## June 12, 2023 Theory Working Group Call

Attendees: Cole Brookson, Jon Borrelli, Mark Champlin Lowell, Freya Olsson, Abby Lewis, Caleb Robbins, Kathryn Wheeler, Jody Peters, Alyssa Willson
Regrets: Shubhi Sharma

Agenda:

1. Get an Update About: Model Development for the NEON Challenge
    a. GitHub repo: eco4cast/Forecast_submissions
    b. Caleb posted a project about developing machine learning models for the EFI Unconference that will be held in Boulder next week
        i. Caleb also had a chance to talk to Quinn and Mike about uncertainty and previous FLARE work that could be applied to the ML models
    c. Abby would love to spend lots of time and add lots of models, but is limited by time. Any input about prioritization - better to add more models or have more model representation?
        i. Think once we get a fair amount of variety which we are getting to, then think focusing on uncertainty is important. How can we reduce the CRPS scores?
        ii. Maybe driver uncertainty is covering many of the scenarios.
        iii. Abby has time series models with high process error that has uncertainty going down to -10C. But these are ones that are not constrained by NOAA forecasts. But either way having better uncertainty quantification is helpful.
        iv. Once we get a set of models, then getting the uncertainty to a reasonable point that is the priority. There are forecasts that are so uncertain that they are useless so they are essentially useless and you can't infer anything about predictability.
        v. It is really great, the workflow is working for the theory models that are getting submitted!
        vi. When workflow does fail it has often been JetStream issues
        vii. Another way to think about increasing the number of models without doing extra models is to generate ensembles of forecasts. Pooling models across - equal weighted linear pooling. Then if you have ensemble members and they are equally weighted.
            1. **Freya has code for this**
            2. Freya submits 3 ensembles for aquatics
                a. Jody didn't catch the 1st one (ensemble of 3 process-based lake models)
                b. Climatology & persistence
                c. ARIMA climatology - like persistence, but different.
        viii. Can we incorporate climatology into all the models because climatology does quite well.

1. Freya incorporates climatology by grabbnign ensemble members and adding them in
2. Or could model the residuals
3. Where is the code
    a. NEON forecast GitHub repo - Freya can help point out this code

ix. **As a group want to set a standard for when to change the name of the model so we can remember when things have shifted**
1. There aren't any EFI wide standards right now. But would help if we had a standard
2. If you make change that is likely to change the scoring of you forecast, then change the name of the model
3. When the changes take place, does the parent model stay running or is the parent model updated?
4. If the parent model is useful, then it might be useful to keep running

x. EFI std is to definitely change metadata

xi. Judgement on when to change name is more up to best judgment

xii. But definitely ID the parent model in the metadata

xiii. Mike thinks you can upload the updated xml file into the bucket

xiv. Need to have an easier way to generate metadata. There are 2 options 1) is to fill out a qualtrics form, or 2) create your own metadata - but the generate_metadata function is not in the package right now (Mike thinks it was removed because it was using a yaml package which created an intermediate file in a different language).
1. Mike thinks that creating an xml in R isn't too hard. Setting up the metadata is not hard. It is setting up a list of lists. The important part to get is the model uncertainty and structure

xv. Check with Quinn - what works best for submitting metadata.

xvi. To Do: Keep the parent forecast running (unless there is a critical error) and then submit a new model for the updated model. And if you can't navigate metadata , definitely write down when you make changes so we can reconstruct the metadata later

**xvii. If anyone feels motivated and qualified for creating XML documents for each model would be great. Abby isn't able to commit time to this, this month, but may be able to set up a Google doc to fill in the information**

2. Get an Update About: Comparative analysis of predictability using an uncertainty framework to decompose how predictable something is and why or why not (Shubhi, Cole, Noel, etc)
    a. GitHub repo: eco4cast/predictability
    b. Shubhi is traveling. Any updates from Cole or others?
        i. Not during this call

3. Thinking through questions about how to compare forecast performance for models that are not submitting for every single date. E.g., New forecasts submitted now, won't have forecasts for January
    a. For Kathryn's analysis for the phenology paper - were there complexities analyzing multiple models.
    b. Freya is struggling with the issue that not all the models are submitted for all times or for all sites
    c. For phenology there were models that performed better because they were forecasting at times that were easy to forecast
    d. Future ideas from the phenology manuscript is to look at key moments of forecast period - analyze all the models for a specific timeframe
    e. We used statistical models to account for issues - didn't use raw summary statistics, but it is an imperfect solution, because it is only as good as the statistical model
    f. Freya has been floating the idea to give the missing the score the climatology would have given, or give it an average of all the models. Don't give it the climatology or model means because those are really good and hard to beat.
    g. Use random walk model because that penalizes you for not submitting.
    h. Or each option has issues, so maybe don't gap fill. That is what the phenology manuscript ended up doing - not gap filling.
    i. For phenology, the day that had the most submissions only had submissions from half of the teams. But since the Theory group is submitting models, then think we can be more consistent.
    j. Wonder if we want an official start date for the Theory analysis? All the models submitted up to this point are all useful, but perhaps leave those out of the Theory paper? And only use what is up and running by Aug 1 and analyze certain forecasts submitted for a chunk of time.
    k. Power of the Theory project is that we can standardize that you can't do with the broader challenge.
    l. In Phenology paper bent the rules "some of the university courses were able to submit hindcasts as forecasts under the condition that they weren't cheating" - all had instructors that could verify that the teams had data that could only be used for forecasting.
    m. Could fill in if needed for a few weeks. Or if people are delayed a few days, there is a bit of a lag with the NEON data release by a couple of days.
    n. A more minor comment on gap filling is that it would be good to create a little leniency for broken workflows. Mike's mechanistic models crash more than my simple statistical models. Abby's time series models crash less because they don't have NOAA forecasts involved.
    **o. Need to have a broader conversation about what the theory paper will look like and what the timeline is and who is leading it. This will inform if we can set a deadline.**

p. Benefit of not setting a start date is that many of us don't have time to get models set up and in shape. But if you do have the start date, with the automated workflow, once it is set up then there isn't much to do and then the group could focus on the intrinsic predictability analyses.
q. Did the Phenology paper look at spatial analysis of how predictable sites were? Did an analysis that was based of the timing of greenup - linear regression to look at the correlation.
r. Aquatic submissions are skewed toward the lakes, but there are only 7 lakes and 6 of them are paired and in the same place. So there isn't much spatial differences. But could add more rivers. NEON put the rivers in different ecoregions so including space in the design would be cool.
s. The whole point of NEON is to leverage the structure in their data so it would be good to use that to do the analyses
t. Caleb is hoping to get a much of the workflow set up and then if there is interest at the Unconference to add new models that would be great. But if someone is working off their laptop, then won't have the computing power to get it fully running, but they can set up the structure.
u. Barrier for Abby to get new models set up is the long time for calibration that is needed.
v. Abby is trying to get the BAGLP running. It works well on her computer, but it is not working with GitHub actions - there is no error to figure out what is going on.
w. MARS-All site - "can't tune all of these issue".
x. Caleb ran into an issue that the trained models were too big for GitHub and couldn't upload them
y. If anyone at the Unconference wants to get MARS-All site running, it is super close.
z. Frustrating to troubleshoot GitHub actions because it takes so long?
aa. Could you use the Docker environment to test line by line?
    i. Yes
    ii. To get it running - go to GitHub. Go to Actions. Go to the action being run and go to Workflow. Check the image: image: eco4cast/rocker-neon4cast (but also looking into the image that works for ML)
    iii. If you have docker installed, launch it and then launch the image.
    iv. Freya has code snippet: docker run --rm -ti -e PASSWORD=yourpassword -p 8787:8787 eco4cast/rocker-neon4cast
    v. It will launch a container environment used on GitHub Actions. The script you run in GitHub actions can then be run line by line. It is an intermediate from your local R Studio and the GitHub container.

4. Discussion questions (Jody left this in here from the March call):
    a. Which variables do we expect to be most predictable, why
    b. How do we expect the relative performance of persistence and climatology to differ across variables/themes

c. Are there certain times of year that we expect to be less predictable? Are these consistent across variables?
d. How does the level of biological aggregation change predictability?
e. Do we want to tackle spatial predictability at all? Our conversations so far have been focused on temporal predictability

Previous Notes and Links for Reference

5. What hypotheses from the manuscript could be explored within one or across the NEON Forecasting Challenge themes or with forecasts listed on the EFI forecasting profiles webpage or from the EFI community
    a. Hypothesis 1: The rate of decline in predictability over increasing forecast horizons differs across variables and scales
    b. Hypothesis 2: Predictability increases with biological and ecological aggregation

6. Resources the group pulled together to test hypotheses
    a. Google sheet with a summary of drivers, data availability, number of sites, etc for the Challenge themes
    b. Lit review of models typically used for the NEON Forecasting Challenge themes
        i. Here is a google doc to compile the models
    c. Figures of hypotheses that can be examined using the forecast challenge output
        i. Google slides with images
    d. GitHub repo with code that lets people drop in models to create forecasts for the challenge: https://github.com/abbylewis/EFI_Theory