August 20, 2019 Theory Working Group Call

Participants: Will Pearse, Peter Adler, Carl Boettiger, Amanda Gallinat, John Foster, Mike Dietze, Jody Peters

Summary of the Call:

The Theory Working Group met on August 20 with 7 participants. Much of the call focused on thinking about how to develop and encourage standards that can be followed by ecological forecasters in order to provide metadata for new forecasts as well as allowing comparative analyses between forecasts. Our goal is to work with the CI and Methods Working groups to develop a solid proposal in time for review and discussion at the first RCN meeting in early summer 2020. In preparation for the next Theory Working Group call, the group will work to develop a data table of the types of metadata of interest, look at current EML examples used by organizations such as NEON and LTER, and think through what would be a useful, prescriptive way to request modelers to describe their forecasts.

Notes from the Call:
Key points for discussion

- Previous meeting left with the aim of contributing to the overarching goal of developing forecast standards
    - This will need to occur in collaboration with other working groups (CI, Methods, Decision, etc)
    - Role of Theory group is to make sure any standard that's agreed upon would allow us to do the comparative analyses of predictability that we are working towards.
    - To do so, we also need to have an idea about what specific analyses we aim to perform
        - Forecast limits
            - Null model?
        - Uncertainty partitioning
        - Complexity
        - Scale
        - Within-system transferability in time
        - Within-system transferability in space

    - Across-system transferability
  - To make this more concrete, we discussed looking at a few specific forecasts (see Mike’s 7/22 email, text below)
- Medium-term goals - timeline is more than a month, but less than a year
    - Have a solid plan & proposal in place by the first RCN meeting (May/June 2020) so that it can be presented, discussed, modified, and hopefully adopted
      - Forecast standard
      - Forecast metadata
      - Core analyses
      - Candidate NEON data for forecasting
    - Go into RCN with a wide list and then be able to narrow down
  -
- Goals for this meeting? Short-term. How to tackle getting to the medium-term goals
  - Reporting standards is a good place to start. Use the examples - see what they have reported and what else would we like to see
    - Amanda: All seem to have estimate of uncertainty. Mike: Maybe only half have defensible estimate of uncertainty
    - Peter was hung up on this as well. 2 were making forecasts of continuous response. Other were forecasts of rates. How do you do uncertainty when forecast is probability? If forecast is 60% is uncertainty between 50-70%? Could still have uncertainty - something captured by beta distribution. Validate probabilities and if forecasts are calibrated to probabilities.
    - Peter has 2 things he isn’t sure how to compare across forecasts:
      - Jody didn’t catch the first issue
      - Spatial aspect - with maps, lots of forecasts for lots of locations, but not all the forecasts are independent. Portal and Mike’s forecasts for one location - do you have to handle forecasts differently when you compare single locations to large scale forecasts
        - Portal is forecasting multiple species, Mike’s is forecasting multiple fluxes. So these are not independent.
    - When Mike’s group thinks about which analyses to perform - which ones only require the forecasts and which ones need the validation data? For spatial forecasts there are only certain locations and certain times that can be validated.
    - How does PEcAn archive the forecasts and the data? If Mike wasn’t here and one of us wanted to go the page - how would we compare observations to predictions? The Shiny app will take you to directory of the forecast every day to get to a netcdf for every ensemble member for every day. Has full states, and Mike thinks the full parameters. Drivers, parameters, initial conditions are available for every forecast every day

(but Mike needs to double check). Flux data that they are using is from Ankur's data from Ameriflux. So this data is publicly archived.

- For Portal probably even more transparent - Ethan has whole workflow that is GitHub based workflow. Includes archive of all data, forecasts (csv tables in GitHub repo). All of those are public. One difference with Portal is that they are saving the mean and uncertainty but not every ensemble member. Mike's group is saving every ensemble member.
    - For which analyses are we fine with mean and st error and which analyses do we need every ensemble member?
- For sturgeon and NOAA - did anyone figure out how to get the numbers? Mike didn't poke around enough to find the numbers in their archives.
- Peter is working on another project trying to get data for other projects/code. It is alot of work for each case study they are doing. We want this effort to be easier.
- Want to make some sort of spreadsheet/table that the sturgeon people, Portal people, (e.g., each forecasting group) will fill out and send to EFI. For example, ask for: mean forecast made an hour earlier, mean forecast made a day earlier. Start simple - ask for observations and predictions and metadata for the model.
- There is no way we can get null model comparison. There will be a different null model for every forecast that is not comparable.
- Define a Multi-tier System
    - Level 1 - here is the absolute minimum needed for forecasts
    - Level 2 - additional info that will be useful for forecast use by 3rd parties
    - Think of things that are necessary to be useful for scientific useful (although these things may not be necessary for the forecasts to be socially useful)
    - What are the additional layers of things we want to get out?
    - Providing numbers is a key thing you have to do. Minimum requirement.
    - Next level - treat the forecast as a black box, but be able to feed in whatever data that is needed. Transferability (level 2) - if you can input data.
    - Top level - have the option to play around with covariates; Black box coefficient thing
    - Black box approach also helps to avoid people dealing with learning each others R code
    - From Mike's experience - the black box option, given current technology, is a Docker container
        - Docker is a new variant of the virtual box concept. It is lighter weight. Doesn't take the full operating system with it. Compartmentalizes the operational needs. Isn't too

difficult to set up and has system that allows for archiving, ability to pull and update by others.

    - Similar to Travis on repositories. Pulls your code into a docker and runs in an isolated system. Tests your code, if it runs you are able to add it back into the rest of the system.
    - Mike has been using Dockers. What goes in and out of the Docker is limited by what can be passed in by a JSON file. Have the option to put some constraints on allowable names and dimensions that can go in - allows for standardization
- Want people to work on the metadata in their workflow. Don't ask them to do it after they have done their forecasts. If there is a software product that allows folks to create their metadata as they are working on their forecast that will be more effective.
- Here is a tool that will make a pretty website, or a tool that will validate it. Want well defined specifications.
- Lots of different forecasts - boutique forecasts with their own standards. How do we avoid this? How do we incentivize standards and what should the standards be?
- How to incentivized the additional work?
    - Useful tool for validation or pretty website outputs
    - Authorship on synthesis paper
    - Building community tools that can be used by the community to lower the entry to creating forecasts
    - Come up with standards and incentives early on to make it less painful to get people to go back to adopt standards
- Need to get future developers to follow the standards (easier then asking NOAA folks to go back to their forecasts to update them)
- Who gets access to the forecast archive
    - Analogy to NutNet evolution of original paper ideas and process for proposing new papers and letting others in the community know about it. If you sit on an idea for too long it becomes fair game again
    - How does this work in terms of process and access
    - Functional traits world is super possessive
    - Demography world is more open with passing around models
    - Would be nice to establish culture of open science
        - What are the expectations on both sides?
- Carl: Piggyback on what's there
    - allow / acknowledge that different communities have different standards

        - A lot is already in public domain repositories but isn’t ready for synthesis easily
        - Let it evolve, steer clear of being too prescriptive
    - Who are you doing that extra work for?
    - Bar is high for how someone might re-deploy Portal forecast to a new population time series
        - But could get there
    - Other examples (occurrence data) got into a format that was reusable because the tools drove it
- Circle back to medium-term goals:
    - Will run forecasting competition with RCN
    - Every MIP has protocol for that MIP. As part of RCN, need to figure out protocol for that forecast competition.
    - How do we leverage that to ask the broader question about creating something that is a longer-term standard so it isn’t just a one off protocol.
    - But by creating the protocol are asking people to create a new forecast that they aren’t already forecasting
    - Also asking them to turn in the new forecasts
    - This competition gets away from some of the challenges with forecasts that are already out there
    - To analyse the forecasts - encourage us to plan from day 1 to write analysis code that will be generalizable past the RCN competition
    - Create the tool that will be part of the incentive
- Methods vs conceptual? Given that we can work with other people’s data/output. Are these standards for Methods working group? Right - we need to have the CI working group involved as well. What is already created that we can adopt and not recreate.
    - This Methods Working group is uniquely suited to provide input on the standards. We provide input that allows us to get forecast outputs that meets the needs of the synthesis we want to do. What is keeping us from doing the synthesis we want to do now?
    - If we know what analyses we want to do then we can plan to have groups provide the output we want. Identify the high priority, additional outputs that would be particularly valuable
        - Partitioning of uncertainty - good for forecasts to do it up front. It requires people to think about their uncertainty early on)
        - Null model - should teams archive their null model and not just their best model.
            - How big of an ask would it be to make this a standard for forecasts?

        - Doing a run with a null model is easy. The harder part is ‘what is the null model’? In Ethan’s case -the null model assumes that rodents are density dependent, So woudl be stupid to build model without density dependency. For spatial output, such as NOAA, the null model would be different.
        - Challenge is how to get people working on different systems/models to think about null models in a consistent way?
        - If you have lots of different kinds of models making lots of different predictions - lean on the black box option. Could address a lot of the issues across comparisons.
        - Different null models may be incomparable
  - What are the additional metadata that we need to know about that black box to do the comparisons? What level complexity do we need? If all population models - would need to know what species are in each of the boxes to do the phylogenetic magic happen.
    - Will wants to have degrees of freedom and what messes around with that:
      - raw number of coefficients (simplest),
      - Ensemble modeling. Ethan’s Portal example
      - Anything that is hierarchical
      - “Describe your model” is too open ended
    - Grain & Extent
      - Spatial, temporal, taxonomic
    - Carl: have things to build on (e.g. EML, google data description)
      - EML is already used by NEON and LTER
      - Don’t have an equivalent semantic description of models (should look that up)
      - Maybe just an identifier
    - What are the new areas of kind of metadata we need
    - Need to think more about what the questions are and whether you could figure that out.
- What to do for next meeting? Schedule meetings on monthly time frame
  - Write down mock guidelines. What would the simplest data table look like? Level 1 metadata. What do the metadata fields look like?
  - Take a stab at this and use PEcAn and Portal as test cases
  - Peter will volunteer to get the ball rolling
  - Nominate Carl to do EML and Will to think about how we describe models
  - Ask PIs to fill out table as a test case

- Peter volunteers to fill out table with column headers. Look at what is would look like if we ask for uncertainty on predictions as well. Then Peter to wrangle the rest of the group.
- Carl to copy/paste EML.
- Will to write out how to describe models.
- Next meeting - Jody poll for the 3rd week of the month

Notes from Mike's July 22 email

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):
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.

* 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

Some things to think about when looking across these examples:

* Assume that the goal is to perform comparative analyses of predictability and transferability. More specifically, we're currently envisioning analyses such as:

* Comparative analysis of forecast limits (point at which the model's doing no better than chance, defined either based on a random walk null or a 'climatology')

* Comparative partitioning of forecast uncertainties: model internal feedbacks / sensitivity to initial conditions; driver uncertainty / sensitivity to external forcings & covariates; parameter uncertainty and sensitivity; parameter heterogeneity / random effects; process error

* Comparative analysis of the level of model complexity required to make parsimonious predictions (i.e. that has been subject to some sort of model selection criteria)

* Comparative analysis of how forecast limits and uncertainty partitioning change with scale

* Within system spatial analyses of transferability — can you take the model from one site and apply at another?

* Within system temporal analyses of transferability — can you use the same model / parameters at different points in time

* Across system analyses of transferability - can you use similar models, with similar covariates, and similar parameters for closely relates systems (e.g. moving a population model from a calibration species to a congeneric validation species)

* What other analyses do we want to perform?

* What would we need to extract from the model outputs for sites like these (or what would they need to deposit in an archive) to perform comparative analyses (let's hypothetically assume that 25-50 such sites existed so that we'd have some power, but not a ton).

* What sort of metadata / covariates would we want to include in the above analyses? For example: taxonomic / phylogenetic information on species being forecast; physical covariates (atmosphere, ocean, geology, hydrology); biological traits (physiology, biomechanics, demography); information about ecological interactions (competition, predation, etc); model basics (lat/lon extent, spatial resolution, temporal resolution, forecast horizon). For all of this, are there existing standards out there we should look at?

* What sort of metadata do we want/need about the models themselves and how should that be structured? Are there existing standards out there we should look at?

* How reliant are our analyses on validation data? Forecast uncertainty quantification? Forecast ensembles? Forecast uncertainty partitioning done by forecast producers? Other analyses that need to be done by the forecast producer?

* What sorts of things in the NEON data catalog might we want to run forecasting competitions on to advance all these goals? https://data.neonscience.org/browse-data