December 8, 2023 Joint Methods & CI Working Group Call
Attendees: John Smith, Chris Jones, Jody Peters, Carl Boettiger, Quinn Thomas, Emma Mendelsohn
Regrets: Dave Durden

Agenda/Notes:

1. Follow up from Justin Welty's visit on the November call (see full list of notes from the call in the link to the Nov calls above)
   a. Databases and tools mentioned
      i. [Geodatabase of wildfires](#)
      ii. [Wildfire Fire Trends Tool](#)
      iii. [Land Treatment Exploration Tool](#)
      iv. [RAD framework](#)
   b. What was new or unexpected? What do we need to consider moving forward?
   c. One need for the challenge is to have a catalog of burn-area polygons. The geodatabase seems like it will be able to provide that. What does the group think?
      i. Any of the caveats about the geodata base that needs to be taken into consideration?
      ii. Targets file would query the database - are there new fires that should be included in the challenge
         1. Participants would get a set of polygons to work with
         2. Also a mechanism for moving out of prediction - e.g., 30 years out to recovery would get flagged as no longer active
         3. Targets file - don't remove fires, but would stop scoring fires. Or could set it up so people couldn't submit. But people could use it to train the model. This still needs to be coded.
         4. People could look at all fires in an
      iii. Big picture - this challenge is a proof of concept. But is it time to write a grant for really moving it forwar?
         1. Any grants in mind?
         2. Different agencies interested in different focus
            a. NASA seems interested in seeing actions and decisions
            b. USGS partner that could anchor this to show this is useful
            c. Would NASA ROSES be a good fit?
               i. Looking for 30% match.
               ii. If USGS is partner, then personnel time could be the match, it wouldn't need to be monetary
            d. Bounce a 1-page off a program officer
            e. Could go elsewhere - there is a alot of basic science and education

          i. Could go NSF if we are more in proof of principle than in the application stage

f. Carl is happy to help with a proposal, but would like to encourage others to be the lead

g. John is happy to be lead-ish, but hasn't submitted proposals before so would have lots of questions and need to read previous examples

h. Chris happy not to be lead, but happy to help

i. Emma can be PI on grants. Doestn' have PhD so for some federal grants can't apply. Has a big grant due in February

j. Timeline for submission
          i. Try a toe in a water - apply for infrastructure grant to show that we have something here,
          ii. Could be part of a big grant as a hail Mary to see what happens

k. Act of writing a grant can be a collaborative tool

l. If for NSF, need to think what the science question is and not that it is just a tool

m. Macrosystems program at NSF is done. So large scale ecology at NSF has to go through the regular programs

n. Hard to see it in DEB. But could be in DIV? DVI?
          i. By itseulf may be small, but if included in a larger infrastructure an dhsowing a clear use case and user

o. Food and Agricultural systems - allow for forestry related submissions
          i. Really focused on AI and data building
          ii. $650k over
          iii. Chris got one of these last year and submitted another this year
               1. All AI and detection based stuff for Chris' project
          iv. https://www.nifa.usda.gov/grants/programs/data-science-food-agricultural-systems-dsfas
          v. Pitch as forecasting is next frontier. AI can identify feature and we haven't built prediction into the future.
          vi. Need this challenge to keep everyone honest
          vii. Due next NOv

3. What do we need to do to get to the proof of concept?

a. Think we are ahead of where NSF sees things coming in for proof of concept

b. Don't have functional action to do target, benchmark, adn scores, and dashboard. Still a push away from that.

            i. Attempt to do this over the spring and then be set for writing grant over summer

    c. Have all the elements built out for tabular data

            i. Have some brainstorming to be done for spatial dashboard

            ii. Would do all the things we do for non-spatial dashboard

            iii. See in aggregate what models are performing better than others

            iv. Have time step as month rather than a day

4. Have we posed the right spatial scale to show this is a interesting?
    a. Grab polygons which are small compared to MODIS LAI
    b. Is there interesting scale at scale of fire polygon.
    c. We could punt and say that what we can do for MODIS then it will be useful when a new smaller scale date product is available
    d. Sentinal and Landsat are not producing the same product?
        i. Could build derivative, but is hard. There is nothing built that is equivalent to what is available for MODIS but at smaller scale
    e. Thinking of ecology, reminded of Justin's comment about how the heterogeneity of the land really influences the ability for the area to recover
    f. IBM NASA Geospatial: https://huggingface.co/ibm-nasa-geospatial
    g. Chris has been using planet data to look at sudden oak death which is similar to the NASA Geospatial info and it has worked really well.

5. Hugging Face -
    a. Think of it like GitHub with no storage limits
    b. AI is hungry for data - push your data here and we will host it infinitely
    c. A lot of AI development moved here.
    d. Wanted to share models but that means sharing model weights which took up a lot of space
    e. Organized free storage - data sets side and model side and spatial side (GitHub actions and landing page thing to run visualization with server on backend with GPU - premium model = free, bigger compute = $)
    f. Organizational part is a readme. Dataset card and model card are Readmes.
    g. It is a place to discover data and models
    h. Think it has better search features where you can search just datasets. Tagging it more than GitHub

         i. They are trying to position themselves as different from GitHub not a competitor

         j. Can push to both Git and Hugging Face

         k. Most teams will link to sources code on GitHub and model and dataset are on Hugging Face

         l. Convenient for a discovery sense

         m. Think it is a low barrier of entry for us.

      6. If GitHub action workflow won't work, then we could explore connections with IBM. Chris has experience with IBM to

   d. Should we make plans to talk with Matt Germino?

2. Project Updates: Forecasting Wildfire Recovery Using MODIS Leaf Area Index
   a. GitHub repo: https://github.com/eco4cast/modis-lai-forecast/
   b. TERN example to use as reference: https://projects.ecoforecast.org/tern4cast/
   c. Tasks to set up GitHub Action Workflow: https://github.com/eco4cast/modis-lai-forecast/issues/10
   d. Check in with
      i. John - random walk model and parametric scoring
         1. Finished the random walk baseline model
         2. Have climatology and random walk baseline model
            a. Random walk - CAR process. Neighbors have a spatial dependence on each other
            b. STRW model - end goal is to fit STRW model to LAI data
            c. Had problem with scoring - resulting distribution is a mixture of Gaussian and ICAR Gaussian. This involved writing 20K geotiffs which we don't want to do.
               i. NOAA GEFS only use 31 ensemble members. So how to reduce geotiffs
         3. John was able to get down to 250 ensemble members. The question is - is 250 ensemble members still too many?
            a. The # of ensembles is the number of geotiffs that we would be writing
            b. If doing log score and don't sample the one ensemble member that has any density then the average
            c. John didn't use the log score because some of the neighbor grid cells never have data collected. Not sure if that is an artifact of how the data is collected.
         4. Things to think about in order to get
            a. Automated data extraction/formatting to feed LAI data into nimble
            b. Do we want to use all historical data? Or last 10 years?
            c. Extraction process script is inefficient
            d. And 250 ensemble members?

5. Put in a process to deal with lack of connection with planetary that leads to NANs
    a. Started by running locally, then switched to cluster and was still dealing with lack of connection
6. John highlights places in the Rmarkdown where things can be made more efficient
    a. Can extract data all at once, the slow process is formatting into a matrix as we go
    b. Tried using the stars package. Problem is that there are grid cells that automatically remove the NAs - so get matrix of values that are smaller than pixel resolution and doesn't keep track of indices to know what pixels are removed
    c. Chris can take a look
    d. Same thing happens for climatology - but would be great to score functions on the rasters rather than ripping out all values and pushing them back in
    e. John does need to get the values into a matrix to be able to push it into nimble
    f. John shared the process he used in Section 2 of his Rmarkdown of what he did. This is the section on formatting data into a matrix
    g. Currently stacking 1 geometry. Think could make things outside the fire boundary NA
    h. Only interested in forecast only in the bounding box. But ICAR works best to throw everything in at once and then mask afterward
    i. Extract geom seems to be the wrong thing to do. But rather than do something very manual, want to do something fast
    j. How many pixel is needed on the edges?
        i. John needs to check
        ii. If just 1-2 to make sure we don't run into edge effects, but don't need much more than that
        iii. Can put bounding box and then crop to the polygon
        iv. Find the minimum rectangle around bounding box and add 1-2 pixels to avoid edge effects
    k. John will share the Rmd, pdf, and slides for the code and is asking for advice. First time for working with raster files.
    l. John has grad student working on scoring rules to extend to a spatial context
    ii. Emma - tiff submission and validation checks
        1. Doesn't have next steps on this, but can check in with Carl on it
        2. Kick off the actions and run script by script
    iii. Dave - target generation

        iv. People can claim GitHub issues - https://github.com/eco4cast/modis-lai-forecast/issues/10

           1. Select an issue from the list in the link above and save it as its own issue
           2. Carl is happy to help get people set up as they work through GitHub Action questions

3. Other Updates
   a. CI Workshop Proposal Update (Jake, Jessica, Chris);
      i. Goal: Identifying and filling gaps in CI/Methods for forecasting. Bringing together people from gov't agencies, academia, and private sector/NGO
      ii. Invitations have gone out for the April 10-12, 2024 meeting.
   b. Standards Manuscript Update - Mike checked and returned the page proofs so it should be coming out soon