

October 1, 2019 CI Working Group Call

Summary of the Call for the Newsletter:

During the past two calls the Cyberinfrastructure Working Group has been thinking about what it takes to make forecasts reproducible. They have come up with a three-tier system of asking forecasters to at a minimum 1) submit forecast output, then if possible 2) provide archived code in a public repository, and the ideal would be to 3) develop a containerized docker system that allows code to be rerun. The docker system would also facilitate the comparative analyses that the Theory group has been discussing. The group further discussed that with the docker system it will be important to have standards for both the input and output in forecasting models. Future calls will include brainstorming options for where to store model outputs and what kind of docker options are currently available. The group will also go through a few of the current forecast projects on the EFI website to look at the inputs to begin assessing what kind of input standards will be useful.

On Oct 29 and Dec 3 at 1:30pm US Eastern, the Cyberinfrastructure group will join the Theory group and any others interested in discussing forecasting standards.

Participants: Mike Dietze, Christine Laney, Kenton McHenry, Carl Boettiger, Jody Peters, Alexey Shiklomanov

Agenda:

- Forecasting standards - will have a joint call with Theory folks on Oct 29 starting at 1:30pm Eastern
- Next steps identified in last call
 - Previous discussion focused on reproducibility
 - 3 tier system
 - 1 - Submitting output
 - 2 - Archived code / public repository
 - Ability for a 3rd party to get rerunning in a day
 - Mike: for the Theory group, this is an unrealistically high bar -- we'll need to be able to evaluate hundreds of models to address questions about predictability, etc.
 - 3 - Docker/singularity
 - Mike: need to have standards/interoperability on inputs and outputs
 - Allows the code to be rerun
 - Want to push people to get to this tier. Don't want them to stick in tier 1.
 - From Heather's penguin forecasting challenge - there were 2017, and 2018 predictions, no 2019 predictions and no way for others to run the model and make those predictions

- For NEON forecasting challenge - will want some high frequency predictions, so will want access to the workflow so others can make those predictions
- There may be some things that are unreasonable to ask folks to provide details. In this case having the Containers would be useful for this. This will allow for comparisons to be made. Will allow forecasts to continue beyond the time the original team submit them on.
- From data provider perspective - want to know if there is an ongoing model using NEON data. This will let NEON know it is important so any changes that are made to the NEON data collection process the correct people can be told
- If the ideal is to have a containerized system - there are 3 things
 - 1 - Output standard
 - 2 - Input standard - the inputs could be different model to model, but where you find the parameters is structured so it is easy to find
 - 3 - are there tools that make it easy to make these containers?
 - Ethan's example - Binder. Kenton thinks it is a good system. He hasn't used it, but if you have GitHub repo Binder will use that and stick it in a docker. Don't want it to stay on the Binder system. <https://mybinder.org>
 - How proprietary is Binder - can we work with it and leverage it?
 - Jupyter Notebooks
 - Could Brown Dog do this?
 - DTS
 - Earth Cube - trying to define standards
 - Schema.org
 - Json LD
 - If done correctly Google can mine that
 - Adam Shephard and Doug Fils geoschemas.org - how to describe locations
 - <https://github.com/ESIPFed/science-on-schema.org>
 - Bioschemas.org
- What are the next steps?
 - **Place(s) to submit outputs**
 - Existing archives: OSF, Zenodo, github
 - Has to be machine pushable. A lot of forecasts will be automated. This rules out a lot of the existing archive systems
 - Mike thinks Zenodo is machine pushable since Ethan uses it for Portal
 - **OSF** is machine pushable. Need specific naming scheme, tracks versions

- Forecast-specific archive - define what the submission will look like
 - What about input data if you are worried about the input source being stable? Are there limits of file sizes? OSF doesn't seem to have size limits given Alexey's experience
 - **Containers** - does Dockerhub remain the default?
 - GitHub has its own **docker** management. Alexey hasn't played with it yet, but if it works, would be worth looking into
 - **Input Standards**
 - Get a few forecasts, then try to define a few input standards based on those?
 - **To Do**
 - What higher level descriptors are needed of the model?
 - What level of uncertainty is included?
 - Mike: high overlap with Theory, will update verbally as the theory call will run right before the CI call
 - Go over Theory's draft of what they have developed so far on the Oct 29 call
 - Potential paper on the current state of ecological forecasting
 - Mike: Stephanie Brodie at NOAA (ecocast team) has a similar paper in mind (at the outlining stage) and would be interested in collaborating with EFI working groups. That said, I've encouraged her to consider splitting off the "tools used" part into its own paper
 - Mike will send names of those interested to Stephanie
 - Include an overview of this in an email to the CI group since not everyone could make the call
-
- Braindump of commonly used resources/packages for ecological forecasting tasks (e.g., visualization, uncertainty quantification/propagation, data cleaning, reproducibility, etc). <https://docs.google.com/document/d/1FPYu0tbOBVRv0hEvpFXIUWO2QpUHZ2k7QnJYR9jnu0A/edit?usp=sharing>
 - This list was started by the Methods group. The goal is that this list will be a resource used by the EFI community to help people getting started with forecasting as well as for those who already are forecasting to see what Other resources are being used
 - Also provide a gap analysis of where it would be useful to create general resources for forecasting
 - **Homework: This group to add to this list. If you are adding tools, add it in editing mode. If you are adding comments to current tools, then add in suggesting mode**
- Points from the August 29 Call Brought up by Quinn and Mike on Slack after the Call:

- Reproducibility - in last call discussed the goal to have enough documentation that a student could, in theory, run the forecast in a day if they wanted to. Also discussed having to balance "full and complete reproductivity on any machine and anytime in the future" with "enough reproductivity so that folks on are too burden that they don't contribute"
- There is a need for a **review of what types of models and CI current forecasts are using** to help identify gaps and form a basis for standards. A review paper on building a community of practice in ecological forecasting would define what ecological forecasts are actually out that, what approaches they are using, and **what a community of practice could coordinate**. Part of these would involve identifying the forecasts and working on **classifying them in different ways**. This process of classifying them would for the basis for the metadata in the standards we are developing.
- From Alexey: The idea of reproducibility that was discussed in the last call was that a good metric is if a student with X background can reproduce it with Y hours/days. There are a number of ways to get to this. Easy - R packages or harder - containerization and how well everything is documented. Some combination of this will get to reproducibility.
- Next Steps
 - Look at archiving options
 - Add resources to Task Views Google doc
 - Before talking about input standards, look at specific forecasts
 - Can we use Project Profiles on the EFI website? Are there complete projects? Here's the code, here's what the project looks like.
 - 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
 - Let Mike know if you want to join Stephanie Brodie's manuscript
 - Anything to present to Theory group on joint Forecasting Standards call? Can someone go through the last set of notes to go through what the groups current thoughts/recommendations for the forecasting competition
 - Blurb for EFI Newsletter?
 - Jody to draft. Christine/Alexey to go over draft

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>