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
   1. Changing climate
   2. Land use shifts
   3. Increasing human population
3. Understanding systems and anticipating their 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 economic loss

**Proposed Research**

*Chapter 1: Developing near-term forecasts of phytoplankton in drinking water reservoirs***Background/Introduction/Objectives**

* Re-iterate need for forecasting
* Why is ecological forecasting developing now?
  + Data availability
  + Changing ecosystems
* What is forecastable
  + “Forecastable ecosystem attributes are ones for which uncertainty can be reduced to the point where a forecast reports a useful amount of information.” Clark et al 2001
    - Too much uncertainty means not easily forecastable
    - “strong nonlinearity and stochasticity” clark
* Current state of ecological forecasting: mini-lit review
  + Estes et al 2013: empirical does better than mechanistic at predicting observed productivity and suitability of crops in south Africa
  + Lindegren 2013: process-based models to forecast Baltic cod
  + Coulibaly 2000: uses empirical (artificial neural network) method to forecast reservoir inflows—not exactly an “ecological” forecast, more of a physical one
  + Fenocci et al 2019: use process-based (GLM) to hindcast +validate phytos in large lake
    - Table 1 includes more studies
  + Thuiller et al 2004: uses GAM for projections of species distributions
  + Araujo, Thuiller, and Pearson 2006: uses multiple empirical model approaches to develop species distributions (70% of data for calibration, 30% for validation) and then develops projections using climate scenarios
  + White and Nemani 2004: mechanistic model framework to forecast soil ecology
  + Araujo et al 2005 (Reducing uncertainty…): uses multiple empirical model approaches to model species ranges and compares to observed and makes projections
  + Araujo et al 2005 (validation of speces-climate impact modesl…): discusses different validation techniques for empirical models on species distribution projections
  + Martinez-Meyer et al 2004: empirical models for current species niches and validation on past data
  + Case studies in Dietze???
  + Maybe before getting into approaches, start off with the state of where we are in the forecasting community- could you skim a few different forecasting studies and see what the researchers are using- empirical statistical models vs process-based models? It would be super cool if you could add in a mini-lit review of “true” forecasting studies and be able to say something like, “85% of all (n=20) forecasting studies are using process-based models; 100% of freshwater studies (n=8) are using process-based models”
* Forecasting approaches
  + Table 3 of Nilsson et al 2010 discusses different approaches
  + Ratstetter 2017
  + Empirical
  + Process-based
* Weighted model averaging as a way to incorporate strengths of multiple models
* FLARE framework
* Objectives of my work
  + Question 1: How effective is an empirical and a process-based model at hindcasting and forecasting near-term chlorophyll-a?

Lakes and reservoirs provide a suite of critical ecosystem services, including recreation, food production, and drinking water1. Despite the importance of these services, fresh waters are increasingly threatened by rapidly changing land use and climate2, two factors that have led to the contamination of drinking water for millions globally by severe harmful algal blooms (HABs). Consequently, there is a pressing need to not only understand the current state of our freshwater ecosystems, but to predict how they will respond tomorrow, next week, and next year. 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.

Because the field of ecological forecasting is relatively new, there is not yet a consensus as to the best approach for predicting water quality3. Ecologists commonly use time series statistical models to predict future water quality based on past and current conditions (e.g., today’s phosphorus load, water temperature)4, while another approach uses process-based models (built from coupled differential equations) such as the General Lake Model (GLM), a one-dimensional hydrodynamic model, to simulate different metrics of water quality5. A simplified time series model enables application in many waterbodies without the intensive monitoring technology or the extensive set of parameters required by GLM, yet GLM may give a more informed forecast with lower uncertainty, allowing managers to make decisions with more confidence. However, it remains unknown as to which of these approaches is most effective for forecasting water quality at both the single and multiple waterbody scale. Further, by developing multiple models to inform forecasts of phytoplankton, we can draw from an ensemble of possible outcomes to produce a suite of forecasts with quantified uncertainty.

I propose to build on my past research studying the drivers of historical water quality to develop new methods for predicting future water quality. My project will **(1) develop hindcasts of phytoplankton** over 4-5 years using an ensemble model approachin a monitored drinking water reservoir, **(2)** **produce near-term 16-day forecasts of phytoplankton** using the developed models**,** and **(3) determine the most effective method for developing forecasts** in a drinking water reservoir. My graduate research will thus**: 1) inform 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, and 2) provide essential information to the research community that will guide ecological forecasting applications across many different ecosystems.**

**Methods**

* **Study Site**
  + My research will primarily focus on Falling Creek Reservoir (FCR), a small, eutrophic drinking water reservoir in Vinton, Virginia. This reservoir has been extensively monitored in collaboration with the water utility that owns the reservoir, the Western Virginia Watershed Authority (WVWA) and Dr. Cayelan Carey’s lab at Virginia Tech (VT). The managers at WVWA have allowed Dr. Carey’s lab to participate in active interventions and adaptive monitoring in response to changes in water quality since 2013. As a graduate student in Dr. Carey’s lab, I have access to these long-term term data as well as new data streams as part of a recently-implemented interdisciplinary Smart and Connected Communities project (SCC; smartreservoir.org) that is deploying multiple next-generation sensors to capture real-time changes in water quality in FCR. The extensive monitoring dataset, new SCC project, and mutually beneficial working relationship between WVWA and VT make this reservoir an ideal setting for my research.
* **Objective I**
  + **Develop hindcasts of phytoplankton over 4-5 years using an ensemble model approach** I will use the historical meteorological, hydrological, biological, and chemical datasets to develop two models: an autoregressive integrated moving average (ARIMA) time series model and a calibration of GLM to assess the drivers of chlorophyll-a and HABs in FCR. Models will be developed over a training period covering 2013-2016 for ARIMA and 2013-2017 for GLM. Explain why different training periods, and give a bit more detail as to how modeling will happen? I would add in three sections here: 1) available data description; 2) ARIMA details and 3) GLM details. To develop the ARIMA, we will select variables which are predictable from the entire pool of available data and create model iterations containing all of the variables. The best model will be selected based on Akaike’s Information Criterion (AICc) and parsimony, choosing the simplest model. What is global pool of drivers? How were these chosen? GLM will be calibrated over the 2013-2017 time period. Both models will be validated using data from 2018, which has been withheld from the training dataset. Model comparisons will be made using R2 and root mean squared error (RMSE) to assess model fit. Of what variables? Why?
* **Objective II: Produce near-term 16-day forecasts of phytoplankton.** 
  + Plugging into the FLARE framework
    - Use near-term NOAA forecasts + sensor-cloud networks to pull current data and forecasts of driver variables to go into ARIMA and GLM models
  + For both models, I will sample from a distribution of parameters using Bayesian techniques to generate uncertainties for each forecast
  + These forecasts will be direct deliverables to WVWA
* **Objective III: Determine the most effective method for developing forecasts.**
* The forecast model outputs will then be compared to the observed high-frequency chlorophyll-a sensor data to determine which approach is most robust under different conditions.
* Model fit will be assess using R2 and RMSE
* If model strengths between ARIMA and GLM are distinct, a weighted model average may be developed to capture the relative strengths of each model and cover as much of the variability in chlorophyll-a as possible
* TEXT TO BE INCORPORATED INTO THIS SECTION: Once both models are calibrated, I will assess how effective each model is at predicting near-term future states of water quality. For GLM, I will use 2-week weather predictions to run the calibrated model to predict chlorophyll-a. For the time series model, if the model indicates, for example, that the previous month’s P load is the most important driver of current chlorophyll-a, I will feed the model with current P loads to predict future chlorophyll-a. For both models, I will sample from a distribution of parameters using Bayesian techniques to generate uncertainties for each forecast. The forecast model outputs will then be compared to the observed high-frequency chlorophyll-a sensor data to determine which approach is most robust under different conditions. I predict that output from the time series model will have a larger uncertainty than the GLM model, given that GLM is a process-based model driven by coupled equations, rather than statistical relationships alone.

**Preliminary Results**

* Objective I: hindcasts of phytoplankton
  + ARIMA
    - Autoregressive model includes chlorophyll-a at the previous timestep, discharge to the reservoir, and shortwave radiation
    - Over 2013-2016, R2 = 0.44, RMSE = 1.71 ug/L
    - Captures quite detailed dynamics at lower chlorophyll-a concentrations (<10 ug/L—in CTD units, maybe need to scale this for general comparison if most instruments read higher and that’s the magnitude people are used to thinking about chlorophyll?)
    - Misses some larger peaks
    - 2018 validation: R2 = 0.44, RMSE = 1.02 ug/L
  + GLM
    - Calibration still in progress
    - Over 2013-2017, R2 = 0.001, RMSE = 3.42 ug/L
      * This includes winter months which ARIMA does not model
    - Does a better job at capture larger peaks in chlorophyll-a
    - 2018 validation: in progress
* Objective II
  + Plug both models into the FLARE framework
  + Fully quantified uncertainties for both ARIMA and GLM
* Objective III
  + Weighted model averaging to incorporate strengths of both models (e.g., detailed changes in concentration and peaks)

Chapter 2 Assessing the dynamics of stream-reservoir linkages across reservoirs

*Question 1: How do in-stream nutrient dynamics influence within-reservoir ecosystem response?*

*Question 2: How do major storm events influence these dynamics?*

***Question 4: How does within-reservoir nutrient processing change along the reservoir flow gradient?***

***Question 3: How do stream dynamics differ between neighboring watersheds?***

**Background**

* Importance of watershed characteristics on waterbody condition
  + Terrestrial-aquatic linkages
  + Stream-lake linkages
    - Further proposed reading: Wurtsbaugh et al 2005, Robinson et al 2007, Goodman et al 2011, Xu & Xu 2018, Stachelek & Soranno 2019, Marcarelli & Wurtsbaugh 2009, Sadro et al 2012, Schmadel et al 2018, Jones 2010, Hotchkiss et al 2018., ETC
* Importance of storm events in stream dynamics
* In the new age of data availability, can we relate stream dynamics of a highly monitored stream to neighboring streams lacking data availability?

**Methods**

* Study site
  + Introduce BVR
    - 4 selected inflows (see Figure 1)
    - Relationship to FCR
      * Watershed history
      * Geographical similarity
      * Outflow to FCR
* Field sampling, summer 2019
  + Weekly (bi-monthly?) sampling of 4 major inflows to develop baseline
    - Nutrient chemistry (total and soluble nitrogen, phosphorus, and carbon)
    - Discharge (using a flowmeter)
    - Physical characteristics (dissolved oxygen, conductivity, temperature)
  + Periodic event-based sampling
    - ISCO sampler to capture high-frequency dynamics during major flow events
  + Weekly sampling at deep-hole of BVR
    - Biological, physical, and chemical sampling
  + Grab samples at strategically chosen sites along the reservoir gradient (?) (see red arrows on Figure 1)
    - CTD + nutrient grab samples

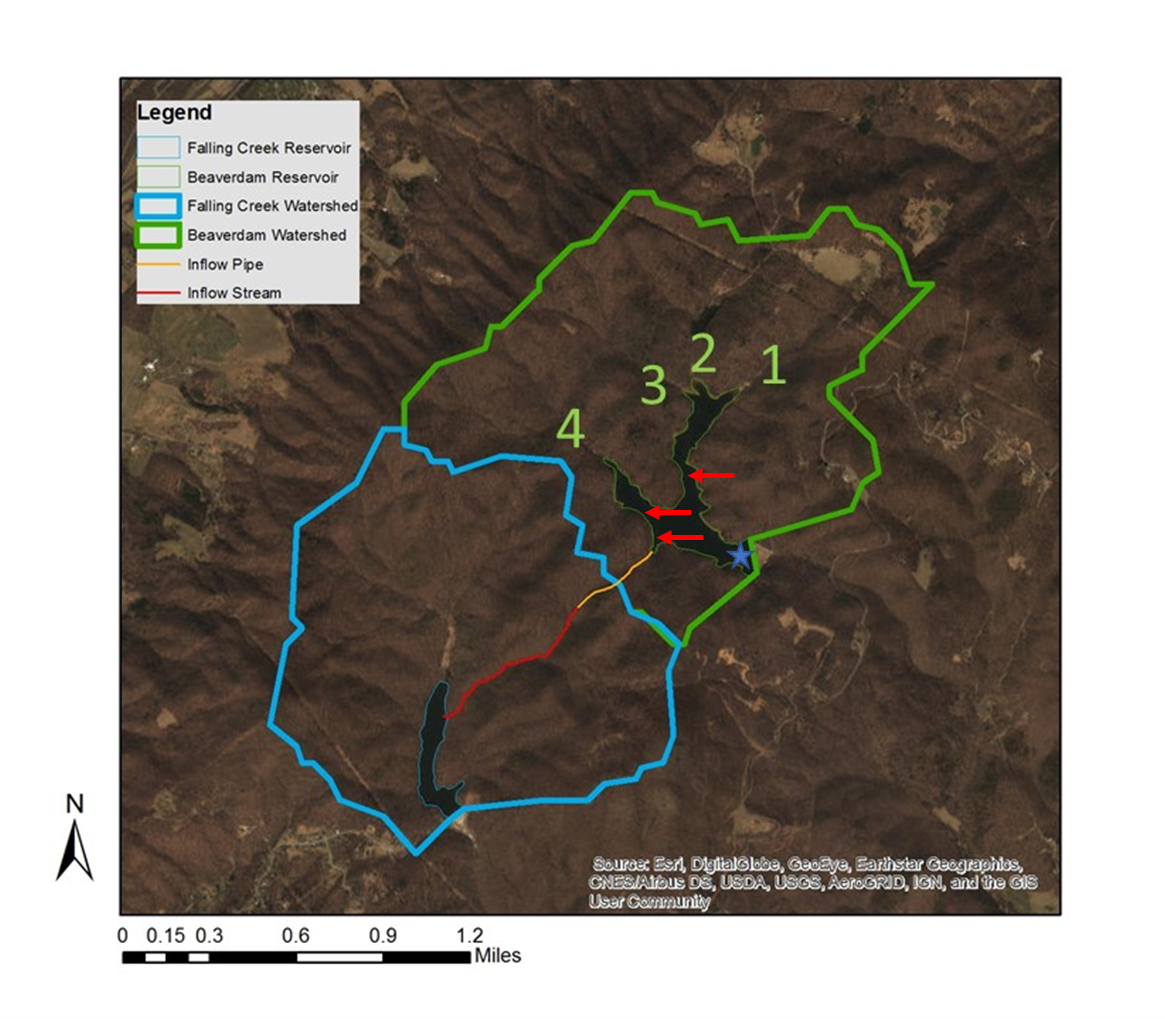


Figure 1. Map of Falling Creek Reservoir and Beaverdam Reservoir, along with watershed boundaries and proposed sampling sites

* Data analysis
  + Generalized linear models to analyze drivers of nutrient and phytoplankton dynamics along a stream-reservoir gradient