Water Research 134 (2018) 74–85

Contents lists available at ScienceDirect

Water Research

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

# 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

*Keywords:*
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 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

* Corresponding author.
*E-mail address:* t.page@lancaster.ac.uk (T. Page).

https://doi.org/10.1016/j.watres.2018.01.046

(Recknagel et al., 2014; Ye et al., 2014); Lagrangian particle tracking model methods (Rowe et al., 2016); and a combination of wavelet analysis and neural networks (Luo et al., 2011; Xiao et al., 2017). The EnKF has been developed to deal with highly non-linear model dynamics which cannot be represented well using the traditional Kalman Filter. Phytoplankton population dynamics are highly non-linear with multiple modes of behaviour that can respond rapidly to threshold-type effects and are prone to rapid changes in their physical and chemical environment (e.g. water temperature, light levels and available nutrients). This makes the EnKF a suitable choice to exploring algal bloom forecasting when coupled with a phytoplankton community model.

Here we assess our ability to make pseudo-real-time forecasts of phytoplankton communities in two lakes in the English Lake District in the north west of England, which are prone to cyanobacteria blooms during the summer. Forecasts were made using a modified version of the phytoplankton community model PROTECH (Reynolds et al., 2001) within a DA scheme using the EnKF. The version of PROTECH employed is appropriate for this problem as it is intermediate in its complexity between physically-based coupled 3-dimensional hydrodynamic-biochemical models and more simplistic "black box models" which have both been used in this context. More complex models are extremely computationally expensive in forecasting (Huang et al., 2013; Recknagel et al., 2014), such that only a limited number of ensemble members can 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; 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; 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 time-step 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

**Table 1**
Study Lakes and primary characteristics.[a]

| 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 | $1.06 \times 10^8$ | 230.5 | 100 |
| Esthwaite Water | 6.4 | 15.5 | 2500 | $5.97 \times 10^6$ | 17.1 | 100 |

[a] Details from Ramsbottom (1976).