March 1, 2021 Joint Methods & CI Working Group Call

Attendees: Abby Lewis, Jody Peters, Ethan White, Jake Zwart, Leah Johnson, Alexey Shiklomanov, Libby Mohr, Mike Diezte, Quinn Thomas, Kelly Heilman, Rob Kooper, Hassan Moustahfid

Agenda/Notes:

1. EFI Task Views - use the Task View 1 on Reproducible Workflows as a guide
    a. Uncertainty Quantification & Propagation, Modeling & Stats
        i. Updates on 3 tasks identified on the last call
            1. Create figure and integrate into the different sections to give direction. Quinn - can work on the figure
                a. Added Fig 1.1. From Mike's book as a way to organize the different tasks. Don't have to use the one from Mike's book, but it is a good place to start to think about the different tasks that the different tools help to solve. Alot of tools are used to solve problem of calibrating data to model. This is part of forecasting but forecasting is different from calibrating a model.
                b. Having a figure like this that fits with all the Task Views could be useful.
                c. Would be good to map the content on to this figure. Should be pretty easy. Ben and Abby could work on this.
                d. Think about doing this after the text is written.
            2. Work on uncertainty section
                a. Ethan can help with uncertainty with frequentist and machine learning. Hassan can help with machine learning as well. Leah can help with the Bayesian section
                b. Ethan worked on a developing an outline and added the frequentist text
                c. Leah added the Bayesian section
                d. Based it on the structure above.
                e. Have nice lists of functions above, but not clear how to make that happen.
                f. Bayes section - is all via Monte Carlo, so everything you are doing is how you get the uncertainty so refer to above.
                g. Propagating uncertainty section is the next to flesh out more. Leah has started it, but Analytic vs ensemble based. Very few options to do it analytically. To what extent do we separate it out?
                h. Bootstrapping isn't in right now. Not sure where it fits. Not either frequentist or Bayes, don't usually do bootstrapping for time series

i. For completion it might be good to include a little bit about bootstrapping for fitting models.
j. Do we want to mention it and say you don't do it for time series?
    i. Mention that you may see this but this is why it is not typically used in forecasting.
    ii. Likewise, analytical stuff could be referenced.

3. Put together the DA section
    a. Jake and Alexey can work on DA section to pare down to the 3 paragraphs
    b. Target set was to have 3 paragraphs. This is a useful upper bound on how long to keep things. 1 short paragraph (4 sentences) introducing vocab of DA (doesn't have to exist for other sections).
    c. Long section is pointing out the common versions you will encounter
    d. Ends with table that cross references the different DAs and the software that can them.
    e. High level conceptual with a few references with a table with materials that can be used
    f. Want to give people what words to Google if something is of interest to them
    g. Also added links to ecoforecasting tutorials
    h. Next step is to flesh out the table - could use help from others for this.
    i. Main concern is that the tools that can go into table are more broad scale weather specific that may not be usable for eco forecasting community and may be overkill
    j. One tool to add NIMBLE sequential Monte Carlo methods geared towards the ecological community
        i. Build your model as state space model and then convert voer to use Ensemble Kalman filter or Particle filter. This is more designed for the small problems. Would fall apart of the large scale analyses, but works well for smaller data forecasts
        ii. Can Nimble do it iteratively? Posterior fit into prior to the next?
            1. Can the Nimble Ensemble Kalman filter restart where it left off with its states. Yes. Have to give it the mean value of parameters, can't give joint distribution of
            2. Can't restart with a joint distribution. Would have to start with a bunch of marginals.

3. Tradeoff is that it is harder to account for parameter uncertainty in the DA methods. Have to figure out how to have the particle filter not collapse on itself. For the enKF also have to figure out how to not have them collapse
4. In practice ecological models the states don't collapse, but if trying to estimate parameters at the same, can be hard. There is a lot of art to estimating sequential models.
5. Point out where it falls apart. It is challenging to do these things (we can't solve it, but can acknowledge where it is hard and where people are going to find places where they are going to be racking their brains).

k. Add a short paragraph (or tack on to the last paragraph) that points out the challenges. You'll apply the methods above to get your parameters, but when you launch forecasting system that involves taking a calibration into the various uncertainties.
l. Anything to add to the table?
    i. There needs to be something for Python and something for Julia. Alexey will poke around to see if he can find the Python/Julia examples
m. This section is almost done.

ii. Think about having 1 box at the beginning to provide a single example that is referenced throughout
iii. Want to clean the Task View as we go. There is some good material on the bottom of the document that could be moved up and cleaned up.
iv. Tension between tools and explanation and having too much explanation.
v. Jumping back to Uncertainty Section
    1. Put in table or bullet points for tools that are useful?
    2. Have tools in the Frequentist section. Bayes section - don't have tools, just do Bayes. Machine learning is so broad, we could pick some arbitrary tools, but doesn't have to do with machine learning as a whole
        a. Is it worth including? Since it is too broad to include specific tools?
    3. If the community is mainly doing Monte Carlo simulation, there aren't a lot of pre-boiled tools for that.
    4. Providing links to published paper or tutorial examples that show how uncertainty gets propagated in frequentist and Bayes ways.

a. If there isn't a tutorial that accomplishes it, then we can highlight this and we can use it to identify gaps that perhaps can be filled with the NEON Challenge and the repository of educational materials.
b. Mike has a tutorial on how uncertainty is propagated - reference this for now.

5. Ethan and Leah - need more direction on what we want this section to look like to work on over the next month.
    a. Trying to make it concise and highlight resources and tutorials more than explaining what uncertainty means.
    b. Tools focused. It is not a book. It is modeled after task views on Cran. It is not there to teach you, but it is there to teach you what people are actually using. If no one is doing something in the community, then mention it and say that it is not being used.
    c. Want this to be used for those moving into forecasting, but also for people in the community so they can learn what others in the community are working on.
    d. This is a working document so can be
    e. For uncertainty quantification - to some degree doesn't matter how you got your approximate or true distribution, but a tutorial on the basics of Monte Carlo simulations for this. The idea of how to differentially get the mean function vs predictive error rate. This is where Leah's students get confused. Any tutorial for this that is available? Or is this where the gap is?
        i. This is what we mean by propagating uncertainty, vs propagating process error. How to do the simulation forward.
        ii. Mike's tutorial isn't bayes vs frequentist. But assume you have a matrix of samples. So us this. Say this is how we do it this way and here is an example.
        iii. In the verbage say that while the tutorial assumes you got your samples Bayes, but it doesn't matter how you got your samples.
        iv. The tutorial sets up the section on global sensitivity analyses.
6. Kelly offered to be a fresh set of eyes to look through and see where there are connections between the different sections

b. Visualization/Decision Support Tools, User Interface
    i. Placeholder while Whitney takes her prelims
c. Data Ingest, Cleaning, Management

    i. Placeholder until we are further along with the other Task Views or have an identified leader for this

2. NEON Ecological Forecast Challenge CI Update
    a. Added a script so when people submit it checks if the file meets the standards. If it doesn't fit, it sends submissions into "Not In Standard" purgatory. So many people have a hard time dotting all the "i"s. E.g., dashes vs underscores, etc
    b. This script is still testing. It is "live" in that the script is working and Quinn emails the person.
    c. Purgatory is here: https://data.ecoforecast.org/minio/forecasts/not_in_standard/
    d. Got to get the Phenology group cleaned up with their submissions
    e. In the next newsletter we have info about purgatory.
    f. When teams fix issues, should they resubmit? If it is changing a file name, Quinn can move it around manually. But if it is something to change in the workflow, then that is more difficult.
    g. Probably better to resubmit so the old one is there so you can compare and say that the only difference is the file name.

3. Forecasting Workflow Updates
    a. Look through the final steps of the Workflow to identify what is covered in the Forecast Challenge and what will be good to keep in mind as a need for the future
        i. Save for next time.

4. Hassan wanted to share - he was at a workshop on ocean modeling where a person was talking about Bayesian intelligent (AI) that was being applied to ocean modeling and predicting ocean acidification. Combined Bayes with AI. Bayesian Intelligent Ocean Modeling: http://mseas.mit.edu/Research/BIOMAPS/
    a. Here is the presentation: https://www.youtube.com/watch?v=nhkRRoIffUk&feature=youtu.be&t=4693