Adaptive forecasting of phytoplankton communities

Trevor Page a, *, Paul J. Smith a, c, Keith J. Beven a, Ian D. Jones b, J. Alex Elliott b, Stephen C. Maberly b, Eleanor B. Mackay b, Mitzi De Ville b, Heidrun Feuchtmayr b

a Lancaster Environment Centre, Library Avenue, Lancaster University, Lancaster, LA1 4YQ, UK
b Lake Ecosystems Group, Centre for Ecology & Hydrology, Lancaster Environment Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP, UK
c ECMWF, Shinfield Park, Reading, RG2 9AX, UK

Article info
Article history:
Received 8 August 2017
Received in revised form
11 December 2017
Accepted 20 January 2018
Available online 5 February 2018

Keywords:
Phytoplankton model
Forecasting
Data assimilation
Ensemble Kalman Filter
Cyanobacteria
PROTECH

A B S T R A C T

The global proliferation of harmful algal blooms poses an increasing threat to water resources, recreation and ecosystems. Predicting the occurrence of these blooms is therefore needed to assist water managers in making management decisions to mitigate their impact. Evaluation of the potential for forecasting of algal blooms using the phytoplankton community model PROTECH was undertaken in pseudo-real-time. This was achieved within a data assimilation scheme using the Ensemble Kalman Filter to allow uncertainties and model nonlinearities to be propagated to forecast outputs. Tests were made on two mesotrophic lakes in the English Lake District, which differ in depth and nutrient regime. Some forecasting success was shown for chlorophyll a, but not all forecasts were able to perform better than a persistence forecast. There was a general reduction in forecast skill with increasing forecasting period but forecasts for up to four or five days showed noticeably greater promise than those for longer periods. Associated forecasts of phytoplankton community structure were broadly consistent with observations but their translation to cyanobacteria forecasts was challenging owing to the interchangeability of simulated functional species.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Algal blooms are a global problem affecting water resources, recreation and ecosystems (Carmichael, 1992; Smith, 2003; World Health Organization, 1999). These problems are particularly acute when cyanobacterial species dominate because of the risk of toxin production that can cause adverse effects to humans and wildlife (Metcalf and Codd, 2009). In addition, water supply companies face associated problems such as poor taste and odour and, in extreme cases, high concentrations of algal-derived toxins which are costly to manage (Pretty et al., 2003; Dodds et al., 2009; Michalak, 2016). Costs associated with implementation of management strategies are growing because of increased bloom frequency (Ho and Michalak, 2015) and because of the effects of widespread nutrient enrichment and climate change (Paerl and Huisman, 2008; Brookes and Carey, 2011; Rigosi et al., 2014). As a result, there is an urgent need for reliable predictions of algal bloom formation to enable timely management interventions to be implemented.

Forecasting algal blooms in lakes is relatively new (Kim et al., 2014) but is increasingly becoming a requirement for lake and reservoir managers (Huang et al., 2013; Recknagel et al., 2014; Xiao et al., 2017) to help inform decisions regarding timely and cost-effective management interventions. The fact that limnology is rapidly becoming data-rich (Marcé et al., 2016; Xiao et al., 2014) means that effective real-time forecasts are increasingly more feasible. However, forecast simulations will be inherently uncertain for a number of reasons including input data resolution and simplifications in model process representation. These uncertainties have implications for the accuracy and reliability of a forecast and therefore effort is required to allow for modelling uncertainty. Data assimilation (DA) is one approach to reducing forecast uncertainty but has, to date, received relatively little attention for forecasting phytoplankton community dynamics. There is hence a need to test different DA methodologies across different lake systems and different models.

There are still relatively few studies for operational lake forecasting systems and various approaches have been taken such as using: Ensemble Kalman Filter (EnKF; Evensen, 1994) schemes and physically-based simulation models (e.g. Allen et al., 2003; Huang et al., 2013; Kim et al., 2014); evolutionary computation
Demonstrating the efficacy of the approach therefore requires a robust appraisal procedure with predictions tested qualitatively and quantitatively against appropriate benchmarks. This approach allows other pertinent questions to be investigated; namely, how does forecasting reliability diminish with time-scale of forecast and, most pertinently, what can be learnt from any forecasting failure regarding future model development and optimisation of monitoring strategies.

2. Methods

2.1. Study lakes

This study considers two lakes in the English Lake District of North West England with differing depths and nutrient regimes (Table 1). The catchments associated with each of the lakes are predominantly hill land, rough-grazed by sheep throughout the year and contain towns and villages that are tourist destinations and are hence associated with seasonal increases in lake nutrient inputs. Windermere is England’s largest natural lake and comprises two basins connected at a shallow region approximately halfway along its main axis. The two basins are usually considered separately as they have different characteristics: both basins are monomictic and mesotrophic, but only the south basin was modelled in this study. Esthwaite Water is a small, generally monomictic and occasionally dimictic, lake that has been subject to eutrophication for many decades because of elevated phosphorus levels (Bennion et al., 2000; Dong et al., 2012); cyanobacterial blooms are common in the summer to early autumn. Previous work has shown that internal sources from the lake sediment form an important component of the P budget of the lake (Hall et al., 2000; Heaney et al., 1992; Mackay et al., 2014).

2.2. Data

2.2.1. Forcing inputs: meteorological forecasts

The primary forcing inputs were meteorological forecasts provided by the European Centre for Medium-term Weather Forecasts (ECMWF) Ensemble Prediction System. The 10-day-ahead forecasts include an ensemble of 50 simulations from perturbed initial states (at 32 km² resolution) and stochastic perturbations of model parameters (see Buizza et al., 1999; Ollinaho et al., 2017). The re-initialisation of model states in the ECMWF forecasting system is implemented using a higher resolution 3-h forecast each day. As this re-initialisation is repeated each day, and as perturbations are random, there is no specific relationship between individual ensemble members in subsequent days. The forecast associated with each ensemble member was hence treated as independent from prior forecasts for this study. Daily averages of forecasts were used (i.e. the average of 3-hourly forecasts for days 1–6 and of 6-hourly forecasts day 6–10) for consistency with the daily timestep of PROTECH. Historic forecasts were obtained for 2008, 2009 and 2010 and used in pseudo-real-time. Given the scale of the forecast grid, each forecast variable was “downscaled” to local data as described in the next section.

2.2.2. Sampling meteorological forecasts

Downscaling relationships were developed for air temperature, wind speed, precipitation, cloud cover, relative humidity and solar radiation (Table 2). For air temperature, a relationship was identified between forecasted temperatures and observed temperatures using linear regression. Residuals from this initial analysis helped identify an additional hysteretic relationship between forecasted and observed temperatures, which was attributed to a lake thermal effect; this effect was implemented as an additional correction for each day of the year. Similarly, wind speed was corrected using a linear correction factor coupled with an additional correction based upon wind direction; this was required owing to complex mountainous topography and lake-axis orientation. A wind-rose with sectors of 30° was used to classify forecasted wind speeds and a sector-specific correction was applied. The uncertainty associated with the corrections was represented by fitting a gamma distribution to the data in each sector. All other variables (precipitation, cloud cover, relative humidity and solar radiation), were corrected using a correction multiplier identified using linear regression, without propagating the uncertainty in the relationship. The

<p>| Table 1 |
| Study Lakes and primary characteristics. |</p>
<table>
<thead>
<tr>
<th>Name/location</th>
<th>Mean Depth (m)</th>
<th>Max. Depth (m)</th>
<th>Max. Length (m)</th>
<th>Volume (m³)</th>
<th>Catchment Area (km²)</th>
<th>Residence Time (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windermere (South Basin)</td>
<td>16.8</td>
<td>41</td>
<td>9300</td>
<td>1.06 x 10⁸</td>
<td>230.5</td>
<td>100</td>
</tr>
<tr>
<td>Esthwaite Water</td>
<td>6.4</td>
<td>15.5</td>
<td>2500</td>
<td>5.97 x 10⁶</td>
<td>17.1</td>
<td>100</td>
</tr>
</tbody>
</table>

* Details from Ramsbottom (1976).
uncertain relationships for air temperature and wind speed were resampled as perturbations of the ensemble members allowing investigation of the effect of different ensemble sizes.

### 2.2.3. Nutrient inputs

Knowledge of diffuse nutrient inputs for the study lakes is relatively poor. Observations available were from approximately monthly frequency routine monitoring and did not cover all river inputs. Both lakes are also impacted by point sources from waste water treatment works (WwTW) and Esthwaite is subject to significant internal P fluxes (Mackay et al., 2014). Diffuse nutrient inputs and WwTW inputs (where included) were treated as reported by Page et al. (2017) and these inputs were modified by a multiplicative parameter included in the EnKF scheme (Table 4). For Windermere, upstream lake inputs of nutrients (and chlorophyll a) were treated as reported by Page et al. (2017) but were not included in the EnKF scheme.

### 2.2.4. Data for assimilation and evaluation of forecasts

Specific years where the observed data were of the highest frequency, were chosen to test the DA strategy. High frequency (4 min) data from the automatic lake monitoring systems (Madgwick et al., 2006; Mackay et al., 2014) were available and were aggregated to daily values. The variables used for DA are listed in Table 3. The “observed” temperatures for the epilimnion ($T_e$) and hypolimnion ($T_h$) used to compare with the modelled variables for these layers were calculated as volume-weighted averages of thermistor chain data, using the simulated epilimnetic depth to delineate the hypolimnion and epilimnion. The “observed” epilimnetic depth ($D_e$) was estimated using a density gradient method (e.g. see Read et al., 2011). In addition to the automatic monitoring, routine monitoring was carried out at the buoy location at a frequency of approximately every 14 days and included chlorophyll a, phytoplankton species “counts”, soluble reactive phosphorus (SRP), dissolved inorganic nitrogen (DIN) and silica (SiO$_2$) (Table 3). These observations were derived from a water sample at the buoy location integrated over 0–7 m depth (Windermere) or 0–5 m depth (Esthwaite Water) (Maberly et al., 2011).

#### Table 2

Forcing inputs and downscaling relationships.

<table>
<thead>
<tr>
<th>Model Inputs</th>
<th>Downscaling factor/relationship</th>
<th>Uncertainty sampled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air Temp ($T_a; K$)</td>
<td>Windermere: 0.095($T_a$) + 279.75$^a$</td>
<td>Y (Regression)</td>
</tr>
<tr>
<td>Solar Radiation (SR; Wm$^{-2}$)</td>
<td>Esthwaite Water: 0.013($T_a$) + 280.16$^a$</td>
<td>N</td>
</tr>
<tr>
<td>Wind Speed (W; m s$^{-1}$)</td>
<td>0.85</td>
<td>N</td>
</tr>
<tr>
<td>Relative Humidity (RH; %)</td>
<td>0.38$^b$</td>
<td>Y (Gamma Dist.)</td>
</tr>
<tr>
<td>Cloud Cover (Cc; eighths)</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>Rainfall (R; mm)</td>
<td>1.25</td>
<td>N</td>
</tr>
<tr>
<td>Nutrient Inputs (P; N; SiO$_2$/mg m$^{-3}$)</td>
<td>3</td>
<td>N</td>
</tr>
</tbody>
</table>

Ta$^a$ is the forecast air temperature (K).

a See Section 2.2.2 for additional wind direction correction.

b See Section 2.2.2 for additional lake-effect correction.

### 2.3. Modelling methodology

The modelling strategy employed was designed to represent the different facets of the forecasting system as simply as possible to reduce computational burden, whilst retaining the requirement to explicitly simulate phytoplankton community structure and, specifically, to estimate the likely concentrations of cyanobacteria given the simulated community structure. Thus, the catchment-lake system was simulated using a suite of models of differing complexity from purely data-based (statistically estimated) transfer function (TF) models and processed-based models which are consistent, in their complexity, with the available data. A schematic of how the models were combined in the forecasting system is presented in Fig. 1 and each model is described in this section. The modelling system is structured around the rationale that epilimnetic depth must be estimated as accurately as possible so that the phytoplankton model, PROTECH, is more likely to provide good estimates of phytoplankton community structure. In PROTECH, community structure is simulated using functional algal types and as classified by Reynolds (1988) and as outlined in the next section. The simple conceptual model that estimates epilimnetic depth is a heat energy “balance” model that requires estimates of epilimnetic temperature and energy fluxes to the epilimnion, including those associated with river inflows and outflows.

The TF models, epilimnetic depth model and PROTECH are run sequentially; the TF and epilimnetic depth models provide forecast estimates of river flow, epilimnetic depth, epilimnetic temperature and hypolimnetic temperature as inputs to PROTECH. Data assimilation is employed for the two primary models (the epilimnetic depth model and PROTECH) using two separate EnKF schemes that assimilate observations at different intervals; the epilimnetic depth model scheme assimilates epilimnetic depth and epilimnetic temperature estimates as well as hypolimnetic temperature estimates on a daily basis and the scheme for PROTECH assimilates nutrient and chlorophyll a concentrations approximately every 14 days.

#### Table 3

Observed data assimilated in the EnKF scheme.

<table>
<thead>
<tr>
<th>Assimilated state</th>
<th>Frequency</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epilimnetic Temperature (°C)</td>
<td>Daily</td>
<td>buoy obs.</td>
</tr>
<tr>
<td>Hypolimnetic Temperature (°C)</td>
<td>Daily</td>
<td>buoy obs.</td>
</tr>
<tr>
<td>Epilimnetic depth (m)</td>
<td>Daily</td>
<td>buoy obs.</td>
</tr>
<tr>
<td>Chlorophyll a (mg m$^{-3}$)</td>
<td>= 14 days</td>
<td>Monitoring</td>
</tr>
<tr>
<td>Nutrient Inputs (SRP; N; SiO$_2$/mg m$^{-3}$)</td>
<td>= 14 days</td>
<td>Monitoring</td>
</tr>
</tbody>
</table>

2.3.1. The PROTECH model

PROTECH (Reynolds et al., 2001) is a lake phytoplankton community model that runs on a daily time-step. It is a 1-dimensional model where the lake is represented by horizontal layers. In the model representation all layers are assumed to be fully mixed throughout the epilimnion. River inputs drive fluxes of diffuse nutrients as well as the flushing of phytoplankton. Upstream lake inputs are treated as river inputs but are given the phytoplankton concentrations associated with the upstream lake, where data are available.

Underwater light for model layer $i$ is calculated using:
the following equation which describes the change in chlorophyll a concentration ($X$) of each phytoplankton species selected to represent the algal community (Reynolds et al., 2001):

$$\frac{\Delta X}{\Delta t} = \left( r' - S - G - F \right) X$$

(2)

where $r'$ is the growth rate, $S$ is settling loss, $G$ is a grazing loss and $F$ is the loss due to flushing. The growth rate is defined for each layer using:

$$r' = \min\{r'_g(\theta), r'_p(\theta), r'_n(\theta), r'_{Si}(\theta)\}$$

(3)

where $r'_g(\theta), r'_p(\theta), r'_n(\theta)$ and $r'_{Si}(\theta)$ are the growth rates determined by temperature, photoperiod ($I$) and daily light extinction ($\epsilon$) and $r'_n, r'_p, r'_Si$ are the growth rates determined by phosphorus, nitrogen and silica concentrations. The final growth rate ($r'_{\text{corr}}(\theta)$) is a corrected rate allowing for dark respiration using equation (4). This is required as the model growth equations are not of basal metabolism but not dark respiration burden.

$$r'_{\text{corr}}(\theta) = R_{d}(\theta) \cdot r'_g(\theta) - \left( 1 - R_{d}(\theta) \right) r'_g(\theta)$$

(4)

Where $R_d(\theta)$ is the dark respiration rate at temperature $\theta$.

PROTECH simulates the dynamics of the species chosen to represent the phytoplankton community of a given lake. Species are represented by their morphology, nutrient requirements (i.e., silica requirement and nitrogen fixing ability) and their vertical movement strategies. The number of species simulated is nominally eight (although unlimited) and they are chosen to represent the dominant functional types of the system. Simulations hence represent the behaviour of the functional algal community rather than the dynamics of specific species. The C-S-R functional phytoplankton classification of Reynolds (1988) is used to classify phytoplankton into morphologically defined groups relating to broad ecological strategies. The primary groups are: C-types, which are invasive, ecological pioneers that are small with high silica requirement and nitrogen fixing ability; S-types which are ‘stress tolerators’ that tolerate relatively low nutrient availability and strong stratification (i.e., Woronichinia, Microcystis and Oocystis); and R-types which can harvest sufficient light at low levels to be able to maintain growth and are hence tolerant of well-mixed, intermittently insolated environments (e.g. Asterionella, Aulacoseira and Plankothrix). Also present, but less important for the lake-years studied here, are CS-types, whose characteristics are intermediate between those of C and S species (e.g. Dolichospermum, Aphaniemenon and Ceratium) and CSR-types (e.g. Cryptomonas) that are intermediate between C- and R-types. The eight phytoplankton used in each lake for this
study are presented in Table Supp. 2.

2.3.2. Epilimnetic depth model

As a way of reducing computational burden, a simplified representation of lake thermal structure was employed to estimate epilimnetic depth (De). The simplified model works on the basis of independent estimates of epilimnetic temperature and lake heat energy fluxes. The estimate of epilimnetic temperature (Te) uses a TF model (see Section 2.3.3) with inputs of air temperature (Ta), solar radiation, wind speed (Ws) and De. Air temperature, solar radiation and wind speed are derived from the forecasts and De estimates are from the previous simulation timestep. The independent estimates of heat energy fluxes are calculated using the PROTECH energy flux function (see Reynolds et al., 2001) for each timestep using Te, river temperature and flow magnitude, day length, cloud cover, Th Relative Humidity and Ws.

These two independent estimates are “balanced” to obtain hypolimnetic volume (Vh) using:

\[ V_h = \frac{E_{i,T}}{\Delta T C_w \rho_w} \]  

(5)

where, \( E_{i,T} \) is the heat energy associated with \( \Delta T \) (the difference between \( T_e \) and the hypolimnetic temperature, \( T_h \)); \( C_w \) is the specific heat capacity of water, \( \rho_w \) is the density of water. Equation (5) is solved to find \( V_h \) where: \( \Delta T \), \( C_w \), \( \rho_w \). \( V_h = E_{i,T} \). Subsequently, the epilimnetic volume (Ve) and hence epilimnetic depth (De) are estimated by difference:

\[ V_e = V_t - V_h \]  

(6)

where \( V_t \) is the total lake volume. The requirement for \( \Delta T \) is satisfied by calculating \( T_h \) using:

\[ T_h = \frac{E_{i,h}}{C_w \rho_w V_t} \]  

(7)

where: \( E_{i,h} \) is the “background” heat energy in the lake (associated with \( T_h \) and \( V_t \) as defined by Eqn. (7)). During the forecast period, \( E_{i,h} \) remains at its previous value until updated during the data assimilation step. This treatment of \( E_{i,h} \) neglects the explicit downward transfer of energy from \( E_{i,T} \) to \( E_{i,h} \) for forecasting and assumes that these are negligible over this timescale: energy is, however, explicitly transferred downwards each timestep when temperatures are updated during data assimilation. The sequence of calculations for each forecast timestep is:

1. Estimate lake surface temperature using TF model
2. Update \( E_{i,T} \)
   I. Radiative energy fluxes
   II. River/ upstream lake fluxes
      • Estimate river input volume using TF model
      • Estimate river temperature using TF model
      • Assume upstream lake temperature = modelled lake temperature
   III. If \( E_{i,T} < 0 \) loose energy from \( E_{i,h} \) (minimum energy set to 0 °C)
3. Estimate \( T_h \) from \( E_{i,h} \)
4. If \( E_{i,T} > 0 \) and if \( T_e - T_h \) is greater than a threshold parameter (nominal set to 1 °C) estimate epilimnetic depth by solving for the volume of water required to match \( E_{i,T} \) given \( \Delta T \): subsequently estimate \( V_e \) and hence \( D_e \) by difference.

2.3.3. Transfer function models

Transfer Function (TF) models were used to estimate lake surface temperature, river temperature and river inflows and outflows. Each model is a discrete-time TF identified directly from the available data. Both the model structures and parameters were identified using the Refined Instrumental Variable (RIV) algorithm (Young, 2015) implemented within the CAPTAIN Toolbox for Matlab (Taylor et al., 2007). The resulting model structures and parameter values are presented in Section (Supp. 1) and are either single input- or multi-input, single-output first order models of the general form:

\[ y_t = \frac{B_1(z-1)U_1 + B_2(z-1)U_2 + \ldots + B_n(z-1)U_n}{A(z-1)} \]  

(8)

where, \( y_t \) is the variable being estimated at time \( t \), \( U_1, n \) are model input vectors, \( A(z-1) \) and \( B_0(z-1) \) are the model coefficients (polynomials in the backward shift operator; defined by \( y_t z^{-1} = y_{t-1} \)) that number 1 to \( n \) in the case of \( B \) but note that in this form for MISO (multi-input single-output) TF the denominator \( A \) is common to all \( n \) TF elements.

2.3.4. The ensemble Kalman filter

The EnKF is a sequential Monte Carlo method which uses a stochastic ensemble of model simulations, and stochastic forcing, to propagate estimates of model states and (or) parameter values between assimilation timesteps. As the ensemble of model simulations is used in place of the linear propagation of an error covariance matrix (as in the traditional Kalman Filter), non-linear model dynamics are retained during model evolution and uncertainties are represented by the variation of the ensemble. When observations are available, each ensemble member is updated individually using a linear update equation (Eqn. (9)) which relies on the assumption that the relationship between states and parameters can be described by multivariate Gaussian distributions. Rather than resampling the posterior distributions of the updated ensemble, the EnKF uses each updated ensemble member such that some of the non-Gaussian properties of the forecast are retained (Evensen, 2009). The procedure for the scheme is as follows:

1. The EnKF is initialised with an \( N \) number ensemble size, sampling states and parameters from a priori specified distributions (see below for specific details of this study) and \( N \) simulations for the forecast period are carried out. Where parameters are varied as part of the EnKF scheme, they are appended to the state matrix to give a state-parameter matrix.
2. When observed data are available for assimilation:
   I. Apply a linear covariance inflation factor (I) to each of the \( i \) states and parameters to reduce the tendency for low ensemble covariance and for spurious correlations associated with small ensemble size (Anderson, 2007; Anderson and Anderson, 1999; Evensen, 2009):

\[ \phi_i^0 = I \left( \phi_i^0 - \phi_i^\tau \right) + \phi_i^\tau \]  

(9)

   II. Generate \( N \) perturbations of the observations (\( Y_t \)); it is essential that the uncertainty associated with the observations is sampled from a distribution with mean equal to the observed value and covariance (\( P^\tau \)) to avoid bias in the update (Evensen, 2009) and to reduce further the tendency for the updated ensemble to have very low covariance (Moradkhani et al., 2005).
   III. Update the model states and parameters individually for the \( j^\text{th} \) ensemble member. This is done proportionally to the deviation of the states in the forecasted state-parameter
matrix \( \phi^0 \) from the vector of perturbed observation \( s \) and the Kalman gain matrix \( K \): note that the timestep suffix is omitted for clarity in the following equations:

\[
\phi^0 = \phi^i + \left( K(Y) - H\phi^i \right)
\]  

(10)

where, \( \phi^i \) is the vector of updated states/parameters and \( H \) is a matrix that maps the model states to the observed states. The appended parameters are updated using the cross-covariance between the predicted states and parameters. The Kalman gain matrix is calculated using:

\[
K = P^e_k H^T \left( H \left( P^e_k \right) H^T + P^o \right)^{-1}
\]  

(11)

where, \( P^e_k \) is the covariance matrix for the ensemble of forecasted state-parameter matrix.

IV. Apply any constraints on states and (or) parameter distributions (e.g. to keep them within physically reasonable ranges). This was implemented using a resampling scheme where if any state/parameter violated specified constraints (Table 4), the ensemble was resampled using a truncated distribution for that state/parameter in conjunction with a Gaussian copula to retain the ensemble’s covariance structure.

V. Make \( N \) number of simulations for the next forecast period using the updated state-parameter matrix.

2.3.5. Ensemble Kalman Filter scheme: epilimnetic model

As the epilimnetic model is very simple, all the main model states were used in the EnKF scheme. The states \( T_e, T_r \) and \( D_e \) were updated using a daily assimilation frequency for the epilimnetic depth model. The “observed” values of these states are those estimated and described above.

2.3.6. Ensemble Kalman Filter scheme: PROTECH

The choice of states and parameters included in the PROTECH EnKF scheme was made based on uncertainty and sensitivity analyses reported by Page et al. (2017). The Page et al., study, which included the lakes studied here, identified that the main challenges for forecasting were uncertainties associated with: representing phytoplankton exposure to light and nutrient inputs (particularly phosphorus). The DA scheme was therefore defined to include the main model states, SRP, DIN, SiO\(_2\) and chlorophyll \( a \), as well the parameters associated with modifying nutrient inputs and underwater light (Table 4). These were updated at an approximately 14-day frequency set by the monitoring data. For Windermere, both point source (WvTW\(_i\)) and diffuse SRP inputs (\( I_{fac} \)) parameters were included in the DA scheme; for Esthwaite Water only the parameter modifying the diffuse SRP inputs was included as simulations which included a simplified representation of sediment-derived SRP inputs did not provide improved results (these results are not reported here).

To investigate the effect of ensemble size and to determine an acceptable ensemble size for the current applications, ensemble member (EM) size was increased sequentially, using the scenarios EM50, EM100, EM200, EM300 and EM400 (where the suffix is the size of the ensemble), until the forecast simulations appeared consistent. These scenarios were generated by resampling the downscaled ECMWF forecast distributions as described above and were used to force the suite of models used. For each of the forecast scenarios, the error associated with the assimilated data and the variance inflation factors were “optimised” manually to provide the best results. For consistency, and in the spirit of the pseudo-real time treatment of the forecast simulations, the variance inflation factors were kept consistent across all lake-years considered. For each of the assimilated variables, the variance was assumed to be proportional to the magnitude of the variable of interest using a percentage. Additionally, a minimum variance was applied to reduce the impact of very small observed values (e.g. where epilimnetic SRP values are observed to be very low or within the limit of detection) where the associated low variance would falsely indicate low uncertainty.

2.3.7. Assessing forecast skill

Different studies have used different benchmarks to evaluate the goodness of fit of forecasts (forecast skill), which are often determined by their aims. Studies tend to use either some form of “reference” simulation or simulations that do not assimilate any observations (sometimes called “climatology”) which serve to quantify the DA effect (e.g. Allen et al., 2003; Kim et al., 2014) or solely a measure of the goodness-of-fit to observations (e.g. the coefficient of determination, \( R^2 \)). Here, as our aim was to assess the value of the model for operational forecasting, we used a more stringent persistence forecast (e.g. see Stumpf et al., 2009) which uses the most recent observation as the forecast for each forecast timestep until the next observation becomes available. In the sections below, forecast skill was assessed by comparing the simulated chlorophyll \( a \) forecast with a persistence forecast for the entire annual timeseries. The goodness of fit of the benchmark and the simulated chlorophyll \( a \) forecasts were determined using the root-mean-square error (RMSE) as a measure. For the epilimnetic depth model, and other sub-models (i.e. TF models), goodness of fit is discussed more generally by comparison with observations using the coefficient of determination (\( R^2 \)). Assessment of the forecasts of phytoplankton community structure and cyanobacteria is made qualitatively as we have much lower confidence in the absolute value of the observations. A discussion of how the phytoplankton species “count” data are used and the associated uncertainties is provided in the relevant section below.

3. Results and discussion

3.1. TF model results

Transfer function models were identified for epilimnetic temperature, river temperature and river inflows and outflows and all models provided good fits to the observed data during model identification: \( R^2 \) values were between 0.86 and 0.98 (Supp. Table 1). Model identification was carried out for the entire period of data available (see Supp. 1) such that they were not year specific models. As detailed above, in each case the models were used to forecast their respective variable deterministically.

3.2. Forecasting epilimnetic depth and the phytoplankton community

3.2.1. Epilimnetic depth forecasts

Epilimnetic depth forecast estimates were made for 2008–2010 for Windermere and 2008 and 2009 for Esthwaite Water within the parallel EnKf scheme. Although very simplistic, the epilimnetic depth model provided reasonable forecasts of epilimnetic depth when compared to those estimated from observations. For both lakes, the forecasts were stable and consistent using the smallest ensemble size of 50 using a variance inflation factor of 1.25. Simulations for Windermere were better than for Esthwaite Water (\( R^2 \) of 0.85 and 0.75 respectively for a 10-day-ahead forecast; Fig. 2a and b) and there were short periods with significant deviations.
from the ‘observed’ depths in both cases. Simulation of the timing of temporary stratification events at the beginning of the year was problematic for both lakes and simulations tended towards overly rapid mixing during autumn turnover, particularly for Esthwaite Water. Where significant deviations exist, they have the potential to reduce the forecast skill and therefore need to be improved, although, importantly, epilimnetic depth estimates for much of the high cyanobacterial bloom risk periods (i.e. during periods of strongest stratification) are reasonable. Given these results, the epilimnetic depth estimates for Windermere appear to be adequate out to 10-days-ahead but for Esthwaite they appear to be adequate for a much shorter lead time; for example, the 3-day-ahead forecast is a much better fit with an improved $R^2$ of 0.81 (Fig. 2c). The adequacy of these estimates is assessed more formally in association with the Chlorophyll $a$ forecasts in comparison to the persistence forecast in the next section.

3.2.2. Chlorophyll $a$ forecasts

For all lake-years, multiple runs of the EM50 forecasts gave inconsistent simulations and a higher EM size was required. Forecasts for Windermere tended towards stability between the EM100 and EM200 scenarios (Fig. 3), which is an ensemble size consistent with previous work with relatively complex models (e.g. Evensen, 1994; Allen et al., 2003). For Esthwaite Water, however, a higher ensemble size appeared to be required with a size of around 400 giving consistent simulations (Fig. 4). Subsequently, in the following, results presented for Windermere and Esthwaite Water are associated with the EM200 and EM400 scenarios respectively. In all cases, the manually “optimised” variance inflation factor was kept consistent for all lake years at a value of 1.1.

Although forecast simulations for Windermere appear to be relatively good visually (e.g. see Fig. 5), they were not always an improvement on the persistence forecasts (Fig. 3). For 2008, the persistence forecast was better than simulated forecasts for all lead times. Conversely, simulated forecasts were better than the persistence forecasts for all lead times for 2009. A lead time of approximately 6 days or less was an improvement on the persistence forecast for 2010 simulations.

For Esthwaite Water, forecast simulations were not as good as those for Windermere (Fig. 5), which is consistent with previous work using PROTECH for these lakes (Page et al., 2017). The forecasts for 2008 were, however, still better than the persistence forecast out to about 5 days ahead (Fig. 4a), but were always worse than the persistence forecast for 2009 (Fig. 4b). The poorer fits for Esthwaite Water are likely to be a result of the complex uncertainties associated with the timing and magnitude of SRP inputs as well as the poorer simulation of epilimnetic depth reported above. In Esthwaite Water, during the period where P limitation dominates phytoplankton growth, it is very difficult to represent SRP fluxes appropriately, even when a representation of sediment-derived SRP fluxes was included (the addition of representation of sediment-derived SRP did not improve forecasts owing to interaction between sources of P: this work is not reported here). The difficulties associated with representing SRP fluxes was helped to some degree by the DA, but remain problematic during times when very low concentrations were present in the epilimnion; at these times, the correlations within the Kalman gain matrix would need to be very well-represented to provide appropriate updates to both epilimnetic SRP concentrations and SRP fluxes simultaneously. The difficulties associated with these updates are compounded by the relatively low frequency of assimilation timesteps. Subsequently, even with relatively large ensemble sizes, the correlations within the Kalman gain matrix have the potential to be spurious. This is not unexpected as the lake system is highly dynamic and non-linear and, perhaps most importantly, the relationships between the states (and parameters in some cases) are not always consistent (e.g. when the nutrient states are not limiting they may have no relationship with the phytoplankton state). The temporal evolution of the nutrient parameter values (modified within the DA scheme) that change SRP fluxes were consistent with these uncertainties and did not show any consistent structure. Given these difficulties, assimilation of higher resolution nutrient observations may be one of the most important ways of improving forecasts. Conversely, for both Windermere and Esthwaite Water, forecasts were improved by the modification of the background light extinction parameter, $ε_b$, within the DA scheme: its evolution over the simulation periods

![Fig. 2. Simulated and measured epilimnetic depth. Results shown for (a) Windermere 2008–2010 10-day-ahead, (b) Esthwaite Water 2008 and 2009 10-day-ahead and (c) Esthwaite Water 2008 and 2009 3-day-ahead: “observed” epilimnetic depth (red line), 50th percentile of the ensemble of simulated epilimnetic depth (black line) and 5th and 95th percentiles (grey lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)](image-url)
was relatively consistent for each of the years considered (Fig. 6) and reflects known simulation artefacts previously reported by Page et al. (2017).

3.2.3. Forecasting phytoplankton community structure

Forecasts of species representing the phytoplankton community structure were made without direct constraint within the DA scheme. Simulations were, however, indirectly constrained by the assimilation of epilimnetic depth, chlorophyll a and nutrients and hence are reliant on the ability of PROTECH simulations to represent phytoplankton community structure where abiotic conditions for phytoplankton growth are simulated adequately. They are also reliant on whether or not the phytoplankton species chosen to represent the community are appropriate (Elliott, 2010, 2012; Page et al., 2017).

Forecasts of community structure are assessed here using

![Fig. 3. Chlorophyll a forecast skill for the differing ensemble size scenarios. Results are shown for (a) Windermere 2008, (b) Windermere 2009 and (c) Windermere 2010, compared to the benchmark persistence forecast. Note that lower ensemble sizes can give “randomly” better forecast performance (e.g. EM = 50 in pane (a)).](image)

![Fig. 4. Chlorophyll a forecast skill for the differing ensemble size scenarios. Results are shown for (a) Esthwaite Water 2008 and (b) Esthwaite Water 2009, compared to the benchmark persistence forecast.](image)
simulations of R- and CS- functional types. These functional types were used as they dominate our study lakes. The observations to which they are compared here are estimated from “counts” of algal species, which are classified into the same functional groups. The “count” data were converted to biovolume using microscope measurements (Centre for Ecology & Hydrology, unpublished data) and subsequently to Chlorophyll a using the relationships in Reynolds (1984). This chain of approximations means that the observed data are associated with significant uncertainty. Accordingly, we used the relative abundance of each functional type for each observation timestep to partition the observed chlorophyll a concentration as our final estimate and estimated the sampling/analytical error to be +/- 25% and the overall error to be +/- 50% in accordance with Page et al. (2017).

A comparison of the uncertain observations of R- and CS-functional types are presented in Fig. 7 where it can be seen that for most lake-years the overall pattern of the simulations are consistent with the observations. There are some periods where the simulations are not consistent, which are associated primarily with the period of transition between the early blooms of R-type species and succession by CS-types (approximately between days 100 and 200). This inconsistency can clearly be seen for Windermere 2008 and 2009 (Fig. 7a and d) and is most likely associated with inadequate representation of nutrient fluxes and subsequent periods of nutrient limitation (Page et al., 2017). There are also some periods where the overly rapid mixing simulated by the epilimnetic depth model made it difficult to simulate the relatively high observed biomass: this is particularly evident for CS-species in Esthwaite Water 2008 (Fig. 7k) and R-species in Esthwaite Water 2009 (Fig. 7l); these inconsistencies are a direct result of the spurious deep mixing events simulated around days 220 and 250 for 2008 and 2009 respectively (see Fig. 2 b and c) and strengthen the requirement to improve the epilimnetic depth model.

3.2.4. Forecasting cyanobacteria

Observations of Cyanobacteria are estimated in the same way as functional species types discussed in the previous section and are associated with similar uncertainty (see Fig. 7). As PROTECH simulates the functional algal community using the dynamics of a number of selected individual species, the philosophy behind this method means that the forecasts of individual species are not as robust as those for functional community structure and are hence more uncertain. This is the case for forecasts of cyanobacteria where they are represented by more than one functional type: e.g.
Fig. 7. Concatenated five-day ahead forecasts of R-species, CS-species and cyanobacteria concentration for all lake years; black line is 50th percentile and grey shaded area represents the 5th and 95th percentiles of the ensemble: EM200 and EM400 for Windermere and Esthwaite respectively. The box and whisker symbols represent the analytical uncertainty and the total uncertainty estimated by the project team. Note that 5-day ahead forecasts are presented as approximately this lead time provided the most consistently acceptable results.
for Windermere cyanobacteria are represented by *Planktothrix*, an R-type species, together with *Aphanizomenon flos-aquae* and *Dolichospermum* which are CS-type species (see Table Supp. 2). In this situation, the interchangeability of species with similar functional behaviour, but which have differing species traits, requires additional interpretation for forecasts of cyanobacteria to be made. For example, the simulations of the R-species *Planktothrix* for all lake-years for Windermere result in overestimations of cyanobacteria concentrations for the periods where *Planktothrix* proliferates (approximately between days 150 and 275: Fig. 7c, f & i). Cyanobacteria forecasts, made for this study, are also a spatial average for each lake, constrained using data collected at one point; they therefore do not necessarily correspond with the risk from near-surface accumulations of cyanobacteria where significant spatial heterogeneity exists, as can be the case for wind-blown cyanobacterial species (e.g. George and Heaney, 1978). Extending point forecasts to spatial forecasts for species that have these characteristics is hence an additional challenge. However, forecasts may be presented as probabilistic or possibilistic risk estimates, such as the likelihood of a cyanobacterial concentration of greater than a given critical threshold: this will be the focus of further research.

4. Conclusions

We rigorously tested the ability of the phytoplankton community model PROTECH to make forecasts of phytoplankton community structure within a data assimilation scheme using the Ensemble Kalman Filter. Some forecasting success was shown for chlorophyll *a*, but not all forecasts were better than a persistence forecast. The results typically indicated a reduction in chlorophyll *a* forecast skill with length of forecasting period with forecasts for up to four or five days showing greater promise than those for longer time-scales. Associated forecasts of phytoplankton community composition, represented by functional algal types, were broadly consistent with observations. Translation of forecasts of functional algal types to forecasts of cyanobacteria are challenging because of functional similarities between species which may or may not be cyanobacteria. Improvements in forecasts are likely to come from higher frequency observations for both chlorophyll *a* and nutrient concentrations. Fluorescence-based field sensors for both chlorophyll and the cyanobacterial pigment phycocyanin exist and while they are not completely quantitative, they would permit patterns of change to be captured. While higher frequency observations for these variables should help improve forecasts, they will also simultaneously improve the persistence forecast. It, therefore, remains to be seen whether or not a modelled forecast driven with improved observations would provide a significant improvement over the associated persistence forecast and the potential to forecast algal blooms in this type of lake.

Acknowledgements

This work was supported by the Natural Environment Research Council projects: the United Kingdom Lake Ecological Observatory Network (UKLEON; grant number NE/I007407/1) and the Consortium on Risk in the Environment: Diagnostic, Integration, Benchmarking, Learning and Elicitation (CREDIBLE; grant number NE/J017295/1); We would like to thank the ECMWF for the historic meteorological forecasts and Mr Bernard Tebay for collecting the rainfall data at Ambleside, Cumbria.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.watres.2018.01.046.

References


