October 29, 2019 Theory, CI and Joint Forecasting Standards Working Group Call

Summary for Newsletter:
The Theory and Cyberinfrastructure Working Groups had a marathon set of calls on October 29. Much of the focus was a very productive joint call between both groups to discuss the type of forecasting standards and metadata information that will be most useful to allow comparisons across multiple forecasting models from the RCN forecasting challenge as well as any other forecasts. Moving forward, the groups will continue to refine the standards and metadata requirements and then will begin to assess how easy it is to get this information from forecasts that are currently running.
In the upcoming November Theory call, the group will talk about opportunities to work on manuscripts that examine ecological forecasting hypotheses and/or a primer on dynamic models for ecologists. The Cyberinfrastructure group will discuss responding to a recent NSF RFI on cyberinfrastructure needs and continuing to discuss the pros and cons of different archiving platforms. Anyone not already in either group is welcome to join for these discussions.

Theory Only Notes

Attendees: Peter Adler, Carl Boettiger, Amanda Gallinat, Quinn Thomas, Mike Dietze, Jaime Ashander, Bryce Mecum, Jody Peters
Agenda:

- Prep for joint CI call on standards - Oct 29 at 1:30
    - Distill into a draft - outline points of consensus and points still under discussion
    - Look at some existing forecasts in light of the emerging vocabulary from Peter's Google sheet, Will's metadata, and Carl's EML update
        - Where are sure things vs where we will run into trouble
        - Do this as a group ourselves - or split up among the group
- Working group scope / "Terms of Reference"? Chairs?
    - Examples of Scopes of Works from 1) an Outline for Decisions Science and 2) Draft from the Partners Group: LINK REMOVED
    - Not required to have Scope of Work, but other groups have done this
    - But one useful thing would be to have 1-2 chair/co-chairs for leadership or organization
        - Can we incentivize the leadership with some kind of product? Can we publish a data paper or forecasting standard.

            - This would be a target with the RCN. Describing the standards for forecasts in general and specifically for the Forecasting Challenge
            - Bring this to a mature (but not finalized state) before the RCN May meeting
            - Example papers of published standards? Yes - Darwin Core and Humboldt Core. MstMIP had paper on intercomparison - came out after MIP was done but was submitted early on in process
            - **Track down links on these papers and share with Theory and CI groups**
            - Next call - discuss other goals for the Theory group in addition to the Standards
- Any other Short to Medium terms goals other than forecasting standards
- Where are we on the decision for what to save (full ensembles or subset of ensemble date?)
    - Has come up in discussions, but no final discussion - see bullet #7 in Forecasting Standards
    - Balance between getting lots of information and needing space to save a large amount of data
        - VT water forecasting example has outputs for: Algae, DOC, oxygen.
            - If uncertainty stat is limited to CI then wouldn't have marginal
            - Trying to flatten out covariance matrix is going to be too complicated
            - Item still for discussion - spreadsheet and long format with csv providing persistent open archive. But there is argument for higher dimensional model output that netcdf has advantage over flat files. Still need to formalize. What are the allowable file formats? This will change the meaning of some things slightly
                - C-Harm has 4 dimensions if including ensembles - netcdf would be useful in this case
                - Dietze lab has been using netcdf
                - But netcdf is not something that is commonly used across the community. But netcdf/high dimensional data has applications that make them web sharable
        - Would Netcdf be a resource for the EFI community/Hackathon?
        - Define standard to be csv or netcdf. And expect people to make rational decisions about what to use. E.g., Ethan's Portal project uses csv. C-Harm spatial forecast with multiple dimensions would be more appropriate to have netcdf
        - Key things Peter laid out - project model, issue date, time, variable, etc are obvious things to be required for forecast
            - If csv - these types of information would be in columns
            - If netcdf - these types of information would be in the metadata for the netcdf

        - How do you assess if model is good?
            - Make sure to have metadata to understand what was done - uncertainty partitioning components
        - Is there a field for input models? E.g., environmental forcing data came from this source
        - Go through EML to figure out what are the tier 1 things that are critical in existing metadata standard. 2nd tier are useful for more complex analyses, but could still understand what the forecast was doing. 3rd tier are everything we want.
            - Since EML was designed for data rather than forecasts, go through to see if there are things missing that would be needed for forecasts
        - Do we want to be able to run the models?
            - If we want to run the model/partition uncertainty, then need drivers, covariates
            - 3 tier again. 1st tier - Have output archived, 2nd tier - have code and 3rd tier - have container to allow for the model to be re-run and for the forecast to be continued to be run
            - Would need to train people to use/containers or teach them how
                - There are programs in place to make containers easy to use
                - Large data/analyses will need certain resources. But there may be groups working on that, e.g, WholeTail

Forecasting Standards Joint Call

Attendees: Quinn Thomas, Mike Dietze, Jaime Ashander, Amanda Gallinat, Carl Boettiger, Peter Adler, Jake Zwart, Christine Laney, Bryce Mecum, Jody Peters, Rob Kooper, Alexey Shiklomanov, Chris

- Goals
    - Multiple models predicting the same quantity from the RCN forecasting challenge
    - Move toward a community standard (which ideally is in sync with existing standards)
    - Make sure that standards support the post hoc analyses we aim to perform (especially comparisons across forecasts of different quantities)
- Theory's proposal
    - Output columns (Peter)
    - EML metadata (Carl)

    - Additional model survey questions (Will)
    - Uncertainty classification (Mike)
- CI's proposal
    - 3 tier system for reproducibility
        i. Submitting output
        ii. Archived code / public repository
        iii. Docker/singularity
        iv. Want people to push for the 3rd tier, in particular for high-frequency forecasts in the NEON forecasting challenge
    - Containers require 4 things
        i. Standardized outputs
        ii. Standardized inputs (important to be able to run uncertainty analyses)
            - Earth Cube, Schema.org, geoschemas.org, Bioschemas.org
        iii. Tools to easily build containers (e.g. Binder https://mybinder.org)
        iv. Container storage - does Dockerhub remain the default? Github?
    - Places to submit outputs
        i. Existing archives: OSF, Zenodo, github
        ii. Have to be machine pushable
            - Input from Ethan: LINK REMOVED

- Not yet discussed
    - How to specify the 'rules' for what any specific forecast is going to be?
    - Application/technical readiness for guiding decisionmaking (probably immaterial for this competition)

CI Only Notes

Attendees: Mike Dietze, Bryce Mecum, Chris, Christine Laney, Alexey Shiklomanov, Rob Kooper

Agenda using Notes from Oct 1 Call:

- Next Steps from Oct 1 Call. Think about for Nov call
    - Look at archiving options and think about **Input Standards**
        - Below are 4 forecasts that Mike shared with the Theory group previously that we can look at for Input Standards
        - Cayelan and Quinn's reservoir forecast is another option
    - Other next steps?

- Anyone willing to provide leadership on flushing out details around archived standards?
- For RCN short term goal - is there a specific single archive to have people use? Or do we want to have people have the option for using any of multiple archives and say let us know where you put your forecasts
    - One option - start with assembling and then distribute a few minor tools. Here is your forecast, then we give script that uploads to Zenodo, OSF, etc. That eliminates decision making and gives them the option to archive elsewhere
    - Before now and RCN in May converge on recommended archive. We may need to check in with the archive to make sure they are ready for forecasts to be added
    - We could pick one archive system now. We can still change it if we test it out before May RCN meeting and it doesn't work, then we can make changes if needed.
        - Things to consider:
            - Size limit and size/cost trade-off
            - Metadata is the other main issue. Discoverability (e.g., put all variables for a forecast into one archive, or have each variable for a forecast into individual archives)
            - Do we want the ability to update datasets? This also goes along with reproducibility
            - Need to make sure we know/archive what the true forecast was vs. subsequent reanalysis efforts
            - Alexey - OSF has versioning and has been able to do machine writes. He has not found any issues with size
    - **Path forward - create a subcommittee to explore archiving options and provide a recommendation**
        - **Create list of requirements needed.** Then for next call discuss/link to example for the different archiving platforms from people who have experience with each platform
        - This could then become a blog post summarizing what we learn to make it more open to the EFI community
            - Alexey volunteers to do this for OSF
            - Check with Ethan and Matt/Bryce for their experiences

- New items
    - NSF RFI on community CI needs (**deadline Dec 16**) [thanks for the heads up Carl]
      https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf20015
    - Would be good to submit an EFI-wide response -- if CI drafts I suspect more will sign

    - Shouldn't be a big lift. 200 word abstract. 400 word response to question about current emerging data challenges. 600 word section and 300 word section.
    - This will be an area that NSF will most likely be spending money on in the next round in 2020. This is an opportunity for the things we want to write proposals about. We won't get money now but will have options for writing proposals later and for shaping what the community needs
    - **Google doc with questions, start outlining**
- **Mike to lean on folks to fill in Forecast Standards document**
- Punt Input Standards discussion to lower priority to RFI. Nailing down outputs and archives is higher priority

Email from Mike to Theory group July 22

Here are some **example forecasts** to look at that I know are running iteratively and making near-term forecasts:

* My own group has a forecast of carbon and water fluxes and pools that's accessible through a Shiny app (takes a while to load):
MIKE NEEDS TO UPDATE THIS URL
http://test-pecan.bu.edu/shiny/Willow_Creek/
This one is definitely still beta, and we haven't started writing it up yet, but we can answer any questions on Slack. FYI, this app shows one of our sites (Willow Creek, WI) but we're actually up and running at a couple more so we could look at multiple sites.
http://test-pecan.bu.edu/shiny/Flux_Dashboard/

* Portal rodent forecast: https://portal.naturecast.org/
This one also has a paper describing it: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13104

* C-HARM -Day Advanced Forecast: Pseudo-Nitzschia, cellular domoic acid, and particulate domoic acid probability, California and Southern Oregon coast https://coastwatch.pfeg.noaa.gov/erddap/griddap/charmForecast3day.graph
Project description: http://sccoos.org/california-hab-bulletin/

* Atlantic Sturgeon Risk of Encounter forecast: http://basin.ceoe.udel.edu/shiny/sample-apps/sturgeon/
Also has a paper: https://academic.oup.com/icesjms/article/doi/10.1093/icesjms/fsx187/4222666