June 29, 2020 Joint Methods & CI Working Group Call

Attendees: Quinn Thomas, Chris Jones, Ashton Griffin, Alexey Shiklomanov, Jody Peters, Matt Helmus, Jake Zwart, Debjani Sihi, Yang Song, Mike Dietze, Leah Johnson, Ben Toh, Rob Kooper, Rich Fulford

Regrets: Bruce Wilson, Abigail Lewis

Agenda/Notes:

1. Introductions
    a. Mike Dietze - EFI Director at Boston University. CI is an area where his lab has focused heavily particularly on terrestrial ecosystem models
    b. Quinn - Virginia Tech - RCN lead and NEON Forecasting Challenges
    c. Jody - University of Notre Dame, works with Jason McLachlan, coordinates with Dietze et al on EFI and other projects
    d. Chris - NC State - tooling and docker containers for deployment
    e. Jake - USGS Postdoc Data Science branch. Lake/stream temp modeling.
    f. Leah - Virginia Tech Dept of Stats. How to implement stats methodology. Working on vector borne disease and working on tick RCN NEON challenge and RCN steering committee
    g. Ash - Data Scientist at Mailchimp background in ecology. PhD at U of GA in marine diseases. Looking to get back into ecological work and applying what learned in tech can be brought to ecology and CI/methods & tools
    h. Alexey - NASA. Strong interest in ecological forecasting and CI
    i. Matt - Temple University, starting an ecological forecasting project in invasive species
    j. Debjani - post doc at OakRidge, going to be faculty at Emory!. Apply CI in modeling GHG
    k. Yang - U of Arizona land surface modeling simulating veg and microbiology community
    l. Ben - PhD student at U of Florida. Bayesian stats of ecological forecasting working on predicting malaria and other vector borne diseases
    m. Rob - research lead programmer at NCSA worked with Mike on PEcAn. background on computer science. Employment of docker containers
    n. Rich - Environmental Protection Agency had focused on quantitative fisheries modeling. Recently transitions more to coastal land use. Interest is in creating forecasting tools for land management

2. Updates for EFI Task Views
    a. Use Task View 1 on Reproducible Workflows as a guide
        i. Update from Jake - there was an internal USGS review of the blog post and added a couple of comments. Jody updated on the website and will add a note that it was updated today

b. General notes/discussion about task views - similar to R task view to give information about R packages. We want to use the same format, but not limited to R packages

   i. The plan is to have 4 rounds of information/blog posts. The post will also be saved as stand-a-lone webpages. These Task Views are not meant to be comprehensive but instead provide new users with a high-level overview of each area and what the most commonly used tools are. And to provide people who have been forecasting for a while a list of tools that they may not be aware of

   ii. Key thing to move Task Views forward is having a lead for the Task View

   iii. We could use the forecast example workflow (see below) to help produce the outline/resources to include in the Task Views

      1. For example, the data ingest/cleaning/management - will work well with the NEON forecast challenge since there will be NEON data that will need to be ingested, cleaned, and managed
      2. **Proceed by getting a forecast example up and running and as discuss how to do each task in the forecast example, add those tasks/notes to the Task Views**
      3. By focusing on the forecast example, it will help to narrow the tools listed in the Task Views

c. Uncertainty Quantification & Propagation, Modeling & Stats

   i. Abby, Ben, Alissa, Leah happy to help (Leah can’t lead right now)

d. Data Ingest, Cleaning, Management

   i. Matt H., Jake Zwart happy to help, Chris Jones happy to help.

e. Visualization/Decision Support Tools, User Interface

   i. Abby, Ben, Alissa happy to help, Chris happy to co-lead with people more familiar with R Shiny and tableau

3. RCN Tasks

a. Note from Rob on last call - He talked to the XSEDE director who said there is the ability to leverage XSEDE resources for the challenges proposed.

   i. There is no reason why we cannot ask for allocation for XSEDE and Jetstream. Need to figure out how much space will be needed

      1. See Slack conversation on the #cyberinfrastructure channel

   ii. This also relates to CyVerse discussion.

      1. Quinn, Mike, Jody had a call with CyVerse folks to talk about synergies
      2. CyVerse provides resources that allow us to log on and create jobs, share data, code, and frameworks. If jobs get too big, they can be farmed out to XSEDE or local universities. Everything is on CyVerse, but can be run on individual universities’ HPC resources.
      3. Null models could be on CyVerse, submissions could go to CyVerse. It provides an option for a shared work environment

4. But it also does not need to be the place where all forecasts are run, if participants in the challenge want to use their own resources or other prgorams that they are more familiar with
5. Thinking about the data part of the NEON Forecasting Challenges - FB's latest challenge with Kaggle had the top 2 winners disqualified because they used data that were not available to others. This is something for us to be aware of for the Challenges.
    a. We need to clarify what data is available so as not to run into this. Rob will look for links to articles about this so we avoid this issue
        i. https://syncedreview.com/2020/06/14/facebook-and-kaggle-face-backlash-after-disqualifying-apparent-deepfake-detection-challenge-winner/ Short story, the top two contenders were disqualified by FaceBook because the claim was that "ensure the External Data is available to use by all participants of the competition for purposes of the competition at no cost to the other participants"
6. Jetstream through XSEDE will allow you to create VM to run forecasts
7. There will be a follow-up meeting with CyVerse. If anyone in the CI/Methods group has any questions that we should ask, let us know
    a. How can we set up an interaction with XSEDE resources?
    b. How does it spin up, how does it work with different queuing systems?
    c. How does it work with the data sources?
    d. Can data sit in different systems, will it be smart about where to run (send to Texas/San Diego to run).
    e. How easy is it to bring my own model as a docker container into CyVerse?
    f. Most NEON datasets are not 3 TB, except for Airborne data. Carl testing NEON store package which may be why his datasets were so large. What is the range of the size of data we expect we will be using?

b. Major Modules for the Forecast Challenges - Forecasting Workflow Example
    i. Review Suggestions on the Google Doc [link removed]
    ii. Use this to work out an example workflow. Step back and think about what the major tasks are that go into an ecological forecasting workflow, even with a simple model. Can we set up a repo and a docker container that has the null model that we will use to compare the other NEON forecasting challenge models to?
    iii. Don't want people to reinvent tools

iv. Want to agree on the big boxes and some small boxes as a first step. Then from that point say how do those things get lined up? Is one expected to submit a workflow with all the components? Or what do we pull back to shared resources that no individual team will need to run, because we will run them for everyone. Example - scheduler. Put docker container in that has a single scheduler module in CyVerse and XSEDE Jetstream. Won't be in the container because it tells the container to go. Data availability or time step driven.

v. Some forecasts will have data that shows up daily with reliability, but others like the tick forecast will show up every couple of weeks and you will want to run the forecasts when the data shows up.

vi. Same with data ingest - will need to ingest data for drivers for models. If that is common across groups, have some of those in shared code. E.g., weather forecasts, sub-seasonal climate forecasts

vii. This probably isn't the case for the 1st round of the RCN NEON challenges, but in other forecasts, the inputs and drivers may be scenario-based

viii. Notes when looking specifically at the Workflow Example

ix. Mike split up the forecast into 2 steps.

1. Reforecast step. Forecast running once a week. When you ran it a week ago, you ran it with the weather forecast, when you run it today, you run it with the actual forecast before running the forecast for the next 16 days. Have to jump backward before going forward
2. Data Assimilation step. If not updating parameters or states, don't have to do this, but in traditional DA mode then grab the state of the model, reading the re-start, adjust state of system, write out state again. Hop test is useful as a tool to run the test that verifies that your model does the same thing if you start and stop as if you don't stop it. This is often for the large models. In simple models people forget to write out seed of random number generator so get different results.

x. Other things to consider:

1. Running forecasts - with ensemble based and larger things can require load balancing work.

xi. There are tools that are done with testing and diagnosing the models, but are not part of the operation forecasting cycle (e.g., calibration, benchmarking, uncertainty, etc)

xii. In the Workflow doc, there is a figure that shows the Unified Approach to developing ecological forecasts - if we had community tools, how might a new team leverage those tools to get a forecast up and running quickly

1. Have larger workflow, individual models live in individual containers, get them to the automated system, getting data

ingested, archiving, and feedback loops to decision making and theory

xiii. Are there things we have missed?
   1. Is the construction of the model separate from the workflow?
      a. The Outside of Operational Workflow section on the Workflow doc with 6 points is the development loop that is more manual and slower than the automated forecast loop. This is outside the forecasting loop.
         i. These steps need to happen for forecast development and there are tools that are needed to support these
         ii. This also highlights that ecological forecasting is not the same as ecological modeling
         iii. We are pushing forward thinking about deploying a model. Automated and using future information that you don't have yet as drivers
         iv. Have R package skeleton for the models themselves is a great idea. Could figure out what are the parts of this workflow that an individual should be responsible vs what should the collective be responsible for?
            1. If we template out a NEON Forecasting Challenge null model to show how the forecasting workflow works that can help with reducing the need to reinvent the wheel.
      b. The workflow lines up nicely with the NEON Forecasting Challenge
         i. The template starts from scratch and the construction, but after that there is alot of code and resources available particularly from Mike's course
         ii. What are the concrete tasks?
         iii. Aquatics RCN discussion - now that the group has met, there is a latency issue with the NEON data and what is feasible in the first phase of challenge forecasted drivers vs observed met.
         iv. May not be ready for the first phases, but could think about people submitting a docker container that forecasts and runs when new data is collected
   2. Want to coordinate the Forecasting Workflow with the Education Working Groups who are also thinking about developing a concrete forecasting example, but thinking about it from the broader context of what is forecasting, why do we care about the forecasting whatever topic we decide to forecast, and what can

the forecasts tell us or not tell us. Planning to use video vignettes to go over this.

a. This group had talked about aquatic sensor data, but after the aquatics group got together may not be such a good option to start with
b. Education group had talked about ticks
c. Now that all the NEON Forecasting Challenge design group shave met - phenology may be the best topic to go with. Phenology - doesn't have data latency problems

xiv. Concrete next steps

1. Skeleton of R code/packages. Create repo for test null model.
2. On Slack and the eco4cast GitHub repo, between now and next meeting we can continue the discussion and work on the code and placeholders for code that match the workflow
3. Ideas for GitHub repo name: phenology null, phenology template, go with "**phen-null**"