October 29, 2019 Theory Working Group Call

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
  - 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, WholeTale

Forecasting Standards Joint Call

See this Forecasting Standards document: https://docs.google.com/document/d/1GHZfs4T-rIgM0ot2s46depE6AQ1D6ytLJdpOiJxqwYo/edit?usp=sharing

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 https://ecoforecast.slack.com/archives/CKWM8CBMZ/p15722633358012400
      -
- 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)