater'20; May 5–7; Adelaide; AWA.

Li W, Qin B, Zhu G. 2014. Forecasting short-term cyanobacterial blooms in Lake Taihu, China, using a coupled hydrodynamic–algal biomass model. Ecohydrology. 7(2):794-802.

Li H, Qin C, He W, Sun F, Du P. 2021. Improved predictive performance of cyanobacterial blooms using a hybrid statistical and deep-learning method. Environmental Research Letters. 16(12):124045.

Liu M, He J, Huang Y, Tang T, Hu J, Xiao X. 2022. Algal bloom forecasting with time-frequency analysis: a hybrid deep learning approach. Water Research. 14:118591.

Myer MH, Urquhart E, Schaeffer BA, Johnston JM. 2020. Spatio-temporal modeling for forecasting high-risk freshwater cyanobacterial harmful algal blooms in Florida. Frontiers in environmental science. 8:581091.

Nürnberg GK. 1988. A simple model for predicting the date of fall turnover in thermally stratified lakes. Limnology and Oceanography. 33(5):1190-5.

Peng WY, Chu CW. 2009. A comparison of univariate methods for forecasting container throughput volumes. Mathematical and computer modelling. 50(7-8):1045-57.

Qin B, Li W, Zhu G, Zhang Y, Wu T, Gao G. 2015. Cyanobacterial bloom management through integrated monitoring and forecasting in large shallow eutrophic Lake Taihu (China). Journal of hazardous materials. 287:356-63.

Recknagel F, Orr P, Swanepoel A, Joehnk K, Anstee J. 2018. Operational Forecasting in Ecology by Inferential Models and Remote Sensing. In: Recknagel F, Michener W, editors. Ecological Informatics. Cham: Springer; p. 319-339.

Recknagel F, Orr PT, Bartkow M, Swanepoel A, Cao H. 2017. Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modelling. Harmful algae. 69:18-27.

Rousso BZ, Bertone E, Stewart R, Hamilton DP. 2020. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. Water Research. 182:115959.

Shan K, Ouyang T, Wang X, Yang H, Zhou B, Wu Z, Shang M. 2022. Temporal prediction of algal parameters in Three Gorges Reservoir based on highly time-resolved monitoring and long short-term memory network. Journal of Hydrology. 605:127304.

Shah V, Turner D, Hancock C, O'Donoghue P, Sneddon A, Stanmore J, Morrow A, Lundmark A, Bates L, Hanson A. 2021. Monitoring cyanobacteria in Grahamstown Dam. Australian Water Association Water e-Journal, 5(4).

Svetunkov A. 2021. Forecasting and analytics with ADAM. Openforecast.org: Bookdown.

Swanepoel A. 2015. Early warning system for the prediction of algal-related impacts on drinking water purification [dissertation]. Potchefstroom: North-West University.

Swanepoel A, Barnard S, Recknagel F, Cao H. 2016. Evaluation of models generated via hybrid evolutionary algorithms for the prediction of Microcystis concentrations in the Vaal Dam, South Africa. Water SA. 42(2):243-52.

Tromas N, Fortin N, Bedrani L, Terrat Y, Cardoso P, Bird D, Greer CW, Shapiro BJ. 2017. Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course. The ISME journal. 11(8):1746-63.

Verpoorter C, Kutser T, Seekell DA, Tranvik LJ. 2014. A global inventory of lakes based on high-resolution satellite imagery. Geophysical Research Letters. 41(18):6396-402.

White MA, Nemani RR. 2006. Real-time monitoring and short-term forecasting of land surface phenology. Remote Sensing of Environment. 104(1):43-9.

Woolway RI, Denfeld B, Tan Z, Jansen J, Weyhenmeyer GA, La Fuente S. 2021. Winter inverse lake stratification under historic and future climate change. Limnology and Oceanography Letters.

World Health Organization. 2003. Guidelines for safe recreational water environments: Coastal and fresh waters (Vol. 1). Geneva: World Health Organization.

Xiao X, He J, Huang H, Miller TR, Christakos G, Reichwaldt ES, Ghadouani A, Lin S, Xu X, Shi J. 2017. A novel single-parameter approach for forecasting algal blooms. Water research. 108:222-31.