Collection and Analysis of Weather Forecast Error Data

Sai Shreyas Bhavanasi, Harrison Lanier, Lauren Schmiedeler, and Clayton Strauch

Saint Louis University Department of Mathematics and Statistics

Project Goals

1. Acquire U.S. weather forecast and city feature data.
2. Organize the acquired data so that it contains weather observations alongside predicted values at regular time intervals.
3. Use this data to create visualizations and models that seek to explain which areas of the U.S. struggle with temperature prediction and the possible explanations for these forecasting difficulties.

Data Acquisition & Organization

The process for weather data collection is as follows:

1. Receive emails from wx-natnl@lists.illinois.edu that include weather data (forecasts and observations) for January 30, 2021 through February 1, 2022.
2. Organize the email data into an R data frame that has a similar format to that of the emails but excludes irrelevant information.
3. Create a reorganized data frame in which each row includes all of the relevant observations and forecasts for a unique (date, city) pair. Also create an expanded data frame that has eight rows for each unique (date, city) pair.
4. Add a possible error column to the expanded email data.

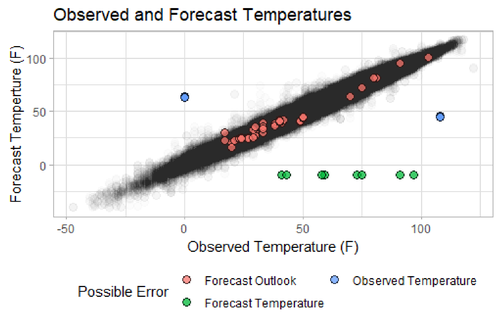
We also gathered the following information on each city for which we have weather data:

1. Longitude and latitude coordinates
2. Koppen classification
3. Elevation
4. Distance to coastline
5. Mean wind speed
6. Average annual precipitation

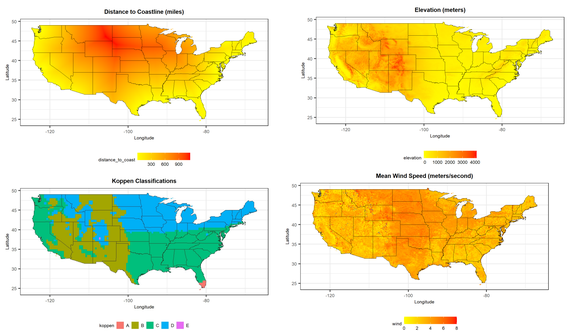
These features help us determine why some areas of the country have larger temperature forecast error than others.

Data Visualization

Below is a scatter plot of observed temperatures vs. forecast temperatures. The colored points correspond to observations that are possible errors, and the color indicates the problem variable.



Below are some maps showing the different features we collected for each city.



Statistical Tests

We perform paired t-tests to determine whether or not predictions improve as the number of hours before observation decreases. In this analysis, we use a 5% significance level and only consider the cities in the continental United States. The results for the tests examining high (top table) and low (bottom table) forecast errors follow.

Note that three numbers are reported for each test: the number of cities in which predictions (1) improved, (2) got worse, and (3) neither improved nor got worse.

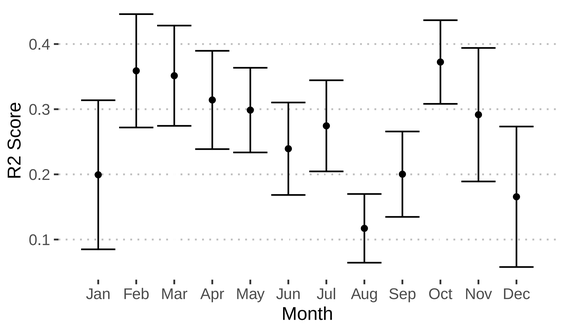
	12 hrs	24 hrs	36 hrs
48 hrs	159 0 0	146 0 13	110 0 49
36 hrs	158 0 1	110 0 49	NA
24 hrs	139 0 20	NA	NA

	12 hrs	24 hrs	36 hrs
48 hrs	141 0 18	134 0 25	56 0 103
36 hrs	121 0 38	85 0 74	NA
24 hrs	57 1 101	NA	NA

Error Modeling

To determine which predictors are the most significant in predicting the forecast error, we use stepAIC - a variable selection method that gives the best model with the fewest predictors that still has good predictive capabilities. We run this experiment at the season level and month level. We perform an 80/20 train/test split repeated 50 times. In the table below, we summarize the percentage of time each predictor was chosen.

Predictor	% of times chosen
Distance to coast	**77.0**
Latitude	70.0
Koppen classification	56.3
Elevation	37.94
Wind	31.67



R2 Score for the month

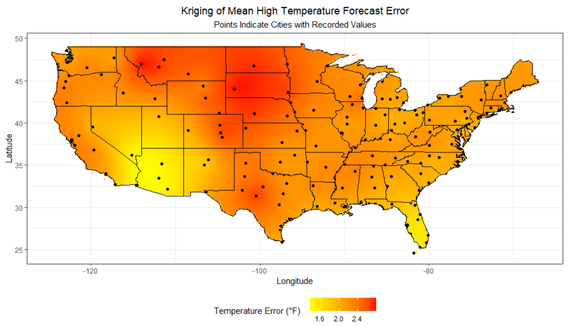
To predict the error in a specific city on a given day, we predict the error on the daily rain, season, and outlook. We observe that this model produced suboptimal results.

Kriging

Kriging is a method of interpolation used in geostatistics. The method uses the preexisting data to create a variogram model of temperature error. A variogram (γ) is a function that describes the spatial dependence of a variable in a space. There are several different models used, but this project utilized the exponential model given by the equation:

$$\gamma(h) = b + C_0 * \left(1 - e^{-\frac{h}{a}}\right) \text{ where } \begin{cases} b & \text{Nugget} \\ C_0 & \text{Sill} \\ a & \text{Range} \\ h & \text{Distance} \end{cases}$$

Through this, the mean temperature error of any location in the U.S. without recorded data can be predicted by its neighbors based on their spatial proximity.



Project Conclusions

1. Temperature forecast errors generally improve over time, and this improvement is particularly obvious when comparing prediction errors corresponding to predictions that were made further apart in time.
2. Latitude and distance to coast are both important features for predicting temperature forecast errors.
3. Because of the lack of such a data set in the academic space, the data we have produced and managed is a valuable contribution for further study to be conducted.