July 27, 2020 Joint Methods & CI Working Group Call

Attendees: Matthew Helmus, Jody Peters, Quinn Thomas, Mike Dietze, Helen Scott, Ben Toh, Jake Zwart, Abby Lewis, Carl Boettiger, Rob Kooper
Regrets: Bruce Wilson, Ash Griffin

Agenda/Notes:

1. Forecasting Workflow Updates
    a. GitHub phen-null repo
        i. There is existing code from Kathryn Wheeler that could be set up as its own repo and generalized to EFI standards https://github.com/akhtarnabeel/ecoforecastCS/tree/master/ExampleCode
            1. Grabs data from phenocam data, runs random walk model, and saves output. Output isn't in EFI standards yet.
            2. For the example, will need to change Kathryn's code to pull data from one of the NEON sites that will be used in the NEON Forecasting Challenge
            3. Mike can work with his lab to get this up and running as a first prototype
                a. Jake happy to help
                b. The Aquatics RCN is thinking about how big of a footprint people will need. If someone chooses to model a total stream watershed to make a point estimate, will there be an archive for NEON met station or can the tool be more general to grab met from anywhere?
                    i. Making it practical and scalable - every forecasting challenge will allow non-NEON data to be used as long as the data is public and the location of the data is disclosed.
                    ii. Want to set precedent for best practice of data. Provide example script for how to get the data.
                    iii. NASA SMAP data example from terrestrial carbon group - if using SMAP to constrain models, then provide script for how to download and organize into basic time series. Then other groups can decide how/if to use.
                c. Pre-staged default is 1x1 degree grid cell that each NEON site is located in. If someone wants all Wisconsin data to make a forecast, that isn't necessarily going to be made available, but the code to be scaled up could be.
    b. 2 things that we want to prep for the RCN NEON Forecasting Challenge
        i. Provide an example to people interested in participating in the Challenge to show how to create forecasts (start with download NEON, driver data,

run simple forecast in EFI standard, and push to where we are going to archive. Would be nice to also have a short tutorial about how to create a container.)

ii. What does the container get plugged into? What are the key things that need to go into the workflow. Then figure out how can we recruit people within this Working Group/EFI?

c. More efficient to have just the model in a container and then have shared resources for data assimilation
d. Common archive - need to finish evaluating what the options are. Which ones meet the requirements that we have? For the RCN - do we pick one? Or do we say any is acceptable as long as you provide where you archived the forecast?
e. Would all the different steps on the Workflow be separate containers? One whole container? How to connect the pieces?
    i. Different containers responsible for inputs
    ii. Containers responsible for scoring/dashboard stuff
    iii. Containers from each of the groups if they are submitting forecasts. Or they can run the forecasts manually or on a CRON job without using containers.
        1. If they do provide containers then it becomes EFI's responsibility to run the forecast
    iv. Quinn's diagram of the different containers (every whale is a container)
        1. Slide 2 - containers for staging the drivers we want people to work. This is high priority
        2. Downscaling is important, but we can get by without it to start with
        3. Highest priorities -
            a. Agree on what/where the archive is
            b. Provide 1 example forecast
            c. Highest priority inputs
                i. Met is the one we want to get up there first
                ii. NOAA download - have enough folks who already have this worked through who need to find a time to meet. Quinn will work on this
        4. Each Challenge Design Team should have a CI person to create a data processing container. This will be unique to each team.
            a. Doesn't need to sit in a theme-specific container
            b. If we provide one example, then the other 4 can work off that one. There will be common steps (e.g., how to push to repo, getting things into EFI standard).
    v. Any interest from anyone in the group to go back to previous meeting notes/Jake's talk to think about the characteristic of archives?
        1. Big thing we don't want is a new DOI for each new forecast
        2. Staging area with DOI after it is fully published. Mint the DOI at the very end?
        3. This could eliminate certain repositories

4. There may be very few repos available that would do what we want them to do
5. Jody will go back through notes before the next call and compile the needs and potential platforms then we can make a plan for evaluating the platforms

2. Quinn added GitHub repo - EFI Software Needs. Use the issue tracker to pose software, tools, coding needs. Then if folks want to work on these tasks they can.

3. Updates for EFI Task Views
    a. Use Task View 1 on Reproducible Workflows as a guide
    b. Task View 2: Uncertainty Quantification & Propagation, Modeling & Stats
        i. Abby, Ben, Alissa, Leah happy to help (Leah can't lead right now)
        ii. Focus on concepts and then put tools in for each of those concepts. Feel empowered to do a big level re-org and tag folks for input. This type of input we can get on upcoming calls.
        iii. For broad modeling and stats categories - can we come up with an example that goes at the beginning of the Task View. Inputs, outputs, equations, parameters, sometimes observations for calibration.
            1. Layout framework for the world you are looking at and layout the language that you will call things in that world. E.g. theta = tools for calibration
            2. Could link other Task Views with this example as well. Organize around the example
            3. Can you highlight how uncertainty goes into each section?
                a. There is the part that is all about the tools for quantifying uncertainty
                b. But there is also the part about where the uncertainty comes from
                    i. For example, it would be nice to have: if you want met and you want ensemble met, here are useful products
                    ii. This would also be nice for land use products, but Mike doesn't think there are any ensemble land use products available
                c. Data inputs, data as constraints (compare to model outputs) for DA or calibration, trait data (observation that maps onto 1 or small set of parameters)
        iv. For the big areas - if there are other good reviews. E.g., Time series analysis links to CRAN's Time Series Task View, then discuss in the Task View what is ecological useful

1. In Task View 1 - analog is the link to the website that has 200 resources but then we highlighted the top 3-5 resources that are most commonly used
2. No need to be exhaustive on the options for machine learning and time series. For Bayes - JAGS, BUGS, NIMBLE cover a lot of ground. GRETA is a new one that is pretty good as well.

v. Put reviewing this doc on the Next Agenda. Abby et al to revise outline and structure over the next month. Then on next call go over the structure and then develop concrete tasks for filling in resources

c. Data Ingest, Cleaning, Management
   i. Matt H., Jake Zwart happy to help, Chris Jones happy to help, Bruce happy to help
   ii. Folks can start working on this if they have time
d. 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.
   ii. The Social Science Working Group will look through these tools and brainstorm/add to them on their next call on August 25.

4. RCN - Update on call with CyVerse (https://www.cyverse.org/). XSEDE/Jetstream options.
   a. A sub-group has been meeting with CyVerse - NSF funded organization that works to help with computation needs for the scientific community
      i. Good for providing container-based support. Get an allocation where you can spin up an RStudio area so you can do your analyses
      ii. Allows for centralized data for NEON Forecast Challenges (covariates and NEON data)
      iii. Would give compute resources available so people don't have to use/find their own resources
      iv. Automate data prep e.g., NOAA forecasts
   b. TACC is another resource (https://www.tacc.utexas.edu/)
   c. Whole Tale is another resource (https://wholetale.org/)
      i. Helps you build a container of your analysis easily
      ii. Import data, do R code, then click and it created container from your analysis that can then be cited.
      iii. Not sure how this works with the triggering we are looking for (want to have analyses triggered when new data becomes available)
   d. CyVerse seems like what we need, but they don't have event-triggered tasks yet. But are willing to work with us on that.