October 20, 2020 Joint Methods & CI Working Group Call

Attendees: Abby Lewis, Alexey Shiklomanov, Bruce Wilson, Mike Dietze, Jody Peters, Rob Kooper, Ben Toh, Quinn Thomas

Agenda/Notes:

1. Updates on EFI Task Views
    a. Use Task View 1 on Reproducible Workflows as a guide
    b. Uncertainty Quantification & Propagation, Modeling & Stats
        i. Updates from Abby, Ben, Ash
        ii. Suggest to reorganize the text so there is a general description followed by bullets/list of tools with description
        iii. Include figures/tables where appropriate
        iv. Think of Cran Task Views as inspiration - the formatting for that gives a general description (6-7 sentences) then long list of tools with 1-2 sentences each and then longer list of packages with or without text
            1. See examples:
                a. Bayes: https://cran.r-project.org/web/views/Bayesian.html
                b. Hydrology: https://cran.r-project.org/web/views/Hydrology.html
                c. Spatial Temporal - nicest example. Has brief explanation and then has outline with tools and example: https://cran.r-project.org/web/views/SpatioTemporal.html
        v. Identify the core concepts. Pick the key tools and provide an overview
        vi. EFI Task View is like a sequential Cran Task View
        vii. Want the first screen worth of Task View to have a brief overview and list of tools.
        viii. Think about it in the framework of: What is the minimum spanning set? Not one tool to rule them all. What is the smallest set of tools that cover 95% of the needs.
        ix. Encourage sentence or two of description and then link to the tool
        x. Highlight the relative advantages or differences between packages at the highest levels.
            1. Bayes - fall into 3 classes (see Alexey's comment from August in the Task View doc)
            2. Providing general guideline
        xi. Go back to Document and reorganize in the above section
        xii. Areas we need input
            1. ARIMA using Python
            2. External executables/black boxes
                a. General section on calibrating/building models. Then have separate section on DA

        b. Ask Istem Fer or Florin (maintainer of Bayes package in R - does Bayes inference with black box models) to write this and if not Mike can
    3. Uncertainty - Abby can help with it but can't lead this section
        a. Mike had mentioned that there are stand alone tool for uncertainty. He hasn't used those tools. He has always coded them up
    4. DA section - will depend on the type of model used
        a. There are stand alone tools and there may be some existing packages that Mike isn't aware of that do DA for simpler models
        b. See if Andy Fox can do any of this since he was part of the DART at NCAR
c. Visualization/Decision Support Tools, User Interface
    i. Update from Whitney - Whitney is taking her prelims in December so she is taking a step back from this. Chris may make some progress. Otherwise, we will get back to this after December.
d. Data Ingest, Cleaning, Management
    i. Jody is leaving this in the Agenda as a placeholder since we are holding off with this until we are further along with the other two Task Views
    ii. Matt H., Jake Zwart, Chris Jones, Bruce offered to help.
    iii. Still need a leader for this Task View.

2. Forecast Data Visualization RShiny App Update
    a. This App is on the To Do List
    b. Quinn set up an operational container for the NEON Ecological Forecast Challenge visualizations to go into
    c. Katie from Mike's lab has set up a container with 3 forecasting examples
    d. Want to combine the two that fits with the Challenge and Standards and allows new forecasts from the Challenge to be added
    e. Katie expects that this will be higher on her To Do list next week. She can take the lead, but wants input from others in the group.
    f. Also want to include a forecast scoreboard
    g. Need backend code created. Currently Katie's container is connecting to mongodb or PEcAn. Needs to connect to the servers Quinn and Carl have set up
    h. The EFISA group did a great job with the Members Shiny app so it would be nice to get their input and expertise
    i. Is there a list of issues like this that need to be worked on anywhere?
        i. Find issues on https://github.com/eco4cast/EFI_software_needs/issues for cross-cutting needs
        ii. Put specific needs within GitHub repos for those tasks
    j. Duck DB is what Challenge CI is running. https://duckdb.org
    k. Could have a step that builds the database. So there wouldn't be the need to parse.

    i. Once a forecast is submitted put it into a database
l. From PEcAn example - they have a database that keeps track of files. This method sounds like a good way to go for the Challenge
m. The neonstore package has a database embedded where an individual can build up a NEON data database on their own computer. This allows people to analyze. Even though the NEON database is still pretty new, there are already issues with database sizes. For example, the soil moisture data is collected at 1 minute time intervals which is already too big for an average computer to handle. So the neonstore is an awesome way to handle this.

3. NEON Ecological Forecast Challenge CI Update
    a. The Challenge was officially launched on 10-19-20
    b. Quinn has been talking to CyVerse to be a partner in the Challenge. They have an extended partnership that needs to be approved. Quinn has submitted that once approved we can have 5% of a CyVerse employee working on this.
    c. They will help with allocations, making sure that the VM is launched in the right place so it is next to the data store to make things smoother. They will help with the backed up data store
    d. Also applied for an XSEDE startup allocation. This will be used for the 1st year which will give us statistics on usage which will allow us to upgrade as needed in future years.
    e. CyVerse will help us use their Discovery environment. Creates containers to grab data. A team can open their script, run a forecast for a certain time period using data on the CyVers system. Then shuts it down. This works for running discrete forecasts. But this won't work for continually running forecasts.
    f. In future years they will help us launch 200 R Studios for training
    g. CyVerse can support monthly forecasts. They can start a forecast and run their forecasts once a month.
    h. This won't work for the Phenology group that will need to run forecasts every day
        i. There will probably be a way to handle this, but we will need to work with CyVerse on how to do it.
        ii. Keep an eye out for participants in the Phenology group and if they need compute support. If so, this will be a good test for getting CyVerse set up.

4. Forecasting Workflow Updates
    a. How does the Workflow mesh with the RCN Challenge?
        i. The Workflow shows example of how to build forecast
        ii. The brainstorm document was started before the Challenge so there are a number of things on the list that are already up and running
        iii. Build multi-model averaging. Put this on the Hackathon, To Do list. Once the infrastructure to calculate score is in place then this will be easy. Open an issue in the scoring repo to do this multi-model averaging

iv. Wish list for next Challenge round - create the container part that forecast. Teams currently have to do this themselves and add the Scheduler part in

v. Moving forward would be fun to have people vet the null models as a group. Want documentation to show why things were done.
    1. Want to harmonize across the null models. Can we make them look like each other?
    2. The terrestrial and aquatics look alike
    3. Want the structure to look alike across all the themes
    4. All are grabbing the data and drivers, using the Standards, etc

vi. Has anyone worked with an R package that you pass in RMarkdown and then it passes back the script
    1. RMarkdown is nice for documentation. But can't use that to run in an automated system
    2. Want to harmonize those so the documentation is run through and spits out the .R file

vii. Need to make a visual map of the CI for the Challenge. Quinn will produce this before AGU (Nov 20 when posters are due)

viii. LiftR - takes RMarkdown and makes it into a docker container. If this would work it would generate a docker container with the forecast and the rendered version of your document and is saved as a container. Then upload to Dockerhub or elsewhere and others have the fully executed code and documentation when you used it
    1. Put this on the list of things to look at