Article (refereed) - postprint

Page, Trevor; Smith, Paul J.; Beven, Keith J.; Jones, Ian D.; Elliott, J. Alex; Maberly, Stephen C.; Mackay, Eleanor B.; De Ville, Mitzi; Feuchtmayr, Heidrun. 2018. **Adaptive forecasting of phytoplankton communities**.

Adaptive forecasting of phytoplankton communities

*Trevor Page[1**], Paul J Smith[1,3], Keith J Beven[1], Ian D Jones[2], J Alex Elliott[2], Stephen C Maberly[2], Eleanor B Mackay[2], Mitzi De Ville[2] and Heidrun Feuchtmayr[2].*

[1] Lancaster Environment Centre, Library Avenue, Lancaster University, Lancaster, LA1 4YQ, UK.

[2] Lake Ecosystems Group, Centre for Ecology & Hydrology, Lancaster Environment Centre, Library Avenue, Bailrigg, Lancaster, LA1 4AP, UK.

[3] ECMWF, Shinfield Park, Reading, RG2 9AX, UK

*** Corresponding Author: t.page@lancaster.ac.uk*

Key words: Phytoplankton model, forecasting, data assimilation, Ensemble Kalman Filter, cyanobacteria, PROTECH.

Abstract

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.

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 limmnology 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

45 reducing forecast uncertainty but has, to date, received relatively little attention for forecasting
46 phytoplankton community dynamics. There is hence a need to test different DA methodologies across
47 different lake systems and different models.

48 There are still relatively few studies for operational lake forecasting systems and various approaches have
49 been taken such as using: Ensemble Kalman Filter (EnKF; Evensen, 1994) schemes and physically-based
50 simulation models (e.g. Allen *et al*., 2003, Huang *et al.* 2013 and Kim *et al,* 2014); evolutionary
51 computation (Recknagel *et al.,* 2014; Ye *et al.,* 2014); Lagrangian particle tracking model methods (Rowe
52 *et al*., 2016); and a combination of wavelet analysis and neural networks (Luo *et al*., 2011; Xiao *et al*.,
53 2017). The EnKF has been developed to deal with highly non-linear model dynamics which cannot be
54 represented well using the traditional Kalman Filter. Phytoplankton population dynamics are highly non-
55 linear with multiple modes of behaviour that can respond rapidly to threshold-type effects and are prone
56 to rapid changes in their physical and chemical environment (e.g. water temperature, light levels and
57 available nutrients). This makes the EnKF a suitable choice to exploring algal bloom forecasting when
58 coupled with a phytoplankton community model.

59 Here we assess our ability to make pseudo-real-time forecasts of phytoplankton communities in two lakes
60 in the English Lake District in the north west of England, which are prone to cyanobacteria blooms during
61 the summer. Forecasts were made using a modified version of the phytoplankton community model
62 PROTECH (Reynolds *et al*., 2001) within a DA scheme using the EnKF. The version of PROTECH employed
63 is appropriate for this problem as it is intermediate in its complexity between physically-based coupled 3-
64 dimensional hydrodynamic-biochemical models and more simplistic "black box models" which have both
65 been used in this context. More complex models are extremely computationally expensive in forecasting
66 (Huang *et al*., 2012; Recknagel, *et al*., 2014), such that only a limited number of ensemble members can
67 be used (Kim *et al*., 2014); simple black box models may not be able to represent phytoplankton

community dynamics driven by ecological strategies that are represented in phytoplankton community models such as PROTECH.

We aimed to determine the efficacy of phytoplankton community forecast simulations, evaluate the EnKF as a DA strategy and investigate the ensemble size required for making consistent forecasts. Ultimately, success will rely on the modelling strategy being sufficiently effective to capture the necessary short-term phytoplankton community dynamics, given the available meteorological forecasts and limitations associated with driving data. 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 and 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 km^2 resolution) and stochastic perturbations of model parameters (see Buizza *et al*., 1999 and Ollinaho *et al.,* 2016). The re-initialisation of model states in the ECMWF forecasting system is implemented using a higher resolution 3-hour 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 degrees 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 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 minute) 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*., 2010).

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 Figure 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 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.

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

166 **2.3.1 The PROTECH model**

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

172 Underwater light for model layer *i* is calculated using:

173 $$l_i = Isurf.e^{(-\varepsilon.d_i)} \quad (1)$$

174

175 Where: *Isurf* is the daily surface light flux, *d* is the depth from the lake surface*, ε* is the light extinction
176 coefficient resulting from the sum of lake-specific abiotic water attenuation (ε_b) and the extinction of light
177 associated with the concentration of phytoplankton at each timestep multiplied by the parameter ε_a. In
178 the layers from the surface to the epilimnetic depth, the available light is represented by the geometric
179 mean of the epilimnetic layers and hence assumes that phytoplankton spend an equal time in each layer
180 at each timestep. Phytoplankton population dynamics are simulated 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):

$$\frac{\Delta X}{\Delta t} = (r' - S - G - F).X \quad (2)$$

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:

$$r' = \min\{r'_{(\theta)}, r'_{(P)}, r'_{(N)}, r'_{(Si)}\} \quad (3)$$

where $r'_{(\theta,I)}$ is the growth rate at a given temperature (θ) and daily photoperiod (*I*) and r'_P, r'_N, r'_{Si} are the growth rates determined by phosphorus, nitrogen and silica concentrations. The final growth rate ($r'_{cor(\theta,I)}$) is a corrected rate allowing for dark respiration using equation 4. This is required as the model growth equations are net of basal metabolism but not dark respiration burden.

$$r'_{corr(\theta,l)} = R_{d(\theta)}.r'_{(\theta,l)} - \left(1 - R_{d(\theta)}.\right).r'_{(\theta,l)} \quad (4)$$

Where $R_{d(\theta)}$ is the dark respiration rate at temperature θ.

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 surface-to-volume ratios (e.g. *Chlorella*, and *Plagioselmis*); S-types which are 'stress tolerators' that tolerate relatively

low nutrient availability and strong stratification (e.g. *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 *Planktothrix*). 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*, *Aphanizomenon* and *Ceratium*) and CSR-types (e.g. *Cryptomonas*) that are intermediate between C-, S- 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 (D_e). The simplified model works on the basis of *independent* estimates of epilimnetic temperature and lake heat energy fluxes. The estimate of epilimnetic temperature (T_e) uses a TF model (see Section 2.3.3) with inputs of air temperature (T_a), solar radiation, wind speed (Ws) and D_e. Air temperature, solar radiation and wind speed are derived from the forecasts and D_e 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 T_e, river temperature and flow magnitude, day length, cloud cover, T_a, Relative Humidity and Ws.

These two independent estimates are "balanced" to obtain hypolimnetic volume (V_h) using:

$$V_h = \frac{E_{\Delta T}}{\Delta T . C_w . \rho_w} \quad (5)$$

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

$$V_e = V_t - V_h \quad (6)$$

where V_t is the total lake volume. The requirement for ΔT is satisfied by calculating T_h using:

$$T_h = \frac{E_{th}}{C_w.\rho_w.V_t} \quad (7)$$

where: E_{th} 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_{th} remains at its previous value until updated during the data assimilation step. This treatment of E_{th} neglects the explicit downward transfer of energy from $E_{\Delta T}$ to E_{th} for forecasting and assumes that these are negligible over this timescale: energy is, however, explicitly transferred downwards each time 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 $E_{\Delta T}$
    - 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 $E_{\Delta T}$ < 0 loose energy from E_h (minimum energy set to 0°C)
3. Estimate T_h from E_{th}
4. If $E_{\Delta t}$ > 0 and If T_e - T_h is greater than a threshold parameter (nominally set to 1°C) estimate epilimnetic depth by solving for the volume of water required to match $E_{\Delta T}$ given ΔT: subsequently estimate V_e and hence D_e by difference.
5.

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:

$$y_t = \frac{B_1(z-1)}{A(z-1)} U_1 + \frac{B_2(z-1)}{A(z-1)} U_2 + \cdots \frac{B_n(z-1)}{A(z-1)} U_n \quad (8)$$

where, y_t is the variable being estimated at time *t*, $U_{1\text{-}n}$ are model input vectors, $A(z-1)$ and $B_n(z-1)$ 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 (Evenson, 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; Evenson, 2009):

$$\varphi_{j,i}^{a} = I.\left(\varphi_{j,i}^{a} - \overline{\varphi_{i}^{a}}\right) + \overline{\varphi_{i}^{a}} \quad (9)$$

II. Generate *N* perturbations of the observations (Y); 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^{e}) to avoid bias in the update (Evenson, 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^{th} ensemble member. This is done proportionally to the deviation of the states in the forecasted state-parameter matrix (φ^{f}) 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:

$$\varphi^a = \varphi^f + (K(Y) - H\varphi^f) \quad (10)$$

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

$$K = P_\varphi^f H^T (H(P_\varphi^f)H^T + P^e)^{-1} \quad (11)$$

where, P_φ^f 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_h 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

311 the lakes studied here, identified that the main challenges for forecasting were uncertainties associated
312 with: representing phytoplankton exposure to light and nutrient inputs (particularly phosphorus). The DA
313 scheme was therefore defined to include the main model states, SRP, DIN, SiO_2 and chlorophyll *a*, as well
314 the parameters associated with modifying nutrient inputs and underwater light (Table 4). These were
315 updated at an approximately 14-day frequency set by the monitoring data. For Windermere, both point
316 source ($WwTW_f$) and diffuse SRP inputs (P_{fact}) parameters were included in the DA scheme; for Esthwaite
317 Water only the parameter modifying the diffuse SRP inputs was included as simulations which included a
318 simplified representation of sediment-derived SRP inputs did not provide improved results (these results
319 are not reported here).
320 To investigate the effect of ensemble size and to determine an acceptable ensemble size for the current
321 applications, ensemble member (EM) size was increased sequentially, using the scenarios EM50, EM100,
322 EM200, EM300 and EM400 (where the suffix is the size of the ensemble), until the forecast simulations
323 appeared consistent. These scenarios were generated by resampling the downscaled ECMWF forecast
324 distributions as described above and were used to force the suite of models used. For each of the forecast
325 scenarios, the error associated with the assimilated data and the variance inflation factors were
326 "optimised" manually to provide the best results. For consistency, and in the spirit of the pseudo-real time
327 treatment of the forecast simulations, the variance inflation factors were kept consistent across all lake-
328 years considered. For each of the assimilated variables, the variance was assumed to be proportional to
329 the magnitude of the variable of interest using a percentage. Additionally, a minimum variance was
330 applied to reduce the impact of very small observed values (e.g. where epilimnetic SRP values are
331 observed to be very low or within the limit of detection) where the associated low variance would falsely
332 indicate low uncertainty.

333 **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 and Kim *et al.*, 2014) or solely a measure of the goodness-of-fit to observations (e.g. the coefficient of determination, R_T^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 observations 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_T^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_T^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_T^2 of 0.85 and 0.75 respectively for a 10-day-ahead forecast; Figs. 2a and 2b) 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_T^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 and 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 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 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 Windemere 2008 and 2009 (Figs 7a and 7d) 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. 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: Figs 7c, 7f & 7i). 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/J017299/1); We would like to thank the ECMWF for the historic meteorological forecasts and Mr Bernard Tebay for collecting the rainfall data at Ambleside, Cumbria.

References

Anderson, J.L. (2007). An adaptive covariance inflation error correction algorithm for ensemble filters. Tellus, 59A, 210–224

Anderson, J.L., Anderson, S.L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Mon. Wea. Rev. 127, 2741–2758.

494 J. Allen, M. Eknes, G. Evensen. (2003). An Ensemble Kalman Filter with a complex marine ecosystem
495 model: hindcasting phytoplankton in the Cretan Sea. Ann. Geophys., 21, 399–411.

496 Bennion, H., Monteith, D. and Appleby, P. (2000). Temporal and geographical variation in lake trophic
497 status in the English Lake District: evidence from (sub)fossil diatoms and aquatic macrophytes.
498 Freshwater Biol., 45(4), 1365-2427, doi: 10.1046/j.1365-2427.2000.00626.x

499 Brookes, J.D. and Carey, C.C. (2011).Resilience to blooms. Science 334 (6052), 46-47; DOI:
500 10.1126/science.1207349.

501 Buizza, R., Milleer, M. and Palmer, T. N. (1999). Stochastic representation of model uncertainties in the
502 ECMWF ensemble prediction system. Q. J. Roy. Meteor. Soc. 125 (560), 2887–2908.
503 doi:10.1002/qj.49712556006.

504 Carmichael, W.W. (1992). A status report on planktonic cyanobacteria (blue-green algae) and their
505 toxins. EPA/600/R-92-079, Environmental Monitoring Systems Laboratory, Office of Research and
506 Development, U.S. Environmental Protection Agency, Cincinnati, OH. 141 pp.

507 Dodds W.K. Bouska , W.W., Eitzmann , J. L., Pilger , T. J., Pitts, K. L., Riley, A.J. Schloesser, J.T. and
508 Thornbrugh, D.J. (2009). Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages.
509 Environ. Sci. Technol, 43, 12–19.

510 Dong X., Bennion H., Maberly S.C., Sayer C.D., Simpson G.L. and Battarbee R.W. (2012). Nutrients
511 provide a stronger control than climate on diatom communities in Esthwaite Water Water: Evidence
512 from monitoring and palaeolimnological records over the past 60 years. Freshwater Biol., 57, 2044-
513 2056.

514 Elliott, J.A. (2010). The seasonal sensitivity of Cyanobacteria and other phytoplankton to changes in
515 flushing rate and water temperature. Glob. Change Biol., 16, 864-876.

516 Elliott, J.A. (2012) Predicting the impact of changing nutrient load and temperature on the
517 phytoplankton of England's largest lake, Windermere. Freshwater Biol., 57, 400-413.

518 Evensen, G. (1994). Sequential data assimilation with a non-linear quasigeostrophic model using Monte
519 Carlo methods to forecast error statistics. J. Geophys. Res., 99, 10 143–10 162.

520 Evensen, G. (2009). The ensemble Kalman filter for combined state and parameter estimation. IEEE
521 Control Sys., 29 (3), pp. 83-104, 2009.

522 George, D. G. and Heaney, S. I. (1978). Factors influencing the spatial distribution of phytoplankton in a
523 small productive lake. J. Ecol., 66(1), 133-155.

524 Hall, G.H., Maberly, S.C., Reynolds, C.S., Winfield, I.J., James, B.J., Parker, J.E., Dent, M.M., Fletcher, J.M.,
525 Simon, B.M. and Smith, E. (2000). Feasibility study on the restoration of three Cumbrian lakes. Centre for
526 Ecology and Hydrology Windermere, Ambleside, UK. 82 pp.

527 Heany, S.I., Corry, J. E. and Lishman, J. P. (1992). Changes of water quality and sediment phosphorus of
528 a small productive lake following decreased phosphorus loading. Centre for Ecology and Hydrology
529 Windermere, Ambleside, UK. 14 pp.

530 Ho, J. C. and Michalak, A. M. (2015). Challenges in tracking harmful algal blooms: A synthesis of evidence
531 from Lake Erie. J. Great Lakes Res., 41(2), 317-325. doi.org/10.1016/j.jglr.2015.01.001

532 Huang, J., Gao, J., Liu, J. and Zhang, Y. (2013). State and parameter update of a hydrodynamic-
533 phytoplankton model using ensemble Kalman filter, Ecol. Model., 263 (10), 81-91.
534 https://doi.org/10.1016/j.ecolmodel.2013.04.022

Kim, K., Park, M., Min, J., Ryu, I., Kang, M., and Park, L. (2014). Simulation of algal bloom dynamics in a river with the ensemble Kalman filter. J. Hydrol., 519(D), 2810–2821. https://doi.org/10.1016/j.jhydrol.2014.09.073

Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J. S. and Schimel, D. S. (2011). Ecological forecasting and data assimilation in a data-rich era. Ecol. Appl., 21, 1429–1442. doi:10.1890/09-1275.1

Maberly, S.C., De Ville, M.M., Thackeray, S.J., Feuchtmayr, H., Fletcher, J.M.,James, J.B., Kelly, J.L., Vincent, C.D., Winfield, I.J., Newton, A., Atkinson, D., Croft,A., Drew, H., Saag, M., Taylor, S., Titterington, H. (2011). A survey of the lakes of the English Lake District: The Lakes Tour 2010. NERC/Centre for Ecology and Hydrology,137pp. (CEH Project Number. Report to: Environment Agency, North West Region and Lake District National Park Authority: downloaded Jan 2015 from http://nora.nerc.ac.uk/14563/2/N014563CR.pdf

Mackay E. M., Folkard A. M. and Jones I.D. (2014). Interannual variations in atmospheric forcing determine trajectories of hypolimnetic soluble reactive phosphorus supply in a eutrophic lake. Freshwater Biol., 59, 1646–1658.

Madgwick G., Jones I.D., Thackeray S.J., Elliott J.A. and Miller H.J. (2006). Phytoplankton communities and antecedent conditions: high resolution sampling in Esthwaite Water Water. Freshwater Biol., 51, 1798–1810.

Marcé, R., George, G., Buscarinu, P., Deidda, M, Dunalska, J., de Eyto, E., Flaim, G., Grossart, H., Istvanovics, V., Lenhardt, M., Moreno-Ostos, E., Obrador, B., Ostrovsky, I., Pierson, D. C., Potužák, J., Poikane, S., Rinke, K., Rodríguez-Mozaz, S., Staehr, P. A., Šumberová, K., Waajen, G., Weyhenmeyer, G. A., Weathers, K. C., Zion, M., Ibelings, B.W. and Jennings, E. (2016). Automatic High Frequency

Monitoring for Improved Lake and Reservoir Management. Environ. Sci. Technol. 50 (20), 10780-10794. DOI: 10.1021/acs.est.6b01604

Michalak, A., M. (2016). Study role of climate change in extreme threats to water quality. Nature 535, 349-350.

Metcalf, J.S. and Codd, G.A. (2009). Cyanobacteria, neurotoxins and water resources: are there implications for human neurodegenerative disease? Amyotrophic Lateral Sclerosis 10, suppl. 2, 74-78 (2009).

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Hauser (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter, Adv. Water Resour., 28, 135 – 147.

Ollinaho, P., Lock, S.-J., Leutbecher, M., Bechtold, P., Beljaars, A., Bozzo, A., Forbes, R. M., Haiden, T., Hogan, R. J. and Sandu, I. (2017), Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble. Q.J.R. Meteorol. Soc., 143: 408–422. doi:10.1002/qj.2931

Page et al., (2017). Constraining uncertainty and process-representation in an algal community lake model using high frequency in-lake observations. Ecol. Model.: http://www.sciencedirect.com/science/article/pii/S0304380017301345

Paerl, H.W. and Huisman, J. (2008). Blooms like it hot. Science, 4, 320(5872), 57-8. doi: 10.1126/science.1155398. DOI: 10.1126/science.1155398

Pretty, J. N., Mason, C. F., Nedwell, D. B., Hine, R. E., Leaf, S., and Dils, R. (2003). Environmental Costs of Freshwater Eutrophication in England and Wales. Environ. Sci. Technol., 37(2), 201-208.

576 Read J.S., Hamilton, D.P., Jones, I.D., Muraoka, K., Winslow, L.A. , Kroiss, R. , Wu, C.H. & Gaiser. E. (2011).
577 Derivation of lake mixing and stratification indices from high-resolution lake buoy data. Environ. Modell.
578 Softw. 26, 1325-1336.

579 Ramsbottom A.E. (1976). Depth Charts of the Cumbrian Lakes. Freshwater Biological Association
580 Scientific Publication No. 33, Ambleside, UK.

581 Recknagel, F., Ostrovsky, I. and Cao, H. (2014). Model ensemble for the simulation of plankton
582 community dynamics of Lake Kinneret (Israel) induced from in situ predictor variables by evolutionary
583 computation. Environ. Modell. Soft., 61, 380-392. https://doi.org/10.1016/j.envsoft.2014.03.014.

584 Reynolds, C. S. (1984). The Ecology of Freshwater Phytoplankton. Cambridge University Press,
585 Cambridge.

586 Reynolds C.S. (1988). Functional morphology and the adaptive strategies of freshwater phytoplankton.
587 In: Growth and Reproductive strategies of Freshwater Phytoplankton (Ed. C.D. Sandgren), pp. 388–433.
588 Cambridge, University Press, New York.

589 Reynolds C.S., Irish A.E. and Elliott J.A. (2001). The ecological basis for simulating phytoplankton
590 responses to environmental change (PROTECH). Ecol. Model., 140, 271–291.

591 Rigosi, A., Carey, C.C., Ibelings, B. W. and Brookes, J. D. (2014). The interaction between climate warming
592 and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa.
593 Limnol. Oceanogr. 59(1), 2014, 99–114. doi:10.4319/lo.2014.59.01.0099

594 Rowe, M. D., Anderson, E. J. , Wynne, T. T., Stumpf, R. P., Fanslow, D. L., Kijanka, K., Vanderploeg, H. A.
595 Strickler, J. R. and Davis, T. W. (2016). Vertical distribution of buoyant Microcystis blooms in a
596 Lagrangian particle tracking model for short-term forecasts in Lake Erie. J. Geophys. Res.: Oceans. 121,
597 5296-5314. doi:10.1002/2016JC011720.

598 Smith, V.H., (2003). Eutrophication of Freshwater and Coastal Marine Ecosystems: A Global Problem.
599 Environ. Sci. & Pollut. Res. 10 (2) 126-39.

600 Stumpf, R. P., Tomlinson, M. C., Calkins, J. A., Kirkpatrick, B., Fisher, K., Nierenberg, K., Currier, R. and
601 Wynne, T. T. (2009). Skill assessment for an operational algal bloom forecast system. Journal of Marine
602 Systems. 76(1): 151-161.

603 Taylor, C.J., Pedregal, D.J., Young, P.C. and Tych, W., (2007). Environmental time series analysis and
604 forecasting with the Captain toolbox, Environ. Modell. Softw., 22: 797-814.

605 World Health Organization (1999). Toxic cyanobacteria in water: a guide to their public health
606 consequences, monitoring and management. I. Chorus and J. Bartram (Eds.). E & FN Spon, London, UK
607 (1999).

608 Xiao X, Sogge H, Lagesen K, Tooming-Klunderud A, Jakobsen KS, Rohrlack T (2014). Use of High
609 Throughput Sequencing and Light Microscopy Show Contrasting Results in a Study of Phytoplankton
610 Occurrence in a Freshwater Environment. PLoS ONE, 9(8), 1-9. doi:10.1371/journal.pone.0106510

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

614 Ye, L., Cai, Q., Zhang, M. and Tan, L. (2014). Real-time observation, early warning and forecasting
615 phytoplankton blooms by integrating in situ automated online sondes and hybrid evolutionary
616 algorithms. Ecological Informatics, 22, 44–51.

617 Young, P.C., 2015. Refined Instrumental Variable Estimation: Maximum Likelihood Optimization of a
618 Unified Box-Jenkins Model. Automatica, 52, 35–46.

Supplementary information

Supp. 1 Transfer Function models for forecasted inputs

The epilimnetic depth model requires forecasts of epilimnetic temperature, river in/outflows and river temperature. Each TF model that provides these forecasts was identified (as outlined above) using the available timeseries data. The epilimnetic temperature (T_e) at day *t* is given by:

$$T_{e(t)} = -a.T_{e\,(t-1)} b1.T_{a(t)} + b2.R_{sw(t)} + b3.\frac{1}{D_{e(t-1)}} + b4.\left(W_{s(t-1)}\right)^3$$

Where, T_a is the air temperature, R_{sw} is SW radiation, D_e is epilimnetic depth and W_s is the wind speed. The model coefficients are denoted *a*, *b1*, *b2* and *b3* (see Table Supp. 1 for values). One model for each lake was identified from the available data (2008 to 2010 for Windermere and 2004 to 2009 for Esthwaite Water).

The lake in/outflow TF model was identified as a 1st order model with a nonlinear rainfall filter (see Young and Beven, 1994) and took the form:

$$Q_{r(t)} = -a.Q_{r(t-1)} + b.P_{(t)}.{Q_{r(t-1)}}^{\beta}$$

where Q_r is the river in/outflow, *P* is precipitation and *a*, *b1* are TF model coefficients where β is the nonlinear rainfall filter parameter. The model for Windermere was identified using Rainfall data from Ambleside and flow data from the Environment agency Gauge at Newby Bridge for the years 2008 to 2010 (National River Flow Archive: http://www.ceh.ac.uk/data/nrfa/).

River temperature (T_Q) was estimated using observed data from Troutbeck (Windermere) for the years 1997 to 2006:

$$T_{Q(t)} = -a.T_{Q(t-1)} + b.T_{a(t)}$$

References (supplementary information)

Young, P.C. and Beven, K.J. 1994. Data-based mechanistic modelling and the rainfall-flow non-linearity. Environmetrics. 5, 3, p. 335-363.

Table 1 Study Lakes and primary characteristics[§]

Name/location	Mean Depth (m)	Max. Depth (m)	Max. Length (m)	Volume (m^3)	Catchment Area (km^2)	Residence Time (days)
Windermere (South Basin)	16.8	41	9300	1.06×10^8	230.5	100
Esthwaite Water	6.4	15.5	2500	5.97×10^6	17.1	100

[§] Details from Ramsbottom (1976)

Table 2 Forcing inputs and downscaling relationships

Model Inputs	Downscaling factor/relationship	Uncertainty sampled
Air Temp (T_a; K)	Windermere: $0.095(T_a^{§}) + 279.75$** Esthwaite Water: $0.013(T_a^{§}) + 280.16$**	Y (Regression)
Solar Radiation (SR; Wm^{-2})	0.85	N
Wind Speed (W; m s^{-1})	0.38[¥]	Y (Gamma Dist.)
Relative Humidity (RH; %)	1	N
Cloud Cover (Cc; eighths)	1.25	N
Rainfall (R; mm)	3	N
Nutrient Inputs (P; N; SiO_2/ mg m^{-3})	See section 2.2.3	Y (Gamma Dist.)

*$Ta^{§}$ is the forecast air temperature (K); ** see Section 2.2.2 for additional lake-effect correction; [¥] see Section 2.2.2 for additional wind direction correction.*

Table 3 Observed data assimilated in the EnKF scheme

Assimilated state	Frequency	Source
Epilimnetic Temperature (^oC)	Daily	buoy obs.
Hypolimnetic Temperature (^oC)	Daily	buoy obs.
Epilimnetic depth (m)	Daily	buoy obs.
Chllorophyll a (mg m^{-3})	≈14 days	Monitoring
Nutrient Inputs (SRP; N; SiO_2 / mg m^{-3})	≈14 days	Monitoring

Table 4. States and parameters included in the ENKF scheme

State/Parameter	Acceptable range	Observational error (%)	Initial distributions (uniform)**
Epilimnetic Temp. (T_e, ^{0}C)	2-25	5	5.5-7 (W); 4-6(E)
Hypolimnetic temp. (T_h, ^{0}C)	2-25	10	5.5-7 (W); 4-6(E)
Epilimnetic depth (D_e, m)	0.5-max. depth	5	41 (W); 15.5(E)
Chlorophyll *a* (mg m^{-3})	1e^{-6}-1e^{3}	10	3-4.5 (W); -4.5-6 (E)
Background light extinction (ε_b, m^{-1})	0.15-0.9	N/A	0.15-0.6(W); 0.45-0.75(E)
Epilimnetic P conc. (P_e, mg m^{-3})	1e^{-6}-1e^{4}	25	10-20(W); 8-15(E)
Epilimnetic DIN conc. (N_e, mg m^{-3})	1e^{-6}-1e^{4}	25	400-700(W); 500-1100(E)
Epilimnetic SiO_2 conc. (Si_e, mg m^{-3})	1e^{-6}-1e^{4}	25	1500-2500(W); 2000-2500(E)
Diffuse P input multiplier (P_f, dimensionless)	0.05-7	N/A	0.01-1.5
Diffuse DIN input multiplier (N_f, dimensionless)	0.1-3	N/A	0.5-1.2
Diffuse SiO_2 input multiplier (Si_f, dimensionless)	0.1-3	N/A	0.5-1.2
Point source P input multiplier ($WwTW_f$, dimensionless)	0.01-2	N/A	0.1-1.4

****** Where distributions are different for each lake W = Windermere; E = Esthwaite Water

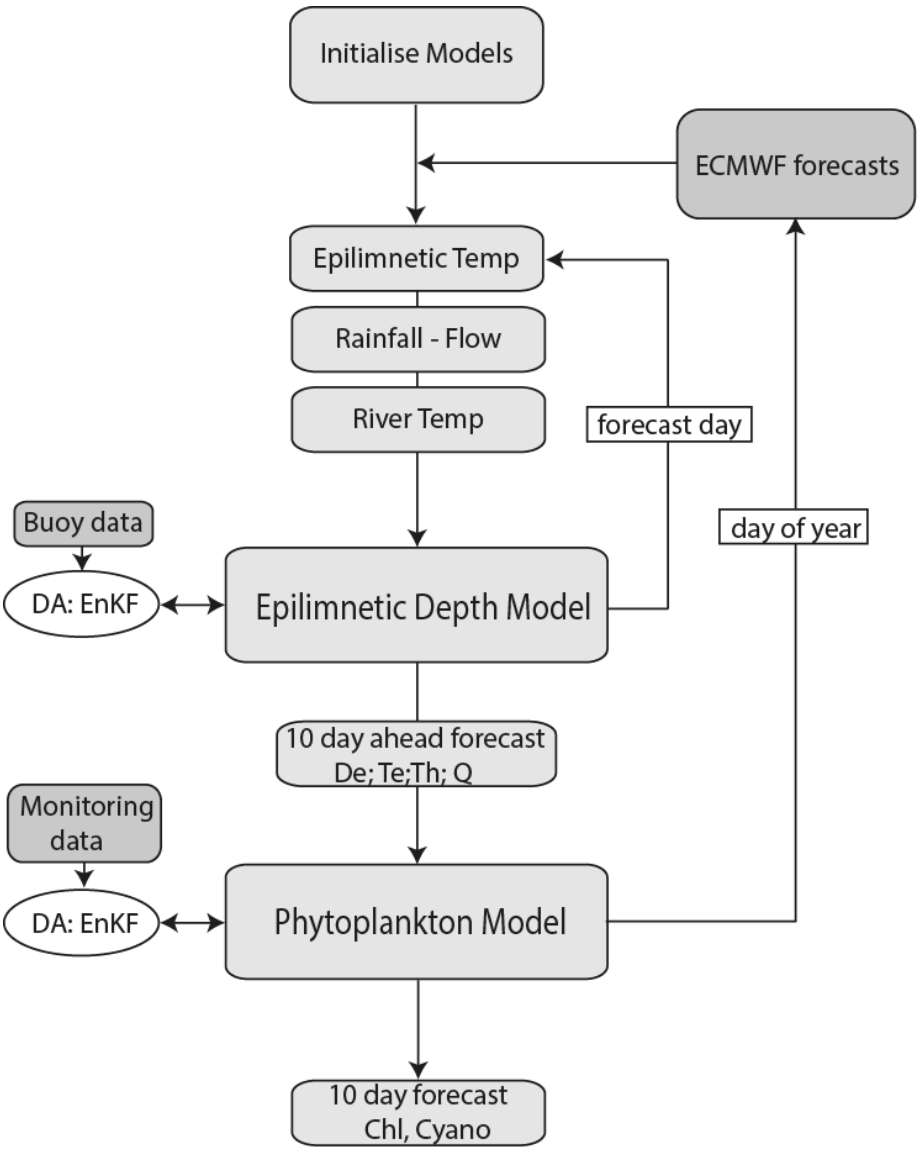
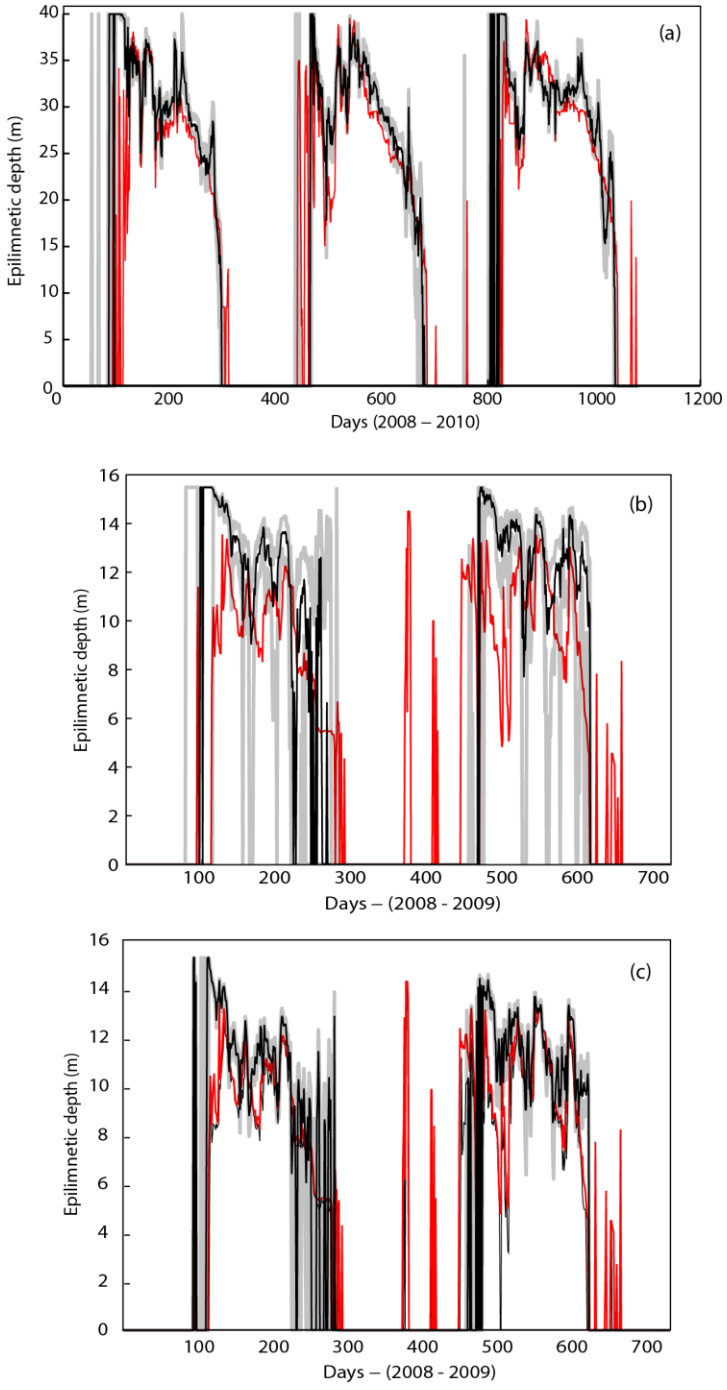
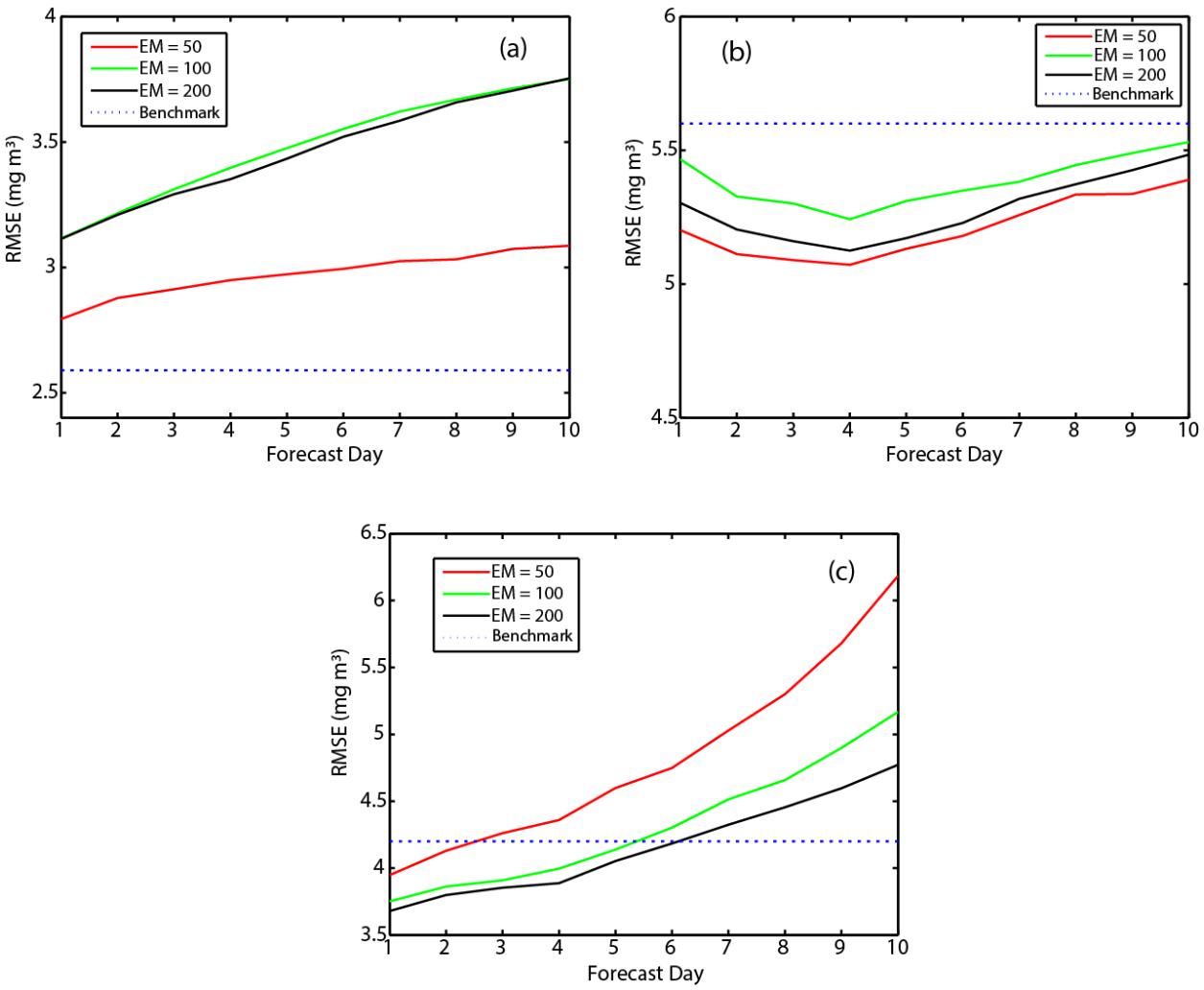


Figure 1. Schematic diagram of the forecasting system. The schematic shows sequential model input-output structure and DA strategy. De is epilimnetic depth; Te is epilimnetic temperature; Th is hypolimnetic temperature, Q is lake inflow/outflow and Chl and Cyano are the concentration of total phytoplankton chlorophyll a and cyanobacterial chlorophyll a respectively.



669

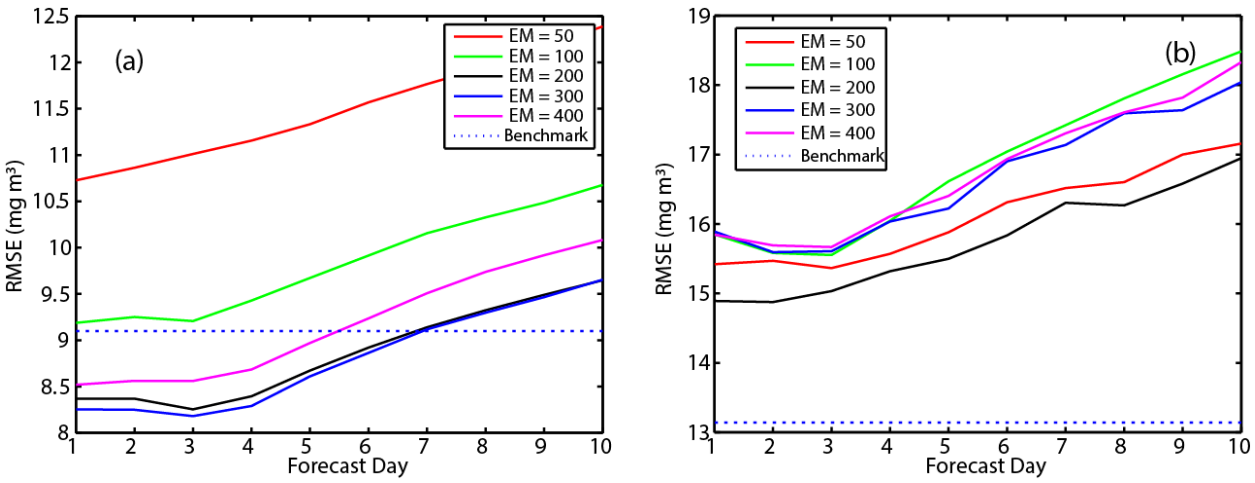
670 Figure 2. Simulated and measured epilimnetic depth. Results shown for (a) Windermere 2008-2010 10-
671 day-ahead, (b) Esthwaite Water 2008 and 2009 10-day-ahead and (c) Esthwaite Water 2008 and 2009 3-
672 day-ahead: "observed" epilimnetic depth (red line), 50th percentile of the ensemble of simulated
673 epilimnetic depth (black line) and 5th and 95th percentiles (grey lines).



674

675 Figure 3. Chlorophyll a forecast skill for the differing ensemble size scenarios. Results are shown for (a)
676 Windermere 2008, (b) Windermere 2009 and (c) Windermere 2010, compared to the benchmark
677 persistence forecast. Note that lower ensemble sizes can give "randomly" better forecast performance
678 (e.g. EM = 50 in pane (a))

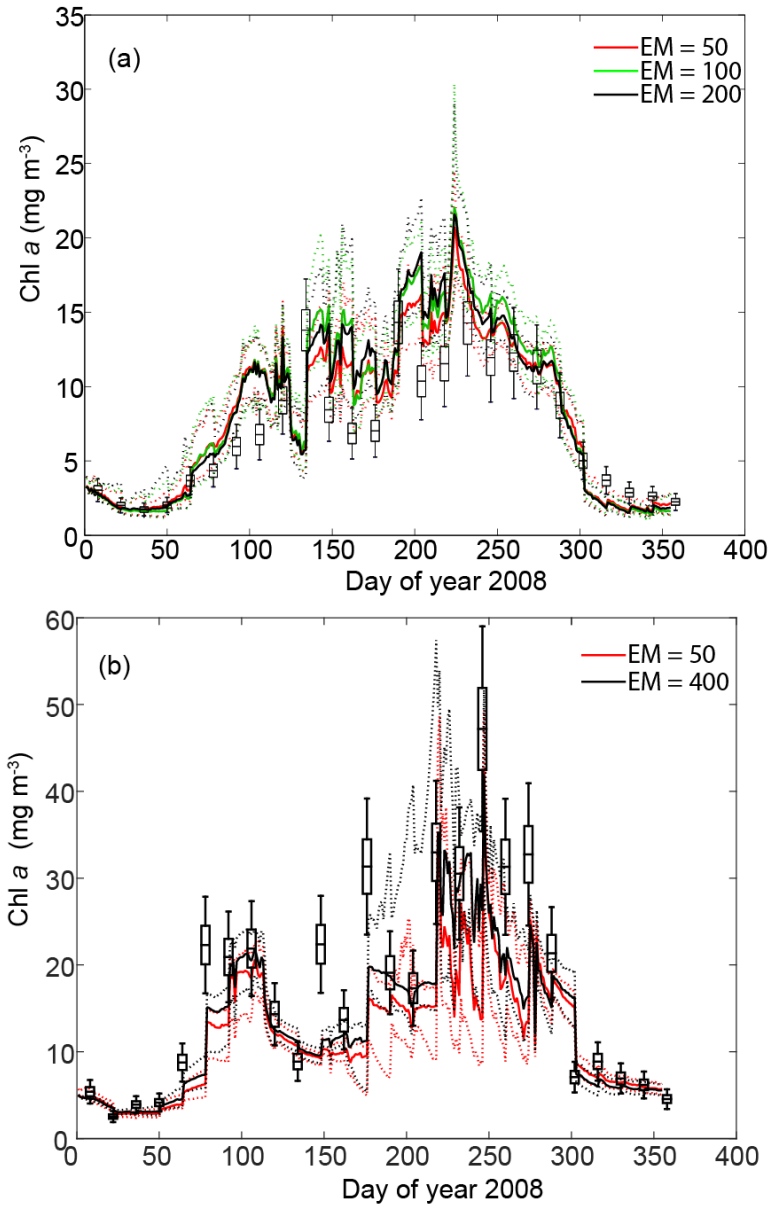
679

680 

681 Figure 4. Chlorophyll a forecast skill for the differing ensemble size scenarios. Results are shown for (a)

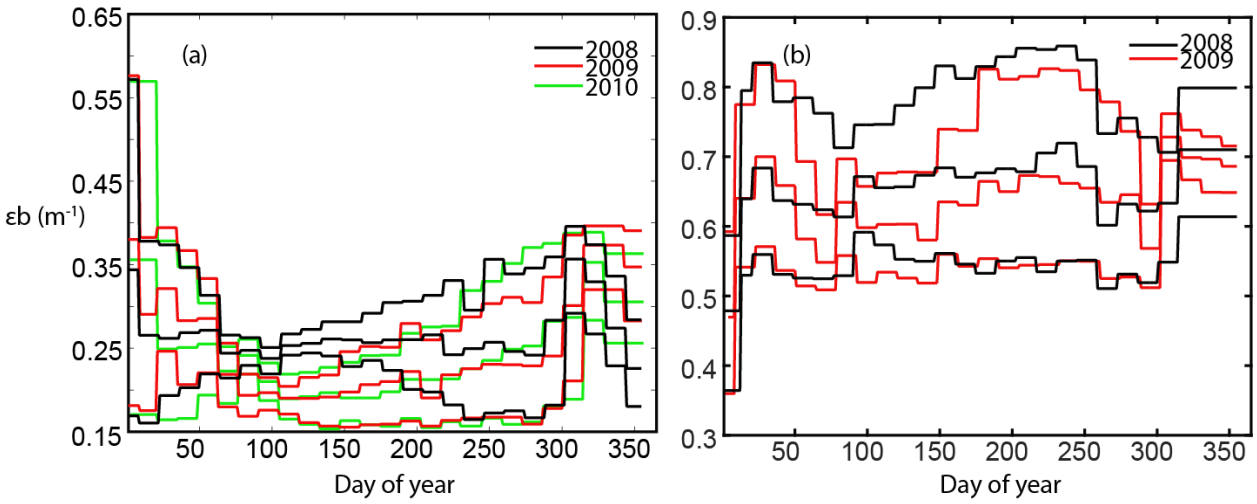
682 Esthwaite Water 2008 and (b) Esthwaite Water 2009, compared to the benchmark persistence forecast.

683



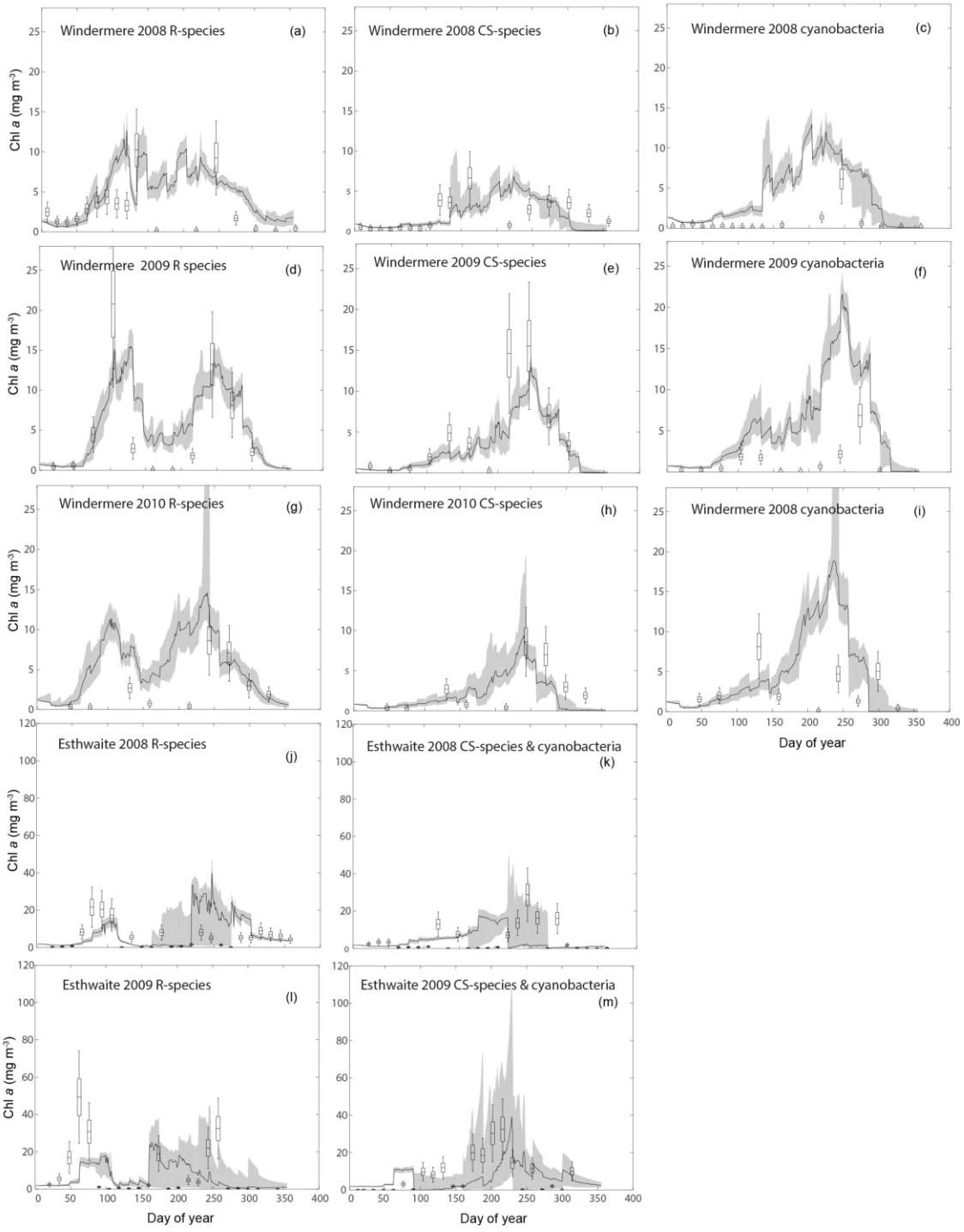
684

685 Figure 5. Measured and forecast phytoplankton chlorophyll a in the two lakes during 2008. Results
686 show concatenated forecasts for: (a) 10-day-ahead for Windermere 2008 for ensemble member sizes
687 (EM) of 50, 100 and 200; (b) 5-day-ahead for Esthwaite Water 2008 for ensemble member sizes (EM) of
688 50 and 400. Solid lines are 50th percentile of ensemble and dotted lines are 5th and 95th percentiles.
689 The box and whisker symbols represent the analytical uncertainty and the total uncertainty of +/- 8%
690 and +/- 25% (see Page et al, 2017).



691

692 Figure 6. The evolution of the background light extinction coefficient parameter (εb). Results are shown
693 for (a) Windermere 2008, 2009 and 2010 and (b) Esthwaite Water 2008 and 2009. The three lines in
694 each colour are the 5th, 50th and 95th percentiles of the EM200 (Windermere) and EM400 (Esthwaite
695 Water) ensembles.



696

697 Figure 7. Concatenated five-day ahead forecasts of R-species, CS-species and cyanobacteria

698 concentration for all lake years; black line is 50th percentile and grey shaded area represents the 5th

699 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.

Table Supp. 1 Transfer Function parameters and goodness of fit (W = Windermere; E = Esthwaite Water)

	a		b1 (β)		b2		b3		b4		τ		R_T^2	
	W	**E**	**W**	**E**	**W**	**E**	**W**	**E**	**W**	**E**	**W**	**E**	**W**	**E**
Lake Surface Temperature (T_s)	-0.9449	-0.899	0.055	0.093	0.0008	0.0025	0.0011	0.0022	-0.0007	-0.0012	[0,0,0,0]	[0,1,1,0]	0.97	0.98
River in/outflow (Q_r)	-0.7717	-0.829	11.141 (0.2)	0.022 (0.3)			-	-			1	0	0.92	0.86
River Temperature (T_Q)	-0.900	-0.900	0.1005	0.1005	-	-	-	-	-	-	0	0	0.87	0.87

Table Supp. 2. Species used to represent algal communities. Functional algal types and an indication of classification as cyanobacteria given are in parenthesis: functional types follow **Reynolds (1988).**

Windermere	**Esthwaite Water**
Aphanizomenon flos-aquae (CS; Cyano)	*Asterionella (R)*
Aulacoseira (R)	*Aulacoseira* - 2008 *(R)*; *Fragilaria crotonensis-(2009 (R)*
Asterionella (R)	*Aphanizomenon flos-aquae (CS; Cyano)*
Cryptomonas (CSR)	*Aphanothece clathrata (CS; Cyano)*
Dolichospermum (CS; Cyano)	*Cryptomonas (CSR)*
Monoraphidium (CS)	*Dictyosphaerium pulchellum (R)*
Paulschulzia tenera (S)	*Dolichospermum (CS; Cyano)*
Planktothrix (R; Cyano)	*Eudorina (S)*