Statistical Learning for Stochastic Tropical Cyclone Simulation

Culminating Experience

Laurette Hamlin*

4/4/2021

Contents

1 Introduction **3**

2 Research Design **4**
2.1 Environment 4
2.2 Data Structure 4

3 Statistical Learning/Spatial Techniques Used **5**
3.1 Smoothing Splines 5
3.2 Random Forests 5
3.3 Spatial Smoothing 5
3.4 Principal Component Analysis (PCA) 6
3.5 Empirical Orthogonal Functions (EOFs) 6

4 Data Processing **6**
4.1 Ingestion and Cleaning 6
4.2 Time Synchronization 7

5 Data Transforms **7**
5.1 Estimation of Pressure Deficit at Sea Level 7
5.2 Calculation of Storm Direction 7
5.3 Calculation of Wind Speed 7
5.4 Calculation of Distance to Coast 8
5.5 Estimation of Storm Centers 8
5.6 Verification of Storm Centers 8
5.7 Estimation of Radius of Maximal Winds 8

6 Analysis of Precipitation **8**
6.1 Projection to Polar Coordinates 8
6.2 Singular Value Decomposition 10
6.3 Empirical Orthogonal Functions 10

*Candidate for Master of Science in Applied Mathematics, University of Colorado, Boulder.

7 Fitting the Statistical Model **11**

8 Conclusion **12**

9 References **12**

References **12**

1 Introduction

For my culminating experience, I worked with Dr. Will Kleiber on his continuing research to model precipitation fields generated by tropical cyclones. My contribution was to apply the research method on a new location: Miami, Florida. From data created through large atmospheric models of 24 past tropical cyclones, I helped to establish a model that accurately estimated asymmetric precipitation fields and simulate tropical cyclones using only a handful of predictors from standard track databases.

The predictors were spatio-temporal measurements of

1. location of storm center (separated into latitude/longitude),
2. storm direction (separated into latitude/longitude),
3. atmospheric pressure deficit at the storm center,
4. radius of maximal winds, and
5. distance from the storm center to the coast.

This area of research is both timely and important. Extreme rainfall during tropical cyclones, already at unprecedented levels, is forecast to increase $10-15\%$ by the end of the century. ("Hurricanes and Climate Change," n.d.) Some the of the greatest rainfall amounts occur inland, for example when a slow-moving storm events stalls, causing a major risk to life and property. ("Hurricane Flooding: A Deadly Inland Danger," n.d.) One of the storm events in this study, Hurricane Katrina in 2005, was the costliest natural disaster in U.S. history at the time, with asymmetrical rainbands south and east of center resulting in localized flooding in Miami that caused over $500 billion in damage and left 1.45 million people without power. It killed 14 people in Florida, and went on to cross the Gulf of Mexico, strengthen, hit New Orleans, and ultimately claim more than 1800 lives.

Statistical learning is an important tool in atmospheric modeling, where stochastic estimation of quantities such as amount of precipitation is quite advantageous. There are several reasons for this.

- Tropical cyclones are relatively rare meteorological events. Baseline data can be difficult to obtain.
- Fine-grained measurements are usually not available because instruments that gather atmospheric measurements are strategically spaced apart.
- Typical physics-based models use advanced algorithms that can require so much computing power and time that they are beyond the reach of fast, localized desktop estimations.

Statistical learning is a robust technique that fills these gaps. At first, I was astonished to think of using stochastic processes to *create* predictors. It felt like putting the cart before the horse. But after working with the data and seeing the models and explanations Dr. Kleiber taught, plus seeing estimation of quantities such as regression parameters and residuals in other courses, I became comfortable and then enthusiastic about being able to apply stochastic methods in many ways. I realized that a true applied data scientist will almost certainly need to apply stochastic processes to make estimations at several steps in a study. This experience helped me learn how to do that.

Spatial statisticians face challenges, however. Stochastic tools have to be used in ways that preserve model correctness. Consideration must be given to the fact that data on the spatio-temporal grid are not independent. Covariances become the major consideration. Models require careful planning

and verification, and multiple techniques may need to be considered and implemented. A full range of supporting skills is also necessary, ranging from Linear Algebra, Trigonometry, and Mathematical Statistics to statistical programming using specialized packages and creation of visual output. It is a full statistical and data science experience.

2 Research Design

The basis of my culminating experience is a research paper entitled 'Stochastic Tropical Cyclone Precipitation Field Generation' (William Kleiber et al. 2020) which successfully modeled rainfall patterns during tropical cyclones in Houston, Texas. I tested the success of the research by replicating and validating the process on new storm data from Miami, Florida.

2.1 Environment

As with the original study, I used an R-based environment to perform the Miami research. R is a friendly mathematical software language and graphics display tool designed specifically for statistical computing. A large catalog of additional packages are also available to address specific situations. In this research setting, the main specialized packages added were: 'fields,' a package of tools for spatial statistics, 'LatticeKrig,' a multi-resolution kriging tool based on Markov random fields, and 'randomForest,' a random forest generator for classification and regression.

2.2 Data Structure

There were 35 named storms near Miami in the study window of 40 years from 1979 to 2019.

David (1979)	Dennis (1981)	Barry (1983)	Isidore (1984)	Bob (1985)
Floyd (1987)	Chris (1988)	Marco (1990)	Andrew (1992)	Gordon (1994)
Erin (1995)	Jerry (1995)	Georges (1998)	Mitch (1998)	Floyd (1999)
Irene (1999)	Gabrielle (2001)	Michelle (2001)	Charley (2004)	Frances (2004)
Jeanne (2004)	Katrina (2005)	Ophelia (2005)	Rita (2005)	Wilma (2005)
Ernesto (2006)	Noel (2007)	Fay (2008)	Bonnie (2010	Bret (2011)
Emily (2011)	Isaac (2012)	Arthur (2014)	Julia (2016)	Matthew (2016)

Two sources of data from these storms were used in our research, one from a numerical forecast prediction model and one from actual historical measurements.

The Weather Research & Forecasting (WRF) model from the National Center for Atmospheric Research (NCAR) is a mesoscale numerical weather prediction system. It was used to simulate the past tropical cyclones and, in so doing, establish a physics-based framework for synthesizing data used to forecast precipitation fields. This is necessary because, while these types of models are excellent for storm center location and intensity estimates, they typically do not include precipitation estimates. In this study, the WRF output was in the Network Common Data Form (NetCDF) data format and included 79 variables tracked over time. This was pruned and cleaned down to ten variables of interest.

The WRF simulation gave us our spatial frame, centering Miami in a coverage area from 20° to 31° North and 75° to 84° West. A 59x59 grid covering this area provided the 2-dimensional matrix of spatial data locations.

Actual data from these storms is archived by the National Oceanic and Atmospheric Administration (NOAA) in the International Best Track Archive for Climate Stewardship (IBTrACS) system. Each tropical cyclone downloaded as a comma-separated value (csv) text file with 21 measurements, including the three of interest here, namely storm center in latitude/longitude coordinates, tracked over time. This data was used to validate the data processing component of our research, where we constructed storm centers from WRF data.

3 Statistical Learning/Spatial Techniques Used

I was fortunate to be able to use a number of methods learned from my PMD coursework in this research study. The four main techniques are outlined below.

3.1 Smoothing Splines

This is a form of non-parametric regression used when covariates are not linearly related, in which case polynomial regression is used. In the case of smoothing splines, separate models are fitted piece-wise as step functions (splines), and then smoothed with a penalty function so that the coefficients are reduced by a relationship with an input parameter λ.

3.2 Random Forests

This is a decision tree-based method for regression analysis. The predictor space is separated into segments or regions, and inferences are made using splitting weights at each step. Random forests create splits by only considering a random sample of fewer options for splits. This removes correlation among the decision trees in the forest.

3.3 Spatial Smoothing

The bulk of my culminating experience was spent working with spatial smoothing, which I had just learned in STAT 5430. I used this frame of thought as I worked: we can calculate an estimate at any point, not necessarily on the pre-defined grid, by taking a weighted function of all other points where the weights decay with distance, perhaps through a powered exponential kernel, a Matern kernel, etc. The value of points close on the grid have significant impact on the estimate, while values further away have much less weight. But all points are taken into account. This combines with the covariance to make a prediction function.

In a Gaussian process, the behavior of a process can be completely defined by the mean and the covariance matrix. For stationary processes, the mean is 0, so the covariance matrix alone is used. In the case of spatial models, the covariance matrix (and mean) are often described parametrically, as in this study, and are often estimated empirically. There are several kernels for these types of models, which can be tuned by setting a bandwidth (sometimes called scale) that is used to

establish the rates at which weights change. These spatial covariance matrix functions are what allow interpolation (and prediction) to occur.

3.4 Principal Component Analysis (PCA)

This is an unsupervised learning technique used to reduce the a set of features down to just the ones that form a basis of linear combinations. In this study, we used the closely related Singular Value Decomposition (SVD) to determine the square roots of the eigenvalues of the basis.

3.5 Empirical Orthogonal Functions (EOFs)

As the name implies, this is a stochastic function technique for estimating an orthonormal basis from a given data set. The EOFs are found by normalizing the eigenvectors from the single value decomposition of the covariance matrix. In the atmospheric sciences, they are commonly used because the time dimension can be interpreted as giving many replicates of the same spatial process. In this study, we interpreted each tropical cyclone as a replicate of the same spatial process.

4 Data Processing

The first part of the study involved putting the ingested data into a form that was compatible with statistical learning and verifying it.

4.1 Ingestion and Cleaning

The WRF data was processed into 10 measurements following proper ingestion considerations that were taught in several APPM and STAT courses. This included watching for measurement units, handling missing data, and correctly combining data for analysis. For example, latitude/longitude measurements could come in as negative values, indicating they were "Eastings" and "Northings," and had codes of "-900" to indicate missing values. Dates needed to be aligned to UTC and trimmed to match the intersection of the two data sets. Precipitation was totaled from two separate measures of total cumulus precipitation plus grid scale accumulation in millimeters. We were not interested in the total amount of precipitation, however, but the additional amount from one point in time to the next, and so transformed that measure from point data to difference data.

The special nature of spatial data really came out during this step. I first had to hone both my internal vision of the data and my R data structure skills. Rather than dataframes, which I usually use, I learned it was best to use equal-length lists for each event so that we could align all measures to a 3-dimensional array (latitude/longitude/time). This was the programmed application of the spatial random variable $Z(\boldsymbol{s}, t)$ where $\boldsymbol{s} \in D \subset \mathbb{R}^2$.

A typical example is tropical storm Bonnie, which blew near Miami in July, 2010. Her sea level pressure, for instance, was simulated by WRF at 59x59 locations on the spatial grid every hour over about 3.5 days, giving a total of 59x59x84 data points.

4.2 Time Synchronization

For each tropical cyclone event, the time dimension needed some cleaning up. The WRF data occurred at one hour intervals and, since it was simulated, the value at time "2010-07-24 18:00:00 GMT," for instance, was actually the integration of "17:00-18:00." This made it preferable to use midpoint times, i.e. "17:30."

The IBTrACS data were real point measurements at particular times, but only occurred at six hour intervals. Therefore, we interpolated the IBTrACS data from 6 hours to 1 hour increments and then back by another 30 minutes to also be centered at, say, "17:30." These interpolated times were then applied to each lat/long vector of the IBTrACS storm centers through a natural spline function so that boundary points (knots) stayed smooth.

5 Data Transforms

The WRF data provided was transformed into response and predictor variables in steps. This was a major learning experience for me and took the bulk of my time because it involved much more than the typical checking over and perhaps transforming of existing variables we had learned in STAT classes. All of the variables had to be modified or created by using advanced analytic and statistical methods. I learned that spatial data, and atmospheric data in particular, requires quite a lot of very careful consideration and preparation.

5.1 Estimation of Pressure Deficit at Sea Level

First, we considered wind-pressure relationship measurement fields, which were initially stored parametrically by lat/lon vectors, plus had a third spatial dimension for levels of elevation. Because of this, a spatial smoothing package was necessary ("fields") due to the extra dimension, which added a third distance measure for pressure differences as elevation changed and caused the covariance estimation function to become computationally complex. Our spatial interpolation used a normal kernel smoother for 2-dimensional vector fields with a default bandwidth of $\lambda = 10$. For each tropical cyclone, a sea level pressure deficit predictor vector was then created by finding the estimated pressure at the nearest grid point to the storm center and subtracting it from the average pressure at sea level (1013) at each point in time.

5.2 Calculation of Storm Direction

Rainfall is not distributed uniformly around the center of a tropical cyclone. Rather, it tends to morph around in bands and uneven masses, and the way the storm moves affects that. Therefore, it was important to calculate a storm direction vector to use as a predictor. This was a simple difference at each time step calculated in lat/long parameters from the storm center estimates.

5.3 Calculation of Wind Speed

This was also a simple, one step vector norm of the smoothed sea level pressure field.

5.4 Calculation of Distance to Coast

In R, the 'map' function is usually used to draw geographical maps on visual output. However, it is also a comprehensive geographical database. We were able to use that to calculate the minimal distance from the storm center to the coastline at each hour.

5.5 Estimation of Storm Centers

The WRF model did not provide simulations of storm center locations; these needed to be calculated by hand. We accomplished this with Vector Calculus by taking the maximum of the curl of the smoothed pressure vector field above at each time point. This allowed us to create simulated storm centers.

Finally, these storm center estimates were pruned to only include measures within 2° of the real IBTrACS storm centers and then were themselves interpolated using a cubic smoothing spline on each long/lat vector with a parameter value of $\lambda = 0.75$.

5.6 Verification of Storm Centers

Not all tropical cyclones storm center tracks were able to be successfully estimated and smoothed, and were dropped. This left us with 24 storms. I verified our estimations by creating graphs of storm center tracks. Figure 1 shows the estimated storm centers matched well.

5.7 Estimation of Radius of Maximal Winds

First a temporal vector was created by looking at the wind speed on each grid point and choosing the greatest value at each point in time. The values were capped to only consider distances within 400 kilometers of the storm center. Finally, a smoothing spline was applied ($\lambda = 0.6$).

6 Analysis of Precipitation

6.1 Projection to Polar Coordinates

This most interesting and challenging aspect of the data processing was also quite intuitive: precipitation in a hurricane does not fall uniformly from the sky. It forms and moves in accordance with the circular, spiral character of the tropical cyclone. Therefore, it makes things much easier computationally to transform the spatial measures from fixed latitude/longitude coordinates to polar coordinates radiating out from the storm center.

The assumption at the beginning of the study, that a lat/long grid system would be used to create those distances, was overly cumbersome. We preferred to calculate distances from the center of a spiraling, basically circular storm system which moved. Thus, Dr. Kleiber's research called for switching the spatial measurement parameters from the fixed rectangular grid (lat/long) to a relative polar grid (r=distance from storm center, θ=angle from 0°). Figure 2 shows the projection.

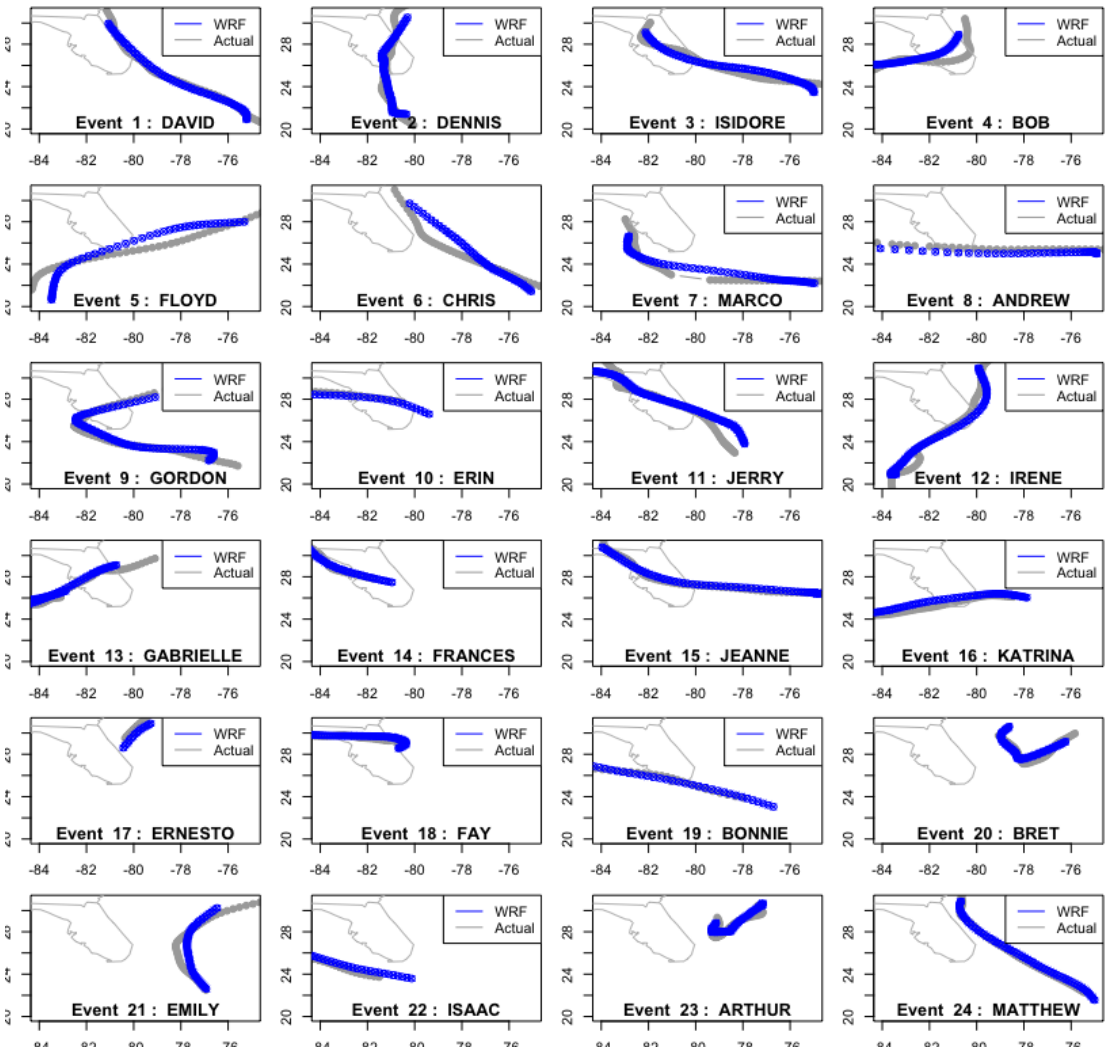

Figure 1: Storm Center Verification

This polar grid was created with a rather fine grain of 100 intersections for each parameter over the study area. The projection was made using typical distance and angle measures (I learned a cool spatial field function called 'rdist.earth' which calculates geographic distance matrices, usually for empirical variograms, but here for polar coordinates).

The precipitation data had to be smoothed against the polar grid. Kriging is common method used for spatial smoothing. It creates optimal predictions stochastically by minimizing the mean squared error of the error. Both the variances and the distances within the random field are used. For smoothing the precipitation data on a polar grid, however, we needed the additional help of a highly specialized function called LatticeKrig. In its documentation, Lattickrig describes itself as, "a variation of Kriging with fixed basis functions that uses a compactly supported covariance to create a regular set of basis functions on a grid. The coefficients of these basis functions are modeled as a Gaussian Markov random field (GMRF)" (Nychka et al. 2016).

We latticekriged the precipitation data on the long/lat grid for every time step. We then ran predictions for every point in the polar grid, giving us a very, very cool field of smoothed precipitation

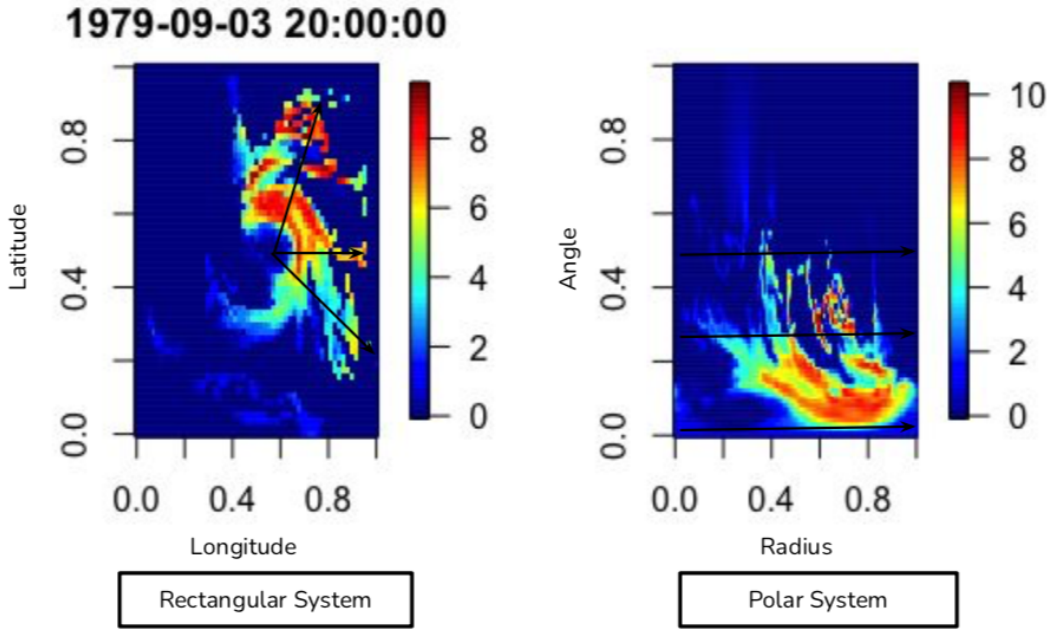

Figure 2: Projection to Polar

estimates in polar coordinates. This was computationally intensive, however, and took an evening to run.

6.2 Singular Value Decomposition

The next step was to take a singular value decomposition of the matrix formed from

- the polar spatial grid matrix ($100 * 100 = 100,000$ rows)
- the smoothed precipitation measures at each hourly time interval.

This gave us what is generally known as the $\boldsymbol{Z} = \boldsymbol{UDV}^T$ matrix where D is the diagonal matrix of singular values that represent the square roots of the eigenvalues. They occur in decreasing order, so can be used to calculate a desired cutoff of the percent of variance explained.

We started with a rough estimate derived by leaving out the first tropical cyclone (for a total of $1,538$ columns in the $\boldsymbol{Z}$ matrix).

6.3 Empirical Orthogonal Functions

At this point, we needed to make a decision. We needed to balance the amount of variance to explain against the number of EOFS to keep. A large number of EOFs would complicate further analysis steps, like the upcoming random forests. For the Miami data, the possibilities from above boiled down to:

Number of EOFs	Variance Explained
12	75%
23	80%
94	90%
231	95%

We went forward with 12 EOFs, confident that explaining 75% of the variability was a good tradeoff. It would have been necessary to almost double the number of EOFs to achieve even 5% more explanation.

At this point, we re-ran the SVD analysis using all tropical cyclones (for a total of $1,610$ columns in the $\boldsymbol{Z}$ matrix) and found that, as expected, the percentage of variance explained had improved slightly and we were still at 75% of variability explained.

The final step was to interpolate the storm centers and EOFs back to the lat/long grid space. This was actually quite fast and simple because we were able to use a function in the spatial 'fields' package called 'iterp.surface' that used bilinear weights to interpolate values from the polar grid to the rectangular grid.

We now had EOFS in the lat/long field to continue working with. Figure3 shows quilt plots of the first 12 for the first event, major hurricane David in August, 1979.

7 Fitting the Statistical Model

At this point we had everything we needed to create a model that predicted principal components and mean value fields for precipitation. The steps used to actually create the model were

- Run a full leave-one-out cross validation for all events
- Fit a Principal Component Model using Random Forests

After running the random forest algorithm, we were able to determine the importance of each of our variables, as shown below.

Predictor	Importance
Pressure Deficit at Sea Level	2814298.06
Storm Center - Polar Radius	1265974.08
Radius of Maximal Wind	1024219.61
Storm Center - Polar Angle	1000045.48
Wind Direction - Polar Angle	800550.80
Wind Direction - Polar Radius	757230.31
Distance to Coast	699380.21
White Noise	95819.21

These were encouraging results. The pressure deficit as main variable importance measure matched the original Houston study, as did radius and angle measures. The white noise variable was added by hand in order to verify the predictors were well-chosen and did, in fact, have significance in the model.

8 Conclusion

One of the nice things about the earth systems science (ESS) arena is that global data is open and readily available, and the scientific community has robust infrastructure in place to communicate and share that data as well as ongoing development efforts and best practices. It is an area of active research. A rich toolset of statistical and physics-based techniques has already been developed, but there is plenty of research still to do. It is critically important for our planet and ourselves that we continue these research studeis.

Especially beneficial for the statistical modeler is that past weather generally provides good training data for current and future predictions. It is reasonable to feed in clean data generated from simulations of historical tropical cyclones and then validate model results against the actual historical data measurements. In this way, we can train for predictions on future tropical cyclones with confidence.

I could not have asked for a more challenging or rewarding culminating experience. While the learning curve was quite steep, seeing the precipitation prediction plots was the thrill of a lifetime. I definitely want to continue forward with statistical learning, and spatial modeling in particular.

9 References

References

Cressie, Noel A. C. 1993. "Statistics for Spatial Data, 2nd Ed."

Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain. 2017. "Fields: Tools for Spatial Data." Boulder, CO, USA: University Corporation for Atmospheric Research. https://doi.org/10.5065/D6W957CT.

"Https://Www.mmm.ucar.edu/Weather-Research-and-Forecasting-Model." n.d. National Center for Atmospheric Research (NCAR). https://www.mmm.ucar.edu/weather-research-and-forecasting-model.

"Hurricane Flooding: A Deadly Inland Danger." n.d. National Weather Service (NOAA). https://www.weather.gov/media/owlie/InlandFlooding.pdf.

"Hurricane Katrina: America's Costliest Natural Disaster." n.d. National Hurrican Center . https://www.national-hurricane-center.org/hurricane-history/hurricane-katrina.

"Hurricanes and Climate Change." n.d. University Center for Atmospheric Research (UCAR). https://scied.ucar.edu/learning-zone/climate-change-impacts/hurricanes-and-climate-change.

"Ibtracs - International Best Track Archive for Climate Stewardship." n.d. NOAA's National Climatic Data Center (NCDC). (n.d.). https://www.ncdc.noaa.gov/ibtracs/.

Kleiber, Will. Fall, 2020. *STAT 5430 - Spatial Statistics Course Textbook*. Department of Applied Mathematics, University of Colorado, Boulder.

———. Fall, 2020. *STAT 5610 - Statistical Learning Course Textbook*. Department of Applied Mathematics, University of Colorado, Boulder.

Kleiber, William, Stephan Sain, Luke Madaus, and Patrick Harr. 2020. "Stochastic Tropical Cyclone Precipitation Field Generation." http://arxiv.org/abs/2011.09918.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Nychka, Douglas, Dorit Hammerling, Stephan Sain, and Nathan Lenssen. 2016. "LatticeKrig: Multiresolution Kriging Based on Markov Random Fields." Boulder, CO, USA: University Corporation for Atmospheric Research. https://doi.org/10.5065/D6HD7T1R.

Pierce, David. 2019. *Ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files.* https://CRAN.R-project.org/package=ncdf4.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

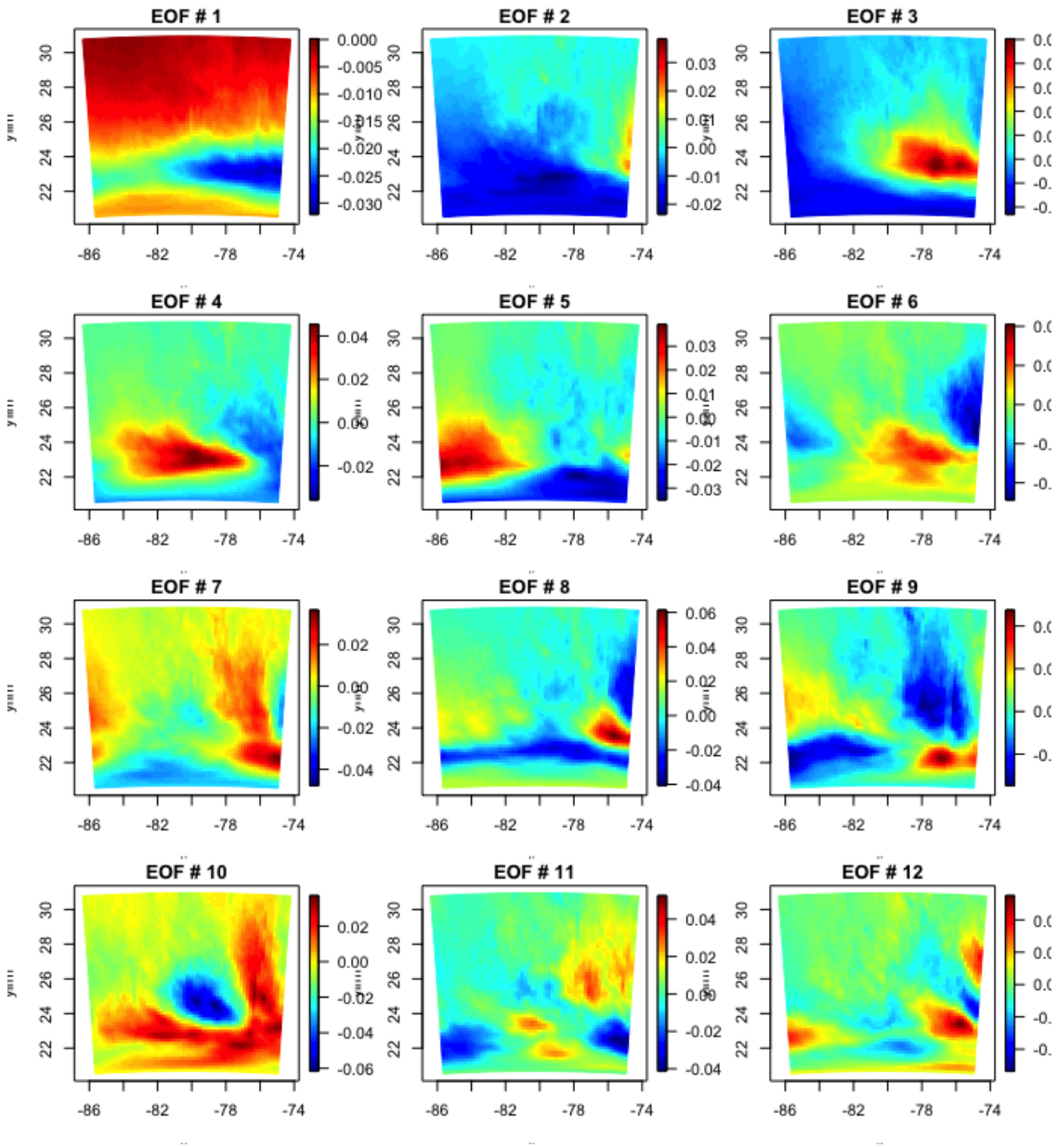

Figure 3: First 12 EOFS