An Analysis on the Accuracy of Weather Forecasts

Benjamin Schweitzer[1] Robert Garret[2] Dr. Thomas Fisher[3]

[1]Corresponding author Benjamin Schweitzer is a third year Undergraduate Math & Statistics Major at Miami University, Email: schweib2@miamioh.edu

[2]Robert Garrett is a third year Undergraduate Math & Statistics Major at Miami University.

[3]Faculty Advisor Thomas J. Fisher is an Associate Professor in the Department of Statistics at Miami University, Email: fishert4@miamioh.edu

Abstract:
One of the most popular usages of predictive modeling is in weather forecasting. We aimed to utilize machine learning techniques to extend provided forecast to sites across the continental United States. The forecasts and observed weather for 113 sites across America (2014 - 2017) were supplied to us by the competition officials. Supplementary data on observed weather was acquired from the National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center. We then created visual displays to analyze the prediction accuracy of the extended forecasts. Our results allow for an in-depth exploration into the accuracy of weather forecast across the nation.

Key Words: Weather Forecasts, Machine Learning, Data Visualization

1. Introduction

Weather forecasts are made for specific locations (i.e., weather station or airport), but are generally interpreted as a regional forecast. For example, the city of Oxford, Ohio is a periphery city for both Cincinnati and Dayton, and people use both city's forecasts to get a general sense of what the weather will be like in Oxford. Therefore, we decided to attempt to expand the 113 location-based forecasts given to us by the competition officials across the United States and then use this expanded forecast to examine the accuracy of forecasts over the station's closest regions.

The data provided had observed temperature maximums and minimums, along with the forecasts for those values. By acquiring more data from NOAA, we were able to then gather observed temperature maximums and minimums for 8258 sites across North America. Through machine learning techniques, we can tune a model to most accurately predict the observed minimum or maximum temperature in one of the NOAA sites using a subset of the observed minimum or maximum temperatures at the 113 provided weather stations. This model can then use the forecasted minimum or maximum temperatures at the 113 provided weather stations to generate a forecast based on a tuned model.

The generated forecast can be assessed for accuracy with the observed minimum and maximum temperature we have for the site. This metric of our expanded forecasts accuracy can be calculated for each day we have records for each site in the NOAA dataset to have a collection of weather errors that will be visualized to explore trends in forecast accuracy across the regions of not only the contiguous United States, but also the southern parts of Canada and the northern parts of Mexico.

2. Data

2.1 Data Expo Data

The Data Expo Data is the data that was supplied to us at the beginning of the competition by the JSM officials. The observations consisted of an Airport Code to identify which site this record is for, along with observed and forecasted maximum and minimum temperature records for the given date. These were the parts of the data that were utilized. The time period ranged from July 2014 to September 2017. The data was filtered and cleaned to remove any outliers, or invalid data points.

2.2 NOAA Data

Now that we have the forecasts we want to extend, we need many more locations to expand our forecast for, along with observed minimum and maximum temperatures to assess the accuracy of our expanded forecasts. By going to the NOAA National Climatic Data Center at ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/ we were able to get observed weather at certain latitude and longitude coordinates by day. The data was filtered and cleaned to remove any outliers, or invalid data points. We now have sites and observed whether to compare our generated forecasts too.

3. Analysis

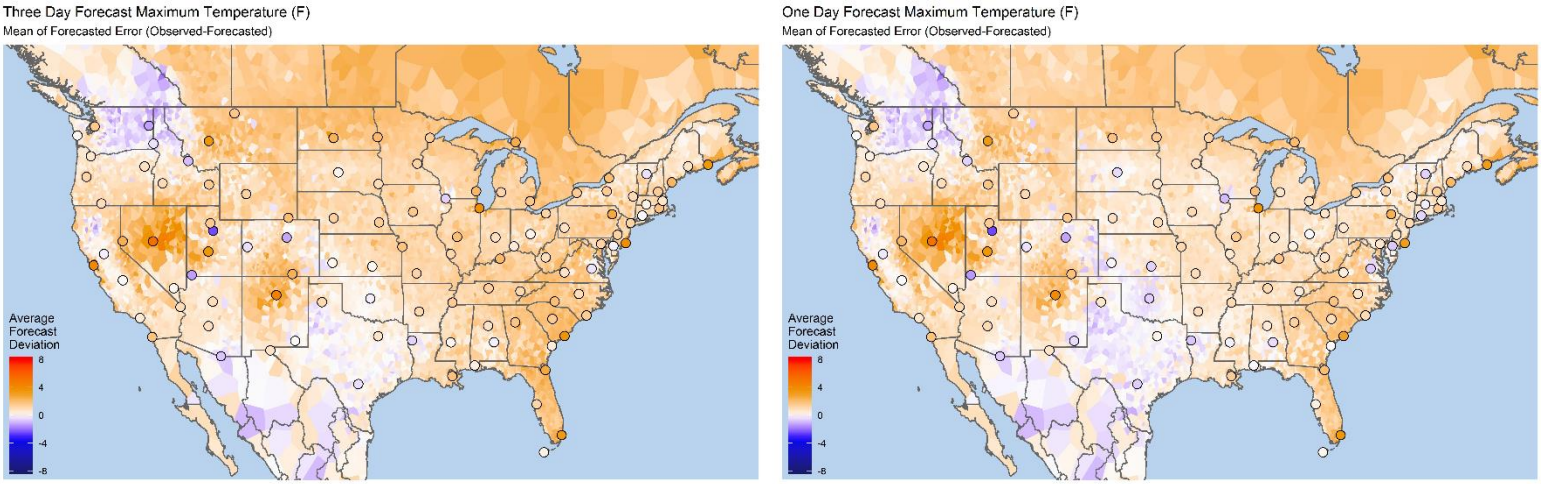
3.1 Model Building

To generate forecasts, we needed to build and tune a model. We decided upon fitting a lasso penalized regression with the response variable being the observed metric (maximum or minimum temperature) on a given day at a site from the NOAA data and the response variables being that observed metric at each of the 113 Data Expo sites. This technique selects a small subset of around five to nine sites whose observed weather metric was deemed as a strong enough predictor of the observed weather at the NOAA site that day. This model was then stored, and this process was used to generate a model for every day that had accurate records for each of the NOAA sites. This resulted in a collection of models that most accurately predicted the observed weather at the corresponding NOAA site on the corresponding day.

3.2 Generating and Assessing Forecasts

Now that we have models that most accurately predict observed weather at a site based off the observed weather at a subset of the 113 Data Expo sites, we need to utilize this model to generate a forecast for all the NOAA weather sites. This was done by using these models, but instead of using the observed metric at the predictor site, plug in the forecasted metric. This tuned model will then result in a valid forecast that can then be assessed for accuracy, and the accuracy can then be visualized in plots to highlight trends and relationships.

3.3 Plots



There are now a set of forecast errors categorized by maximum or minimum temperature forecast error, along with the numbers of day out the forecast was made for each of the NOAA sites and Data Expo sites that can be explored and visualized. Below are four conjoined plots that show the forecast deviation and forecast error standard deviation for the three-day and one-day forecast to show the effect days out has on the plots.

Figure 1: Comparison of the three-day and one-day maximum temperature average forecast deviations. The polygons represent the closest areas to each NCDC weather station, and the circles represent the Data Expo sites.

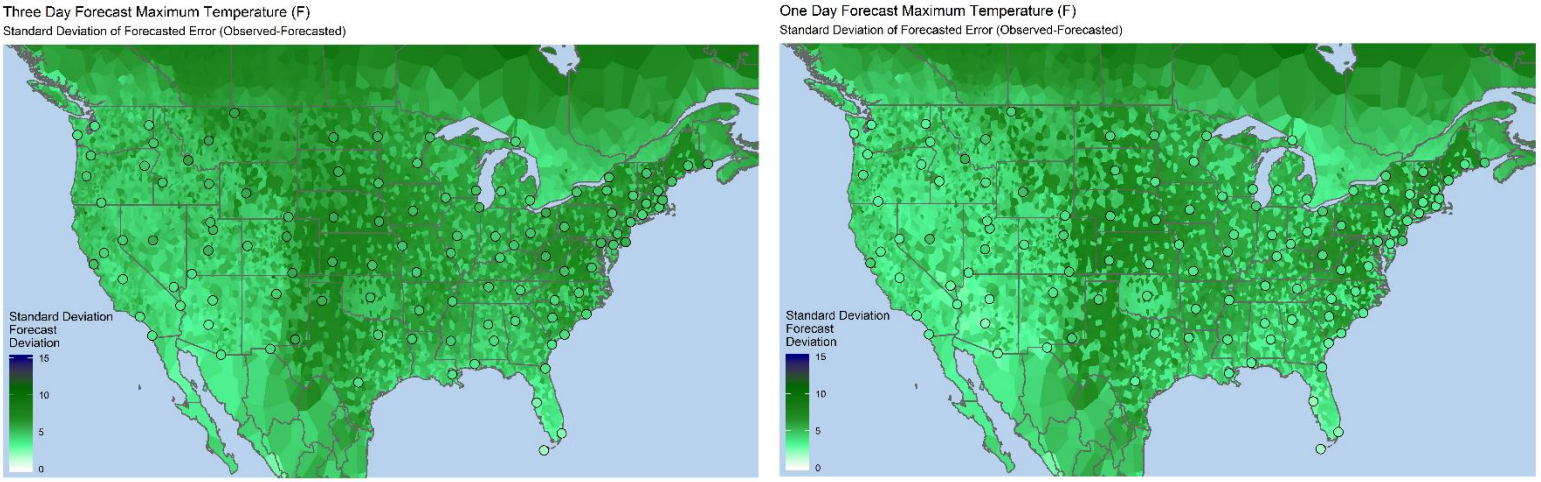


Figure 2: Comparison of the three-day and one-day maximum temperature average forecast deviation standard deviations.

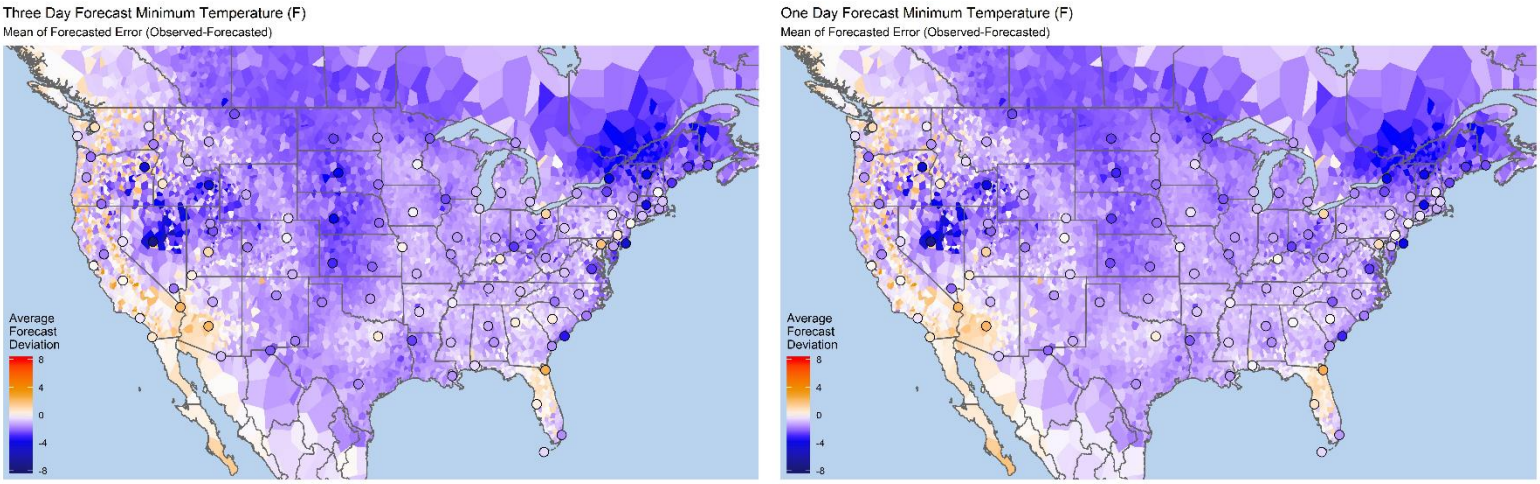


Figure 3: Comparison of the three-day and one-day minimum temperature average forecast deviations.

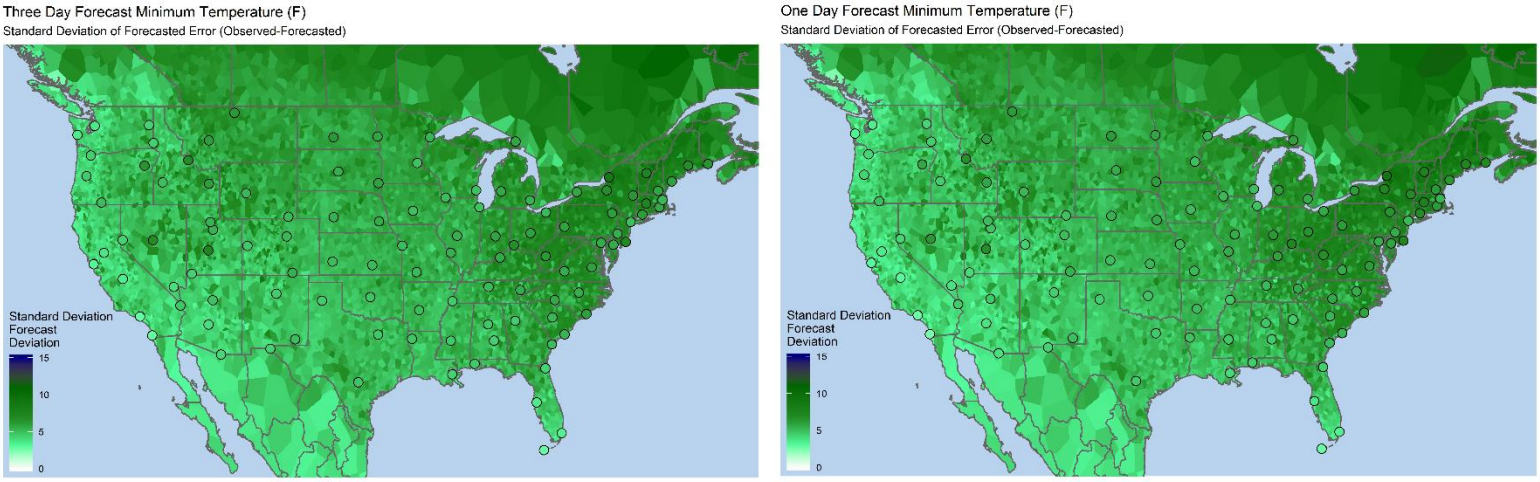


Figure 4: Comparison of the three-day and one-day minimum temperature average forecast deviation standard deviations.

4. Conclusions

In conclusion, forecasts tend to underestimate maximum temperatures across the examined section of North America,, with exceptions in the areas of Texas and Oklahoma, Washington, and into Mexico In addition, forecasts tend to overestimate minimum temperatures across the examined section of North America, with exceptions in the areas of the West coast and Florida. This results in the conclusion that overall, forecasts tend to regress to the mean. The much more intuitive conclusions that forecasts are more accurate closer to the date of forecast, and variability in forecast accuracy decreases closer to the date of forecast were also discovered.