Whitney Woelmer, Master’s Prospectus Outline

**Overall Introduction**

1. Importance of freshwater ecosystems and the many services they provide (setting up why their water quality is so important, as well as what services are the highest priority- e.g., clean water for drinking, etc.)
2. Threats to freshwater systems in a changing world have been increasing the occurrence and severity of eutrophication in freshwaters
   1. Watershed effects which cascade down to freshwater systems
      1. Land use shifts can lead to quicker deliver of nutrients to water bodies and increases in the amount of nutrients water bodies receive
      2. Increasing human population also leads to increases in nutrients through human waste and urban pollution, which negatively impact water quality
      3. Nutrient effects cause rapid changes in phytoplankton populations, understanding how nutrients are processed within a reservoir will help to forecast phytoplankton. By studying nutrient and phytoplankton dynamics at a high-resolution spatial scale, we can gain an understanding of chemical and biological hotspots of activity and further identify areas of greater risk of eutrophication under a changing climate.
   2. Further, global climate change in unpredictable, meaning we don’t know exactly what pressures will effect freshwater systems
3. Research to understand how systems are currently changing and developing efforts to anticipate their future response is more important than ever
   1. to allow society to coexist with changing ecosystems
   2. for managers to preemptively manage/anticipate impending poor water quality events
   3. Lake Erie 2014 example, major social and economic loss

**Proposed Research**

*Chapter 1: Developing near-term forecasts of phytoplankton in a drinking water reservoir***Introduction**

As our world continues through an era of change and human populations continue to grow, understanding not only the current state of our freshwater ecosystems, but predicting how they will respond tomorrow, next week, and next year is of utmost importance. The current shift in the scientific community towards publishing data in publicly accessible venues, as well as the development of high-frequency sensors which result in massive amounts of data have both made it possible for the field of ecological forecasting to develop now. Therefore, the ability to forecast the future state of our drinking water sources is of utmost importance to society and freshwater ecology as a discipline.

Forecasting as a practice has been developing for decades in many disciplines and has substantial breadth. Uses of forecasting can be found in many fields and applications, ranging from the well-known and widely-used weather forecasts to epidemiological forecasts of population mortality (Lee and Carter, 1992) and global Alzheimer’s trends (Brookmyer et al 2007) to forecasts of bankruptcy declaration in the United States (Shumway et al 2001). Many of these fields have been priming their forecasting abilities for decades (Shumway et al 2001). For example, SENTENCE ABOUT HOW ADVANCED WEATHER FORECASTING IS AND THE TRIAL AND ERRORS GONE THROUGH. However, the development of forecasts specific to ecological systems and ecosystem services is still in its relative infancy (Dietze book YEAR).

While being an emerging field within forecasting, ecological forecasting currently spans numerous disciplines, with a variety of purposes. To better understand the current state of studies using ecological forecasting, we conducted a small literature review. Studies were found using the Google Scholar database and the search term ‘ecological forecasting.’ Studies which met the following requirements were selected as using ecological forecasting methods: 1) use models, 2) quantify uncertainty to make a probabilistic forecast, and 3) run the model outside a given training period (Table 1).

Because the field of ecological forecasting is relatively new, there is not yet a consensus as to the best approach for predicting water quality. Empirical methods are favored a majority of times (65%, n= 17, Table 1) of current forecasting studies selected. Empirical approaches are popular because of their data-driven character; they are inherently suited for a particular system because they are based on past trends within that system. Input data for empirical models are commonly available through routine monitoring of a system, and these models do not require specific input data, but can be developed using a variety of ecological important driver variables. Additionally, empirical models are relatively easy and quick to develop and implement. However, because empirical models are built on the historic conditions of a system, if future conditions are outside the realm of past conditions, models might no longer be able to capture the mechanism responsible for changes. In contrast, process-based models were favored in just over a third of the studies examined (n=17, Table 1). It is not surprising that process-based models are currently less frequently used in the ecological forecasting literature given the hefty amount of input data often required to run the model, as well as the amount of time and expertise required of both the system being modeled and the model itself in order to properly calibrate the multitude of parameters within the model. However, these models may be especially useful in the current era of global change given their ability to explain underlying mechanisms which cause a given response, and are likely more generalizable to other systems because they are based on a set of fundamental equations rather than the historical pattern of a system.

The studies examine in our literature study spanned the realm of basic and applied research, showing motivations to understand the current state of science or to apply a forecast product to better informing management. While a majority of the studies in our literature review were conducting basic scientific research (73%, n = 15 ), all of the studies touting applied intentions were focused on informing management of freshwater systems in particular. This illustrates not only the interest in the scientific community to understand changing freshwater systems, but the need by managers and stakeholders for probabilistic forecasts in order to cope with and adapt to changing freshwater system

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 1. Summary of literature review targeting studies using ecological forecasting methods | | | | |
| Authors & Year | Best Model | Forecast Product | Use of forecast product (basic or applied) | Reference |
| Araujo et al. 2005 | Empirical | Bird species ranges | Basic |  |
| Araujo et al. 2006 | Empirical | Amphibian and reptile species distributions | Basic |  |
| Brown et al. 2013 | Process-based and empirical | Chesapeake Bay water quality | Applied |  |
| Dean et al. 2004 | Process-based | Carbon sequestration in forests | Basic |  |
| Estes et al. 2013 | Empirical | Productivity and suitability of crops in South Africa | Basic |  |
| Fenocci et al. 2019 | Process-based | Phytoplankton dynamics | Basic |  |
| Hazen et al. 2017 | Empirical | Blue whale density | Applied |  |
| Lindegren et al. 2013 | Process-based | Baltic cod dynamics | Basic |  |
| Liu et al. 2006 | Empirical | Coral Reef Bleaching | Applied |  |
| Martinez-Meyer et al. 2004 | Empirical | Ecological niches of mammal species | Basic |  |
| Perretti et al. 2013 | Empirical | Species abundance | Basic |  |
| Stow et al. 2003 | Process-based and empirical | Estuarine water quality | Applied |  |
| Thomas et al. 2018 | Empirical | Phytoplankton dynamics | Basic |  |
| Thuiller et al. 2004 | Empirical | Tree species distributions | Basic |  |
| White and Nemani 2004 | Process-based | Soil water | Basic |  |

My first chapter will focus on addressing the knowledge gap in ecological forecasting regarding how best to forecast water quality. I will produce hindcasts and near-term iterative forecasts of phytoplankton in a drinking water reservoir using both an empirical and a process-based approach. Model performance will be assessed by a suite of metrics addressing both the ability of the model to capture overall dynamics, as well as extreme events. My work will specifically address the following questions:

* + Question 1: How well can an empirical model and a process-based model hindcast observed chlorophyll dynamics?
  + Question 2: How well can an empirical model and a process-based model forecast near-term chlorophyll-a dynamics over a 16-day period (assessed through quantified uncertainties of a probabilistic forecast compared to observed dynamics)?
  + Question 3: What information does an ensemble model approach provide for chlorophyll forecasting that cannot be explained from a single-model approach?
* This Chapter 1 will thus span the field of applied and basic forecasting science by**: 1) informing managers and decision-makers about which variables are most important for routine monitoring and management of water quality in the face of land use and climate change, 2) providing essential information to the freshwater research community regarding modeling approaches to forecasting water quality, and 3) expanding the scientific field of ecological forecasting that will guide forecasting applications across many different ecosystems.**

**Methods**

*Study Site*

Falling Creek Reservoir (FCR) is a small (~12 ha), shallow (maximum depth < 9.3m) dimictic drinking water reservoir located in southwestern Virginia. FCR is owned and operated by the Western Virginia Water Authority (WVWA). The watershed of FCR is almost entirely forested, although the reservoir continues to exhibit incidences of poor water quality as a result of historical eutrophication of the reservoir (Gerling et al 2016). The major water source to FCR comes from a single tributary which flows from Beaverdam Reservoir (BVR; Figure 1).

A close up of a logo

Description automatically generated

Figure 1. Map of Falling Creek Reservoir and Beaverdam Reservoir and their watersheds

*Historical and sensor dataset*

An extensive, routine monitoring dataset of water quality in FCR has been collected since 2013 in collaboration with the WVWA and Virginia Tech. This dataset includes meteorological, physical, chemical, and biological data collected both at the deep hole of the reservoir and at the major inflow to FCR (add lots of lab citations here). The inflow dataset also includes discharge to the reservoir measured every 15 minutes at a weir installed at the stream site. More recently, as part of the Smart and Connected Communities (SCC) project, FCR has been outfitted with numerous high-frequency sensors to capture real-time changes in water quality. These data are streamed wirelessly to an online server and appended to the ongoing datafile. Sensor data includes meteorological, physical, chemical, and biological data.

*Model Development (Question 1)*

We will develop both an empirical and a process-based model to forecast chlorophyll-a in the surface water (1.0 m) at FCR during the summer stratified period (May-October) when phytoplankton populations are at their highest. The training period for both models will be 2013-2016 during which we have regularly, weekly coverage of both response and driver data. Both models will be validated using data from 2018.

Empirical: ARIMA

We chose an autoregressive integrated moving average (ARIMA) generalized linear model for our empirical approach to forecasting chlorophyll-a in FCR. The appropriate time step to be included as the autoregressive term in ARIMA will be determined by selecting the previous measurement of chlorophyll-a with the highest Pearson’s r correlation coefficient with the current measurement of chlorophyll-a. From a pool of 53 meteorological, physical, chemical, and biological driver variables, we will select only driver variables which are predictable in nature (e.g., meteorological and inflow variables). From within the pool of predictable driver variables, we will further select only variables which are not correlated with each other through the use of a Pearson’s correlation analysis. Using these variables, we will develop all possible ARIMA combinations with the selected driver variables, and the best model will be determined by AICc (Akaike’s Information Criterion) and parsimony.

Process-based: GLM

We used the General Lake Model, Aquatic Ecodynamics (GLM-AED) as our process-based model. GLM-AED is a one-dimensional hydrodynamic model, etc.

Model Assessment

Model performance will be assessed using a number of metrics currently being used in the ecological forecasting field. Because there is no one metric which can determine that one model performs better than another, a holistic approach examining several model performance metrics will be used instead.

Comparisons of observed and model descriptors will include the mean, standard deviation, quantiles (Dietze book), and kurtosis (Bennett et al 2013). Forecasts are said to be in consensus if they are insignificantly different from the sample mean (Gregory et al ., 2001)

Model outputs will also be compared with each other using the following performance metrics:

* RMSE
* R2 (variation from the 1:1 line)
* Bias
  + SD(model)/SD(data)
  + Base R bias calc mean(observed – model)
* Slope of the regression btw model and OBS
* Pearson’s correlation between the model and OBS
* Area under the curve (Araujo, Thuiller, and Pearson 2006)

Lastly, model performance will be examined through visual assessment of the following figures (Dietze book):

* Observed and model predictions over time
* Predicted vs. observed + 95% confidence interval

*Forecasting Framework: FLARE (Question 2)*

Both models will be integrated into an existing forecasting framework, Forecasting Lake and Reservoir Ecosystems (FLARE), to produce iterative near-term 16-day forecasts of chlorophyll-a. FLARE is designed to pull 2-week weather predictions from the National Oceanic and Atmospheric Administration (NOAA) to force both ARIMA and GLM-AED. Additionally, ARIMA’s autoregressive term will be forced using our established sensor-cloud network to provide yesterday’s chlorophyll-a concentration. Lastly, inflow discharge data for both models will be estimated using an autoregressive linear model of discharge based on the previous day’s mean discharge at the major inflow to the reservoir. Driver uncertainty will be calculated for both ARIMA and GLM-AED forecasts by propagating the uncertainty associated with the NOAA weather forecasts. Parameter uncertainties for both ARIMA and GLM-AED will be calculated by sampling from a distribution of parameters within each model. Forecast effectiveness will be assessed using a suite of performance metrics calculated by comparing the forecast model outputs to the observed high-frequency chlorophyll-a sensor in FCR. Assessment metrics will be calculated for various conditions (summer stratified period, fall mixed period, winter ice period, and following storm events) to determine which approach is most robust under different conditions.

*Integrated Model Averaging (Question 3)*

In order to develop the most informative forecast of chlorophyll-a, we will develop a weighted model including both empirical and process-based model outputs average to incorporate the strengths of multiple models. This will allow the forecast to capture as much variability as possible in chlorophyll-a if for example, one model does a better job of prediction chlorophyll-a at lower concentrations, while another model is more effective at capturing large bloom events. Particularly for low probability yet high impact events such as large blooms in phytoplankton, preliminary results show that a third model may be necessary to capture these peaks in phytoplankton concentration that may operate under different mechanisms. Incorporating multiple models will enable us to develop an early-warning system which will alert users when conditions that indicate a peak in phytoplankton is likely to occur.

**Preliminary Results**

Hindcasts of chlorophyll-a

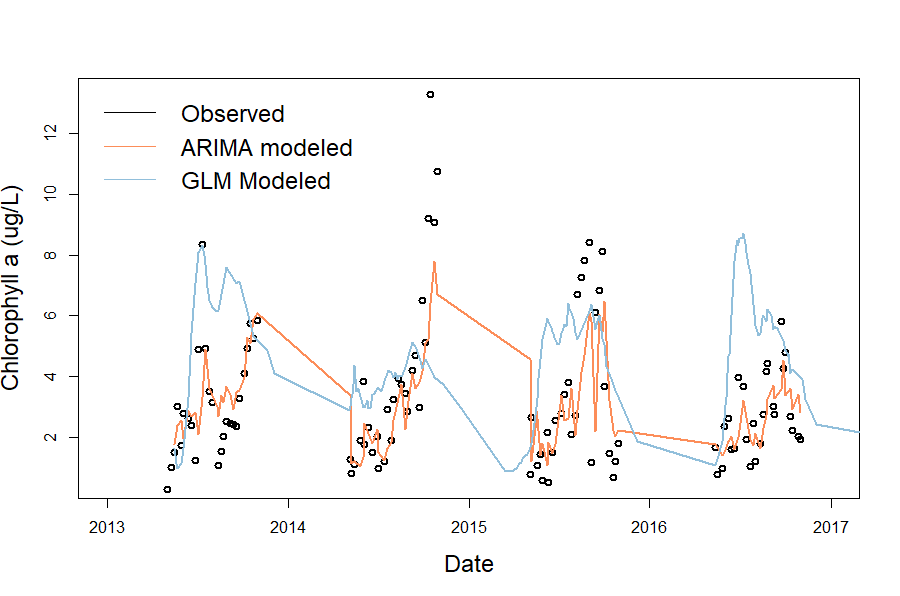
Using the summer period of 2013-2017 as our training period, we developed and calibrated both an empirical, ARIMA, and process-based model, GLM-AED. Our best-fitting ARIMA model over 2013-2016 included discharge to the reservoir and shortwave radiation. ARIMA hindcasted chlorophyll-a over 2013-2016 with an R2 = 0.44 and RMSE = 1.71 ug/L. The ARIMA model was able to successfully capture fluctuations at lower chlorophyll-a concentrations (<10 ug/L) (Figure 2). However, when chlorophyll-a reached values above ~10 ug/L, the model the model was unable to recreate these observed dynamics. GLM-AED was calibrated over 2013-2016 and hindcasted chlorophyll-a with R2 = XXX and RMSE = YYYY. While GLM-AED does not capture as much detail in chlorophyll-a dynamics, it does capture some large peaks in chlorophyll-a that ARIMA does not. 

Figure 2. Observed and modeled chlorophyll-a data over 2013-2016.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 2. Proposed timeline for Chapter 1** *Black boxes indicate completed tasks, gray boxes indicate proposed completion* | | | | | |
|  | Fall 2018 | Spring 2019 | Summer 2019 | Fall 2019 | Spring 2020 |
| Develop ARIMA GLM-AED |  |  |  |  |  |
| Calibrate GLM-AED |  |  |  |  |  |
| Produce hindcasts |  |  |  |  |  |
| Validate model (Q1) |  |  |  |  |  |
| Adapt FLARE code for ARIMA (Q2) |  |  |  |  |  |
| Create integrated model average (Q3) |  |  |  |  |  |
| Run forecasts in FLARE |  |  |  |  |  |
| Submit manuscript, *target journal: Ecological Applications* |  |  |  |  |  |