New discoveries in the GLEON-PRAGMA Lake Expedition: modeling advances and scaling up

Paul Hanson (University of Wisconsin)

Cayelan Carey (Virginia Tech)

Renato Figueiredo, Ken Subratie, Saumitra Aditya
(University of Florida)

Goals for supplemental proposal:

1. Make available and usable new services, software and approaches to freshwater scientists
2. Expand the overlay technology beyond the existing PRAGMA-GLEON expedition
3. Grow the next generation of users by training students how to use GLM-AED to study lake water quality with hands-on modules
4. Assess the efficacy of integrated technology-science-education development in the classroom and in GLEON working groups

Our approaches to scaling

1. GRAPLEr team
2. Products
3. Technology
4. Science

Our Team is Growing

Additional Products

- Subratie KC, Aditya S, Mahesula S. Figueiredo R, Carey CC, Hanson PC. in review. GRAPLEr: A distributed collaborative environment for lake ecosystem modeling that integrates overlay networks, high-throughput computing, and web services
- Carey CC, Gougis RD, et al. In prep. Your brain on GRAPLEr
- Ruan G, Hanson PC, Plale B. In prep. Lake Time Series Mining with Symbolic Representation
- Hanson PC, Carey CC, et al. In prep. A tail of two distributions
- Snortheim CA, Hanson PC, McMahon KD, Read JS, Carey CC, Dugan HA. In prep. Meteorological drivers of hypolimnetic anoxia in a eutrophic, north temperate lake
- GRAPLEr software
- Project EDDIE modules

Technology

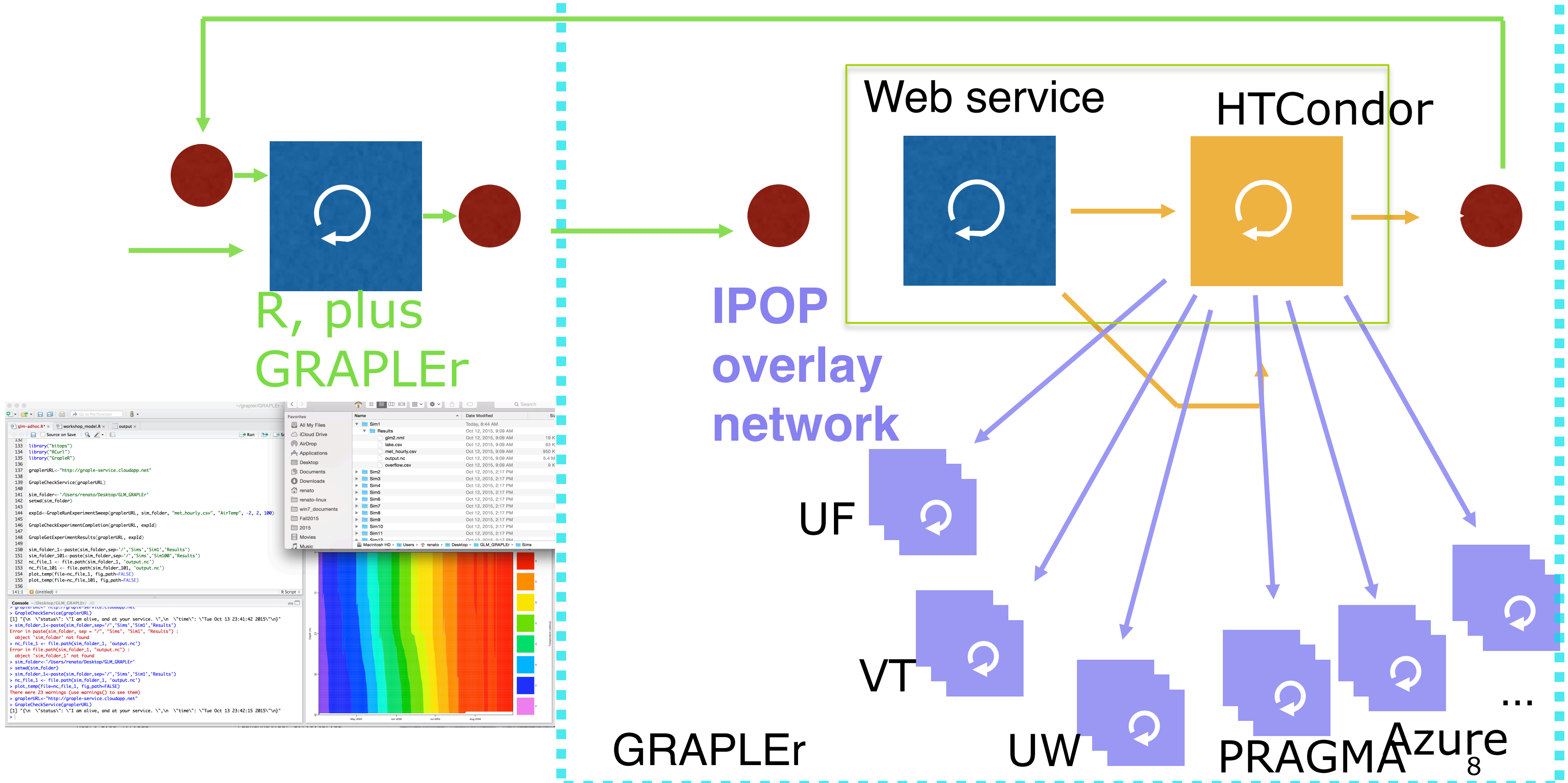
- Distributed resources are available
  - Desktops, servers, cloud; how to aggregate in user-friendly manner, with low administrative overhead?
- Local, familiar computer environments
  - Windows/MacOS, GLM, R
  - Present user's preferred environment for ease of use
- Technical approach
  - IPOP: overlay automatically creates virtual network; only requirement is to run software on endpoints
  - HTCondor: high-throughput computing job manager
  - GRAPLEr: Web interface conveniently accessible from a user's desktop – R

IP-over-P2P (IPOP) Overlay

- "User-Defined Networking"
  - SDN that users can deploy
    - User-level; easy to use
  - Software only needed at endpoints
    - Cloud VMs, personal computers, mobile devices
  - Peer-to-peer VPN links
    - No centralized gateway
- Also being used in ENT testbed
  - Collaboration with Kohei Ichikawa
    - Demo earlier today
  - http://ipop-project.org

Technology Updates

- GRAPLEr Web Service
  - Expose simple R desktop interface to users
  - Handle packaging of multiple simulations into Condor job, compression, job submission, post-processing



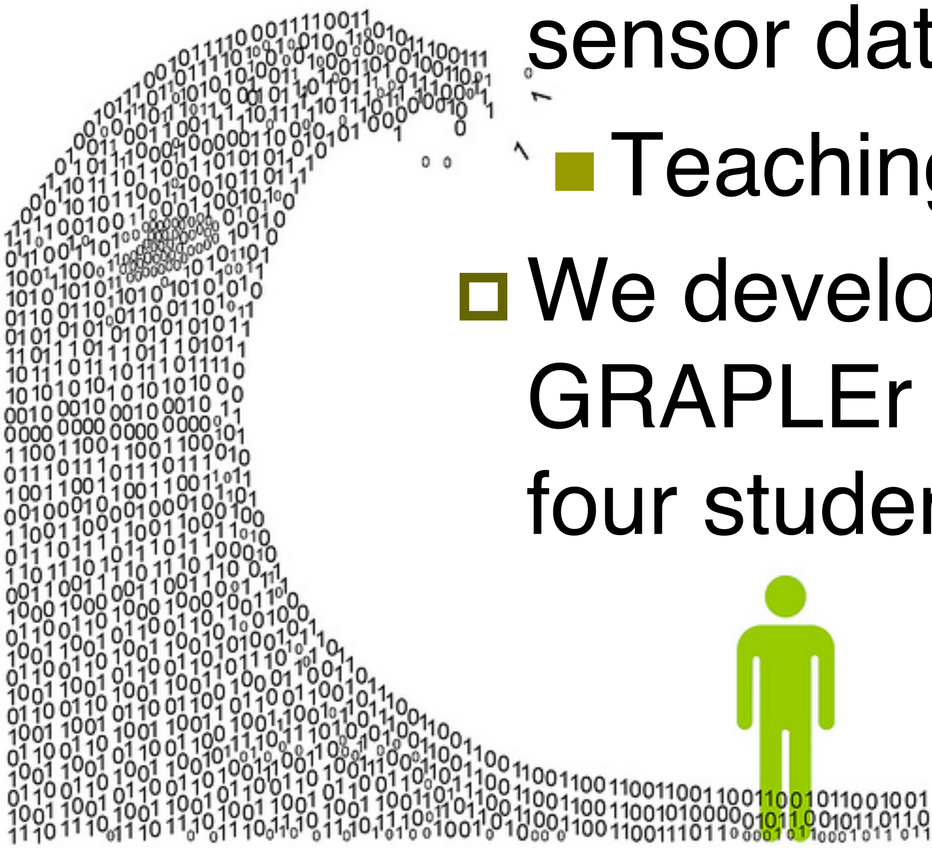
Demonstration

Meta-Goals

- Exploit lake sensor network observations (GLEON) with overlay network technology (PRAGMA) to expand our modeling capacity =(GRAPLE)
- **GRAPLE** = **G**LEON **R**esearch **A**nd **P**RAGMA **L**ake **E**xpedition: integrating disparate technology, knowledge, and researchers to advance lake science
- <u>Our ultimate long-term goal</u>: expand the lake expedition to include cohorts of graduate and undergraduate students through targeted training in both computer science and limnology

Integrating the GRAPLEr into classrooms

- How best can we prepare students to understand and use computer modeling and big sensor datasets?
  - Teaching modules!
- We developed and tested new GRAPLEr teaching materials in four student workshops/classes, n=75 students total



Modules
- Ice Phenology

Meetings

Research

About
- Goals
- Data Sets
- Implementation
- Evaluation

Contact

Forms

Project EDDIE Home

CeMaST Home

Project EDDIE: Ice Phenology Module

Project Summary

Lakes are changing worldwide due to altered climate. Many lakes that were historically frozen in the winter are now experiencing fewer days of ice cover and earlier ice-off dates. In this module, students will explore long-term ice-off datasets from several lakes and use linear regression to make predictions about ice-off dates in the future. Project EDDIE modules are designed with an A-B-C structure to make them flexible and adaptable to a range of student levels and course structures.

Acknowledgment

We recognize that many users might wish to modify this activity. Please include text in each file that you use that acknowledges the original development of this module. We suggest this text, which is currently included in each file:

> *"This module was initially developed by Carey, C.C., J.L. Klug, and D.C. Richardson. 1 April 2015. Project EDDIE: Lake Ice Phenology. Project EDDIE Module 1, Version 1. http://cemast.illinoisstate.edu/data-for-students/modules/ice-phenology.shtml. Module development was supported by NSF DEB 1245707."*

Citation

Please cite this module as: "Carey, C.C., J.L. Klug, and D.C. Richardson. 1 April 2015. Project EDDIE: Lake Ice Phenology. Project EDDIE Module 1, Version 1. http://cemast.illinoisstate.edu/data-for-students/modules/ice-phenology.shtml."

Class Time

This entire module can be completed in one 2-3 hour lab period or two 50 minute lecture periods for introductory or intermediate level students. Activities A and B could be completed with upper level students in one 50-60 minute lecture period. Students will need 1-2 hours outside of class to prepare for the exercise and complete the homework activities.

Files

- Student Handout [docx]
- Student Datasets [xls]
- Instructor's Manual [docx]
- Instructor's PowerPoint [pptx]
- Instructor's Homework Part C Answer Key [xlsx]

Project EDDIE is supported by NSF DEB 1245707.

projectEDDIE.org

article

A Model for Using Environmental Data-Driven Inquiry and Exploration to Teach Limnology to Undergraduates

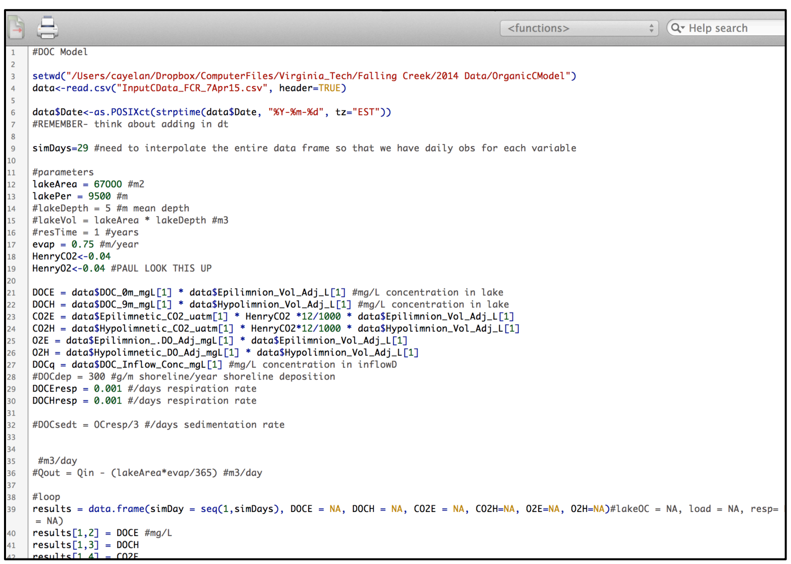
Cayelan C. Carey, Rebekka Darner Gougis, Jennifer L. Klug, Catherine M. O'Reilly, David C. Richardson

Limnologists are increasingly using large volumes of data, both from high-frequency sensors as well as long-term studies, to address new research questions. Undergraduate students, i.e., future limnologists and informed citizens, need quantitative reasoning skills and tools to be able to analyze these large datasets. However, most undergraduate curricula typically remains focused on small-scale local studies, potentially contributing to many students' inability to see the applicability of their classroom experiences (Prokop et al. 2007). In response, we have developed undergraduate teaching modules that integrate the use of high-frequency and long-term datasets from many lakes around the world. Here, we describe two modules that are designed to committed to improving environmental data literacy in undergraduate classrooms as part of the Environmental Data-Driven Inquiry and Exploration Project (Project EDDIE; http://www.projecteddie.org). In addition to describing the modules, we also share both the students' and instructors' experiences during module implementation, and highlight the potential for scaling these modules across different skill levels, both within and across different types of institutions. Our experience suggests that students appreciate the value of high-resolution and long-term data, and that working with large datasets cements the "real world" application of basic freshwater ecology concepts.

The emerging approach of using large and Langen et al. 2014). In response to this challenge, we have developed sensor-based and time series data analysis activities that can be integrated into undergraduate classrooms to improve quantitative skills and reasoning and increase student engagement. Each exercise has a modular "A-B-C" structure with three student activities that build from relatively simple to more complex (Fig. 2). The full ABC module allows students to complete a learning cycle involving data exploration, explanation, and extension into a new situation (Bybee et al. 2006). The flexible format of the module enables instructors to choose the activities most appropriate for their classroom, as some activities of the module can be completed in a standard one-hour lecture

Developing a GRAPLE module

- We have written simple GRAPLE R scripts to use distributed computing to run the GLM lake model for classroom use
- Students create climate scenarios and then run 1000s of model simulations to study their effects on the lake *in silico*



5 E's pedagogical framework

Pedagogical connections:

Phase	Functions	Examples from this module
Engagement	Introduce topic, gauge students' preconceptions, call up students' schemata	Pre-class reading, short introductory lecture
Exploration	Engage students in inquiry, scientific discourse, evidence-based reasoning	Development of new climate scenarios and hypotheses of how altered weather affects lakes; testing of these hypotheses with lake models
Explanation	Engage students in scientific discourse, evidence-based reasoning	In-class discussion of the effects of the different climate scenarios
Expansion	Broaden students' schemata to account for more observations	Using the GRAPLEr software to create hundreds of different climate scenarios
Evaluation	Assess students' understanding, formatively and summatively	In-class discussion of how small changes in weather variables can affect lake thermal structure

From Bybee et al. 2006

MODELING CLIMATE CHANGE EFFECTS ON LAKES USING DISTRIBUTED COMPUTING MODULE
Student Instructions

This module was initially developed by Carey, C.C., S. Aditya, K. Subratie, and R. Figueiredo. 20 September 2015. Project EDDIE: Modeling Climate Change Effects on Lakes Using Distributed Computing. Project EDDIE Module 4, Version 1. http://cemast.illinoisstate.edu/data-for-students/modules/lake-modeling.shtml. Module development was supported by NSF DEB 1245707 and ACI 1234983.

Learning Objectives:

1) Set up and run the General Lake Model (GLM) in the R statistical environment to simulate lake thermal structure.
2) Understand the structure and function of GLM configuration files, driver data, and output files.
3) Modify the input meteorological data for one GLM model to simulate the effects of different climate scenarios on lake thermal structure.
4) Interpret model output from GLM simulations to understand how changing climate will alter lake thermal characteristics.
5) Use the GRAPLEr R package to set up hundreds of model simulations with varying input meteorological data, and run those simulations using distributed computing.
6) Explore the application of distributed computing for modeling climate change effects on lakes.

Why this matters:

Lakes around the globe are experiencing the effects of climate change. Because it is difficult to predict how lakes will respond to the many different aspects of climate change (e.g., altered temperature, precipitation, wind, etc.), many researchers are using models to manipulate weather scenarios and simulate lake responses. Lake models provide a powerful tool for exploring the sensitivity of lake thermal structure characteristics to weather. In this module, you will learn how to set up a lake model and "force" the model with climate scenarios of your own design to

LakeModelingModule_29Sep15

```r
#Modeling Climate Change Effects on Lakes Using Distributed Computing Module
#This module was initially developed by Carey, C.C., S. Aditya, K. Subratie, and R. Figueiredo. 29 September 2015.
#Project EDDIE: Modeling Climate Change Effects on Lakes Using Distributed Computing. Project EDDIE Module 4, Version 1. http://cemast.illinoisstate.edu/data-for-students/modules/lake-modeling.shtml.
#Module development was supported by NSF DEB 1245707 and ACI 1234983.


#R code for students to work through the module activities A, B, and C.
#This module consists of 6 objectives. Activity A consists of Objectives 1-4, Activity B consists of Objective 5, and Activity C consists of Objective 6.
#Ths script was modified last by CCC on 29 September 2015

######################################################################

#ACTIVITY A - OBJECTIVE 1: Download the GLM files and R packages successfully onto your computer.  The example code below is for a Mac operating system, but should work perfectly on a PC- if you're a PC user and having trouble, check the direction of your \ / marks- sometimes this may be different between operating systems.

install.packages('sp') #NOTE: you'll get output that says "There is a binary version available but the source version is later... Do you want to install from sources the package which needs compilation? y/n" Type 'y' (without the quotes) and hit enter. This should now successfully load- when it's done, it should say 'DONE(sp)' if it worked successfully.

install.packages('glmtools', repos=c('http://cran.rstudio.com', 'http://owi.usgs.gov/R')) #you need to be connected to the internet for this step- this step enables you to access the USGS website and download the R packages that allow you to work with GLM in R. Note: if you install this package before Korea, you can skip this step at the GSA workshop.

library(glmtools) #load the two packages that you need to run GLM and manipulate its output
#note: you may get lots of output messages at this step- if this worked successfully, you should read: "This information is preliminary or provisional and is subject to revision. It is being provided to meet the need for timely best science. The information has not received final approval by the U.S. Geological Survey (USGS) and is provided on the condition that neither the USGS nor the U.S. Government shall be held liable for any damages resulting from the authorized or unauthorized use of the information. Although this software program has been used by the USGS, no warranty, expressed or implied, is made by the USGS or the U.S. Government as to the accuracy and functioning of the program and related program material nor shall the fact of distribution constitute any such warranty, and no responsibility is assumed by the USGS in connection therewith".
```

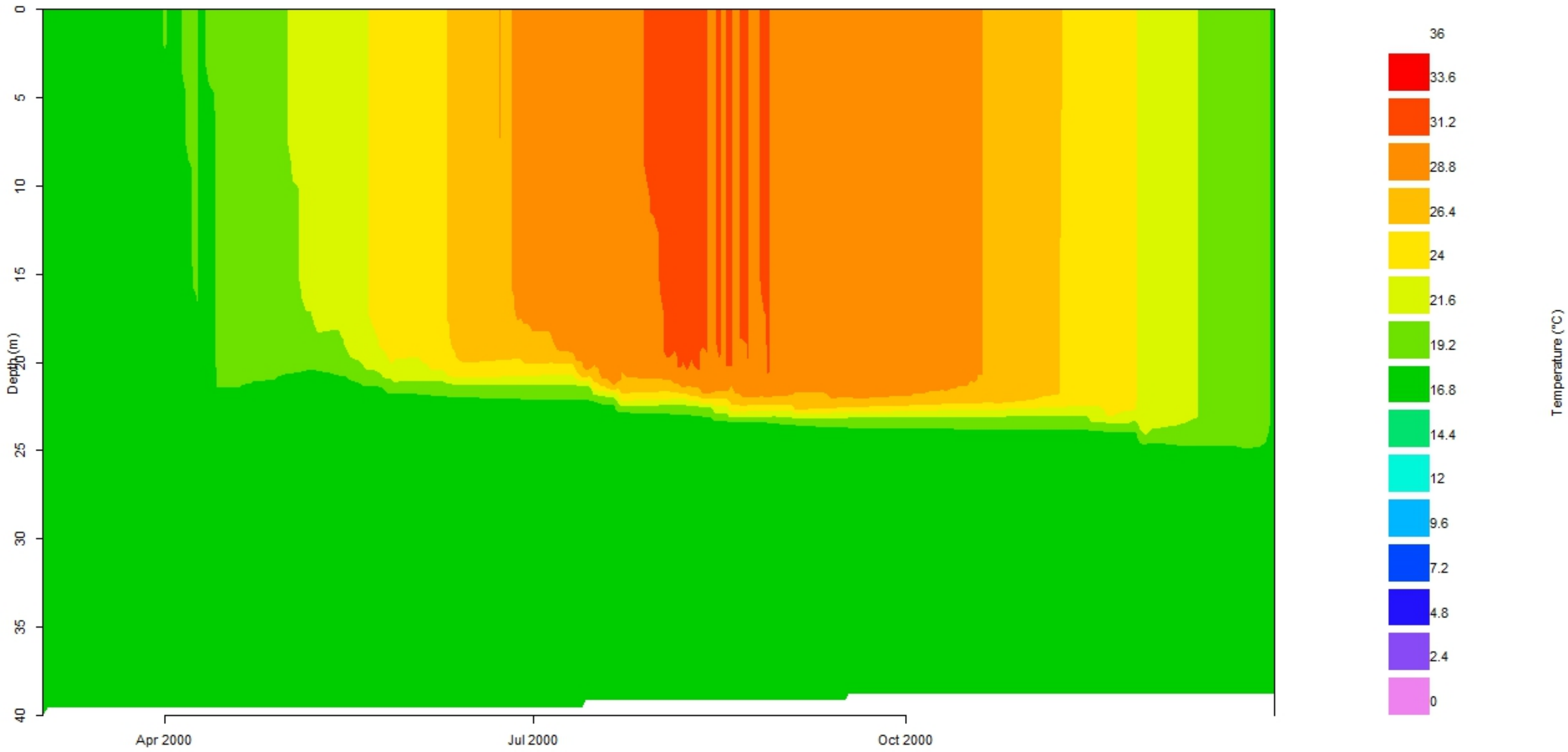
Overview of module

- Short PowerPoint overview of lake modeling and some background on the tools we will be working with
- Activity A: get a lake model to run!
- Activity B: develop one climate scenario, and see how your lake responds!
- Activity C: use distributed computing and overlay networks to run 1000s of lake model simulations!

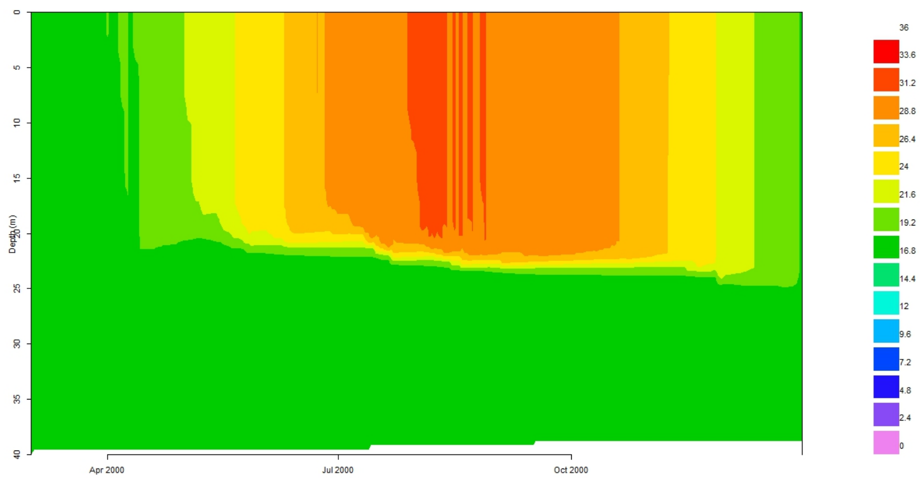
Learning objectives

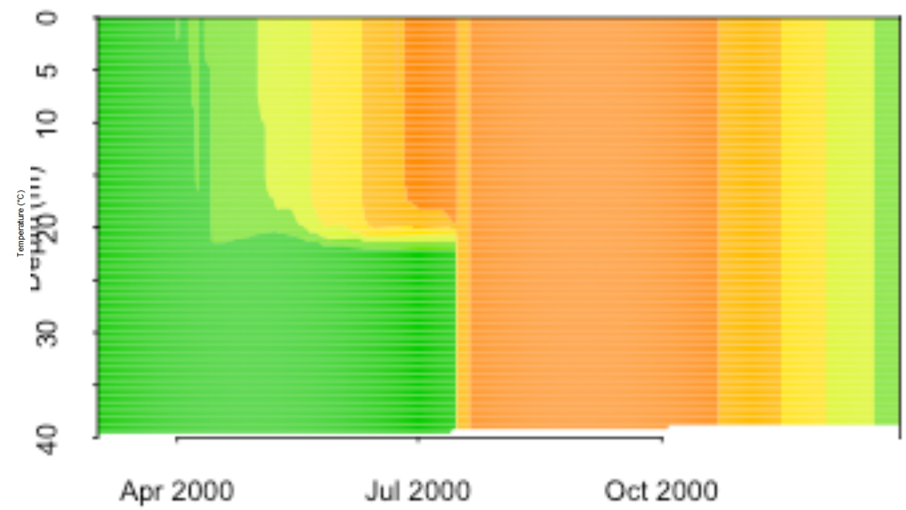
- Students are challenged to:
  - Develop hypotheses about climate change and test them with computer models
  - Learn about distributed computing, IPOP, and overlay networks
  - Integrate technology tools to answer ecological questions
  - Present their findings to classmates

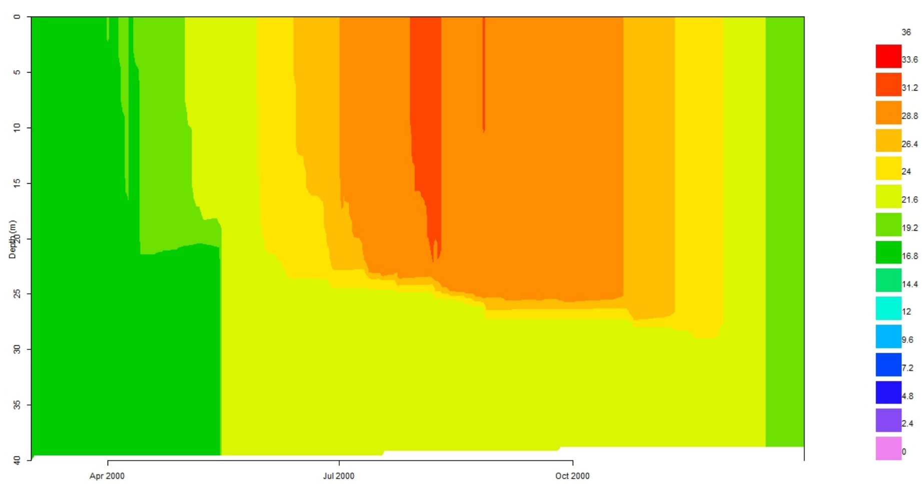
Baseline scenario

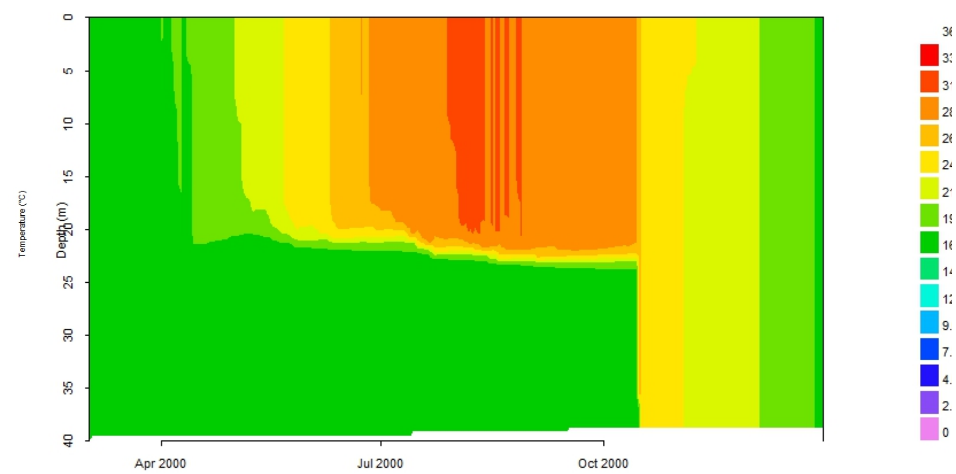


Manipulating climate *in silico* as a teaching tool across disciplines









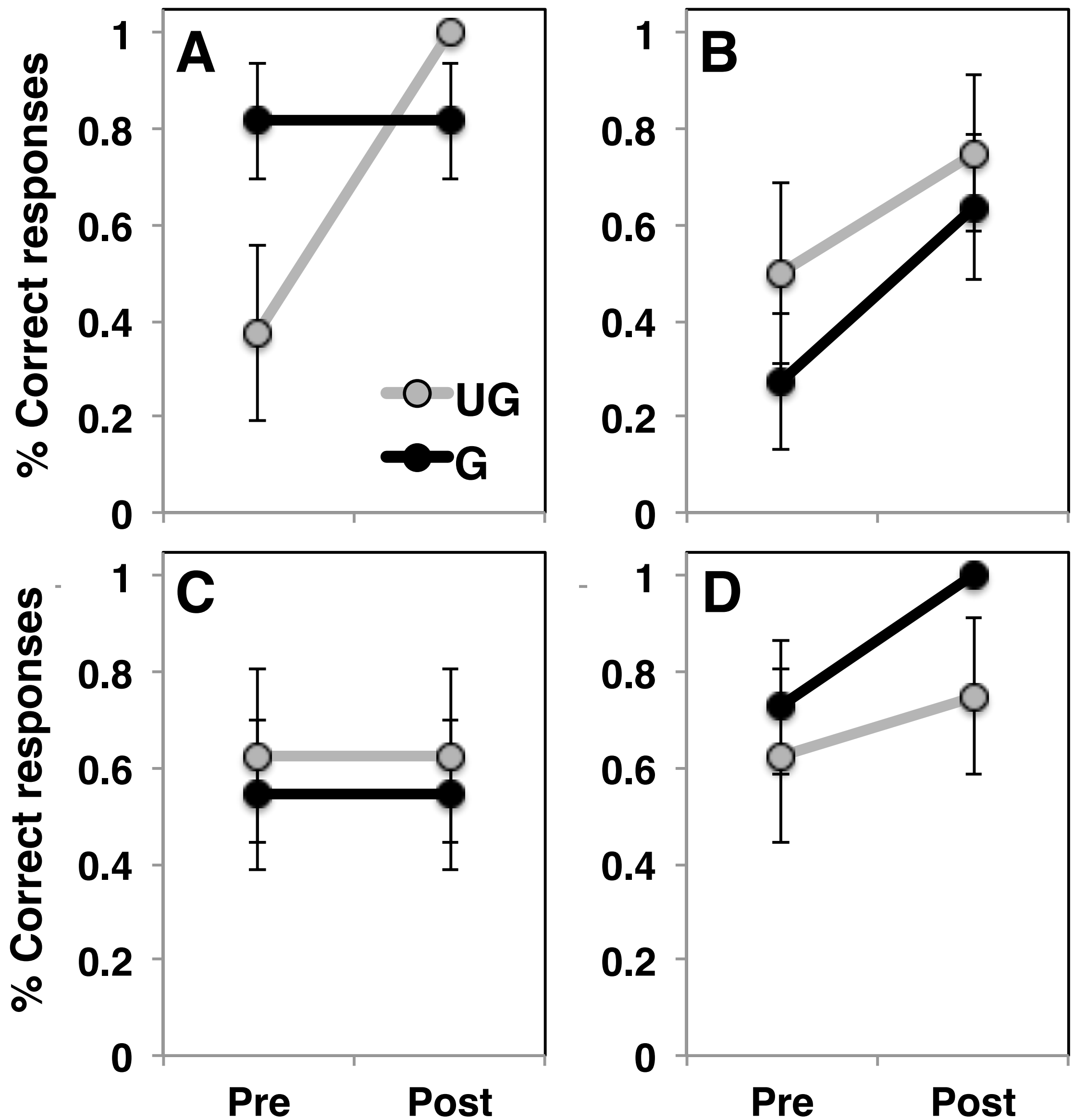
Next steps

- Three cohorts of students are being formally assessed to track their responses to using GLM and overlay technology
- This module has been disseminated to 8 universities for testing as part of the U.S. NSF-sponsored Project EDDIE and will be released publicly in spring 2016
- We have partnered with pedagogical experts to study how non-CS students use this technology → will enable broad dissemination
- All GRAPLE materials will be available open-source for instructors and students to access

Initial assessment results:

- Participation in the module significantly increased ecology UGs' experience and comfort level with:
  - CSV files, R, GLM, computer programming, distributing and overlay networks
- **UGs' interest in computer models significantly increased** as a result of the module, as did **grads' interest in using distributed computing & numerical simulation** for their PhD dissertations
- Students' conceptions of science also significantly changed, in terms of how they perceived variation in environmental data, and the value of observations in the scientific process

Improved understanding of ecological concepts: climate change effects on lakes



Q1: How do mixing events alter thermal stratification?

Q2: How will modeled air temperatures affect water temperatures?

Assessment summary

- Integrating the GRAPLEr into classroom has:
  - Increased students’ proficiency in understanding ecological concepts
  - Increased their comfort level, experience, and interest in computer models, distributed computing, and coding
  - Changed the way that they view the scientific method and hypothesis-testing using models

Expanding GRAPLE beyond this initial expedition

- Our work to date has demonstrated that CS-limnology synergies can provide new tools for tackling local to global water issues
- We are poised to lead a new wave of technology-enabled, multi-disciplinary research that will transform the science we can accomplish
  - Make distributed computing accessible to domain scientists by simplifying the steps needed to participate in the overlay network
  - Grow the next generation of users by training students how to use GLM & overlay networks by teaching these scripts in classrooms worldwide

Acknowledgments

- GRAPLE students
  - Craig Snortheim, Jon Doubek, Alex Gerling, Kate Hamre, Ken Subratie, Youna Jung
- IPOP developers and Project EDDIE collaborators
- This material is based upon work supported in part by the National Science Foundation under Grants No. 1234983 and 1339737