

August 29, 2019 Cyberinfrastructure Working Group Call

Summary points of the call from a discussion between Mike and Quinn on Slack after the call:

- 1) We focused what is needed to draft standards to describe the models used in forecasting (rather than what form the forecasts need to be saved). Overall our conclusion is that rather than having a separate CI group that the Theory and CI (and Methods if appropriate) group temporarily merge to a "standards" group. Folks don't have the bandwidth to tackle more than that right now so merging would be a good way to focus efforts
- 2) We never discussed actually having a 3rd party re-run a forecast. It was just a thought experiment on the level of reproducibility. One goal is that there is enough documentation that a student could, in theory, run the forecast in a day if they wanted to. We 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"
- 3) We identified 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.

Participants: Quinn, Christine, Alexey (Kenton had the wrong Zoom link and this didn't make it on to Carl's calendar)

Notes from the Call:

Tasks from July 9 call:

- Examples of output formats
 - Collect from folk
- Application of uncertainty classification matrix to some existing forecasts
- Ask existing members to 'archive a forecast on [Zenodo](#)'
- Ask existing members to fill out basic Google Dataset description a la: <https://developers.google.com/search/docs/data-types/dataset>
- Then ask for their feedback on whether this was easy/hard useful/not?
-

New items:

- We primarily discussed what is needed to document a forecasting model for the NEON Forecasting Challenge so that we can have a draft plan by the first RCN meeting.
- We focused on two things
 - What is needed for model reproducibility?

- We need to decide what level of reproducibility is necessary for the challenge
 - On one extreme of reproducibility is the idea that the submission includes enough code and documentation that someone could run the forecast with low effort
 - I.e., could a graduate student run a few scripts and get the output
 - This does not mean that we require a 3rd party to review the submission - just that we consider what is needed so that a forecast can be rerun on a different computer system.
 - Requirements: github or zip of code, list a software requirements (i.e. R packages etc.), a master execution script that runs the code, and a REAMDE.
 - Easy for simple models but could be harder for more complex forecasting systems.
 - If Dockers are part of the requirements we need to create a very clear how-to-guide for dockerizing a forecasting models. Are there existing standards to document model runs and the models themselves?
- If so, will any of them suit the needs of ecological forecasters?
 - What level is bare minimum, that won't put off people just entering the modeling space. Especially for people that will participate in the first EFI workshop.
 - **What level is needed for reproducibility?**
 - NASA - levels of reproducibility
- What's needed for the NEON forecasting challenge?
 - What format is required?
 - Idea - a grad student should be able to open and run a model within a day
 - Not actually do this - this was thought experiment for guiding what level of reproducibility is required
 - Are there common system libraries that we can assume forecasts can use without providing documentation?
 - Complexity and level of documentation required are related - the more complex, the more documentation should be provided to assist users to set up and run the model.
 - Look at the [OpenSci](#) platform - how packages are peer reviewed, published
 - Start by providing a few students with a handful of forecasts and have them run them - see what falls out in terms of lessons learned. Perhaps students of EFI members.
 - When it becomes available, look the the NEON community code sharing platform.
 - Minimum - github link or zip package, readme, code (master scripts), the data or links to the data.

- Speaking of data. Do the forecasts need to have the data used to develop the forecast. Ideally each forecast just references the relevant NEON data but since NEON data are not yet versioned it is hard to reference the exact NEON data - they are provisional and subject to change. Time stamps are at the end of each file name for IS and OS data and can be used to indicate whether data have changed since last download - but users can't reach back and request data with older timestamps through the data portal or API - would have to be a manual request.
 - Is this a major barrier to users? Shouldn't be especially once versions are out - only need to track provisional data. AOP data should be pretty stable. OS data are small. IS is probably the biggest challenge.
- Docker
 - If Dockers are part of the standard we need clear training tools so that it is very easy for folks to dockerize their forecast.
- Singularity - closely related to Docker but works better with HPCs. Can handle large data files. Fewer people know about it.
- Two pathways
 - Path for "open-source" forecast: Github, code, etc as described above
 - Path one for "closed-source" forecast: Readme, docker, etc. (this allows reproducibility without shared all the code)
- What higher level descriptors are needed of the model?
 - What level of uncertainty is included?
- Pull apart description of the model vs the forecast.
- We discussed that we need to current state of ecological forecasting
 - Ask EFI members to contribute what they have done in the past or are actively working on - basic description, link, permission to share on EFI website, etc.
 - Write paper (review) describing current state -
 - What is being forecasted?
 - What tools are being used?
 - What is needed to standardize across forecasts
 - revisit after 5 years and update. Look for domain-specific reviews.