February 26, 2024 Theory Working Group Call

Attendees: Caleb Robbins, Marcus Lapeyrolerie, Jody Peters, Abby Lewis, Cole Brookson, Shubhi Sharma, Saeed Shafiei Sabet, Alyssa Willson
Regrets: Freya Olsson

Agenda:

1. Marcus Lapeyrolerie - overview of the machine learning forecasts he is working on
    a. Has been building a model comparison of different ML models with NEON data
    b. LSTMs (long short term models) - has been around since the 90s and Marcus wanted to explore newer models
    c. Has relied on Darts library - times series made easy in Python
    d. Took 8 ML models (some include LSTMs, some other neural networks).
    e. There is missing data and ML require you fill in the missing data. This was the most important work Marcus has worked on.
    f. Had been using Gaussian process to fill in missing data. But had been giving unrealistic estimates for missing data so was misleading the models
    g. Benchmarks used = climatology model and persistence model
    h. Couldn't meet the performance of the climatology model - so used the climatology hindcast to fill in the missing data.
    i. Performance - compared CRPS from ML models - CRPS from climatology model for oxygen, temp, and chla variables. If better than climatology then should be negative and found that was the case for oxygen, equally well or a little worse for temp, and for chla all outperform climatology. But there wasn't one model that performed best for all of them.
    j. Did the same thing for persistence model - looked at RMSE. Same thing if negative then better than persistence. Oxygen and temp generally performed as well. Chl a - all models did as well or marginally worse.
    k. Looked at how the models are performing at the different NEON site types - wadeable stream, lake, non-wadeable river. Nothing jumped out in the pattern
    l. Looked at how model performed over the year. Nothing jumped out in the pattern
    m. Clustered by lat and long - didn't see one region perform worse than the others
    n. For every target variable at each site. Looking at site/target model combination there is a lot of variation. Some sites model performs well but at other perform worse
    o. Next steps - can take a more involved ML approach to train and tune the models.
    p. From ML want to do hyperparameter tuning to see increased performance. So think there could be better performance. But the issue with tuning is that there are so many sites and there is variance across sites.
    q. If Marcus tunes model on one site and one variable it won't translate to other sites.
    r. Think you have to tune for every site - there are 34 sites so that is a lot of tuning and tuning on just one site takes a long time. 5 minutes to train a model on a

time series. Tuning would take hours/day time frame so will be a large investment in time

s. This gels with the analysis that Caleb is doing. There isn't a particular model that is doing particularly well.
t. Covariance that Marcus is using varies per model - some models don't allow past covariance, only future covariance. In that case, using day of year. For ones that accept past covariance use air temp and something else. Could explore this more
u. Interesting to see what happens when you throw in meteorological forecasts as the future covariate. Would hope that including air temp would make prediction of water temp better.
v. Has the theory group looked at climatology vs persistence?
    i. Think it is in Kathryn's phenology paper
    ii. From another presentation Cole has seen Quinn present - there are some conditions where they do well and others where they do not do as well.
w. Freya has looked at the proportion of the time that a model is performing best out of all the models and the proportion that a model is performing worse. Found with ensemble modeling that it doesn't tend to be the best, but it is rarely the worst. But if you rely on something that is really, really good but can be bad then that can be hard for management.
x. Marcus is working on a manuscript right now with the results he shared.
y. Cole thinks that in Kathryn's thesis - had 2 ensembles, one predicting outliers and one predicting the main distribution and then joined them together to see which ensemble member was predicting better.
    i. Marcus' models are mixed for where they perform poorly. Looks similar to climatology forecast. Models sometimes miss the outlying points.
    ii. Marcus will continue to look into the ensemble analyses
z. Hyperparameter tuning - any plans for this?
    i. Did hypertune at one site and it did improve at that site, but it didn't increase performance at the other sites.
    ii. Think this is an important results to share in a manuscript and the hypertuning at an individual site that is important to a manager so they may be interested in putting in the time to do the hypertuning to work on their site

2. Check in with Cole and Shubhi - simulations, weighted permutation entropy and handling data gaps
    a. Calculating WPE for all challenge variables. WPE does poorly when you have gaps in the data. There are significant data gaps across all the challenge themes. Have been looking at where the gaps exist and how they are patterned across the challenges.
    b. Have been looking into different gap filling methods.
        i. Use process based model is one option
        ii. Rolling window averages is another option

c. Testing with the terrestrial challenge data
    i. Showed plot with the data gaps for the ABBY site
    ii. Tried closing it with rolling average of 2 days - this does close the gap and reduce the number of small gaps, but can't reduce the number of large gaps.
    iii. Tried with rolling average of 7 days - still have the longest gaps, but have reduced the number of gaps.
    iv. Looked at what happen across all sites in Terrestrial Challenge for WPE
        1. When you have gap closure method it increased your predictability, but there is no affect with increased rolling average. 2 day rolling window works as well compared to the 3, 5, or 7 day windows
    v. Next step - look at all the challenge themes.
    vi. Is the loss in predictability due to big long gaps?
        1. As you reduce your # of small gaps that doesn't seem to affect the WPE
    vii. You can fill gaps with model based method, but question is who predictable are these data (not data that we would get from a model)
    viii. So want to see if they find the same thing across all challenge themes and variables
    ix. And want to find out why you can't get improvement past 2 days
d. This gap filling sets up the workflow for other WPE analyses.
e. Have you compared the gap filling methods you used with the gap filling methods developed by the eddy covariance community? I feel like those folks spend a lot of time thinking about gap filling and so may be worth looking into
    i. Cole has not
    ii. Gap filing carbon dioxide and latent heat - eddypro is the standard, but think there are lots of people looking at long term data gaps since that is a problem when the tower goes down.

3. Check in with Caleb - GAM and LSTM analyses
    a. Working on ESA abstract - if you want to be included, add your name and comment. If you are listed on the abstract and don't want to be included take your name off. Send edits by Wed evening.
    b. Last time showed GAMs that fit poorly - but fit again with a categorical variable. But analysis is computationally tough and not feasible for each of the sites for each variable.
    c. Now working on an independent fit of the forecast
    d. Fit an independent GAM to each forecast - ran in 30 seconds (awesome!)
    e. Example fits
    f. Using Normalized NSE (Nash Sutcliffe Efficiency) for estimating fit (like R2)
    g. NSE are coming from ensemble form all ML models - the prediction means from all the models. NSE is calculated off the predicted mean

h. Density plots - is the degrees of nonlinearity. Overwhelming the edf is way away from 1 (1=linear). Most smooth forecasts from GAMS are non-linear.
i. Caleb will play around with the autocorrelation
j. Rate of change - how much did NSE change over 10 days. Most density is peaking at 0 or toward the negative so more forecast decline. Density becomes more 0 at longer forecasts.
    i. Will be fun to think about this
k. Next steps - want to do this with CRPS, but need to look into how that might be possible, particularly across ensembles, check to see how it looks across ecoregions
l. Could add basic time series models for a comparison - those are in the ensemble. When they were added in the ensemble the declines happened more and got more prominent.
m. Models may not include initial conditions but they perform well. So important to understand predictability for models without initial conditions.
n. Abby's idea: cool team thinking about predictability. Would be fun to have in person work time. There is NSF workshop funding that postdocs can apply for.
o. Can also check in with Quinn, if thinking about predictability as the group has been doing using the NEON Challenge, then think there could be RCN funds to support a group to get together.
p. Will also have a contingent at ESA.

4. Blog post idea for code review materials (Jody)
    a. On January call the group talked about Jody drafting the blog post and running it by the group. There is no definite timeline for this, but hopefully within the next month or two