Project 3: Final Project Statistics 8330 (DAIII; Wikle) November 29, 2021

DUE: Written – December 15 , 2021 (noon); Oral Presentation – December 16, 2021 (12:30-2:30pm)

Instructions: You will be assigned to a group. Your group will then analyze the following real-world problem/data.

Water is rapidly becoming one of the most precious resources that we have. Indeed, fresh water availability is crucial for irrigation, human use, and for ecosystem sustainability. One only need consider the massive wildland fires and floods that have increased in frequency in many parts of the world in recent years. We are interested in precipitation over much of North America for this project.

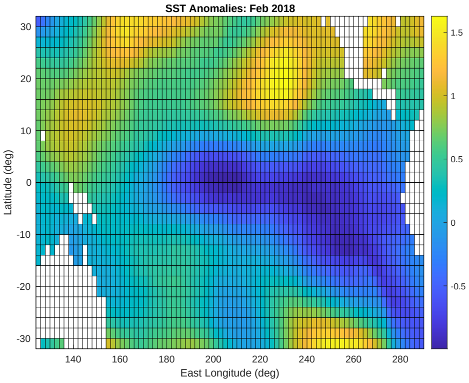
It has long been known that much of North American weather is driven on longer time scales by tropical convection over the Pacific ocean. This convection is in turn driven by fairly persistent quasi-periodic modifications of sea surface temperature (SST) in the tropical Pacific ocean. This project will be related to the investigation of relationships between monthly SST anomalies (see below) in the tropical Pacific ocean and monthly precipitation over North America.

Some things to consider:

- Are there relationships between SST and precipitation contemporaneously (meaning time t in SST corresponds to the same time t in precipitation). Does it help if one considers precipitation to be "categorical" (i.e., based on quantiles or perhaps "low", "normal", "high"). Which regions in space are the relationships strongest (i.e., are some regions of North America more related to certain areas of the tropical Pacific?).
- Are there relationships between lagged SST and future precipitation values. That is, can you develop an effective forecast model that predicts precipitation 6 months in advance (say use SST in October 2010 for a forecast of precipitation in April 2011)? With long-lead climatological forecasts, it can actually be the case that one has more predictive skill at longer lead times than shorter lead times (for some regions), yet some regions are likely to show better predictive skill at other lead times or, in some cases, no predictive skill regardless of lead time. So, are these predictions better in certain months/years than others? Are certain spatial regions better predicted? What parts of the SST domain seem most helpful? Are categorical forecasts better than predicting the precipitation values directly? Note, there are two baseline simple forecasts that must be out-performed for such a long-lead forecast to be considered effective. The first is "persistence", which is just that the prediction at time $t + \tau$ of precipitation is just the value of precipitation at time t. The second is the "climatology" forecast - which is the historical average for that month (e.g., the average of all Aprils would be the forecast for April 2011 – although be careful about using training data more than once in this setting!).

The following datasets were compiled and are available on the canvas class website:

- `SSTdata_011948_022018.nc` – This contains monthly SST anomalies (differences from a long-term average) of SST anomalies on a grid for the period January 1948 through February 2018. Note, this file format is in the self-describing "netcdf" format - which can be read directly into R. For example, the land values are given by -99 (and, you h



do not want to include those!). A plot of the SST anomalies for the last time period is shown here so you can see what it is supposed to look like.

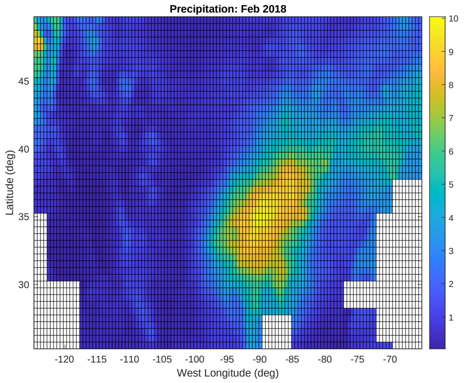
- `Pdata_011948_022018.nc` – gridded precipitation data for much of North America. Again, this is a netcdf file and details are given within the file. Note, in this case, make sure you don't include any "data" that is negative – this is considered missing data. A plot for Feb 2018 is shown below.

You can analyze these data and forecast them with any method you like, but a minimally sufficient project must consider (1) dimension reduction, (2) clustering, (3) regularization, and (4) "explainability". In your analysis, you must convince me that your prediction model works well for "out of sample" observations. One reason for considering this as a group project is so you can divide and conquer the data – that is, you can consdier these data from different perspectives and with different methods.

Grading rubric: Each group will be responsible for

- A common write-up (approx 10 pages): Introduction, Exploratory Analysis, Methods Considered, Results, Conclusion (good visuals are essential - these data have space and time structure – use it). Final R or Python code must be uploaded and runnable.
- Methods: dimension reduction, clustering, regularization, explainability, appropriate methods for considering the contemporaneous relationships and long-lead forecasts (appropriate validation); novelty will be highly valued
- Common short (10-15 min max) presentation of methods, results and conclusions to the class

In addition, each individual will be responsible for: (1) Evaluation of other members of your group in terms of their participation in the project (I will provide these confidential surveys); (2) Evaluation of other group's presentations (form to be provided) h



Groups: Your groups are assigned below. Note, I fully recognize that not everyone likes to work in groups – but, it is increasingly the nature of how things are done in the business and academic world. You should very quickly elect a group leader (after chatting about the problem for a bit) and the group leader should try to facilitate group members working on components of the problem about which they are most comfortable and skillful. As you can tell from the nature of the problem, it is a very big problem with many avenues of exploration – hence the reason for the fairly large groups. Note that you will be asked to fill out a (mandatory) evaluation form for the other group members and the group leader. These are completely anonymous and I take them very seriously – they can affect your grade. If you have problems with your group, please let me know. I don't want the experience to be horribly unpleasant for anyone! (Also, if I missed you - let me know immediately!)

- Group 1: Ping Liao, Jay Zhang, Bujingda Zheng, Fangda Wang
- Group 2: Richard Byfield, Kianoosh Sattari, Harpreet Kaur, Suhwan Lee, Jian Liu
- Group 3: Katie Price, Emily Scully, Ben Graves, Ellen Fitzimmons, Mira Isnainy
- Group 4: Abi Bigham, Justin Marrs, Dan Vedensky, Aaron Bogan