Let's Talk About the Weather

Jill Lundell, Brennan Bean, Jürgen Symanzik - Department of Mathematics and Statistics, Utah State University

Motivation

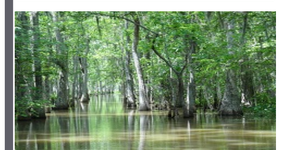
How do different regions of the United States experience forecast error? (click on images or tabs for details)

Clusters

Cali-Florida

Intermountain West

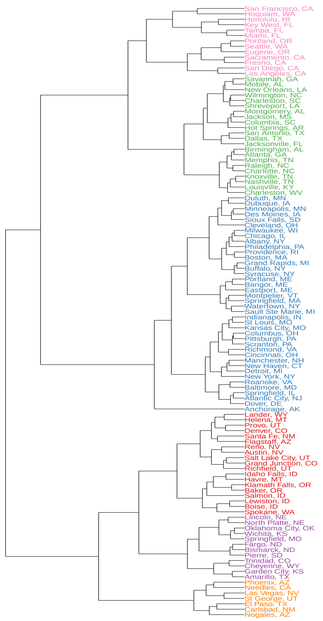
Southeast

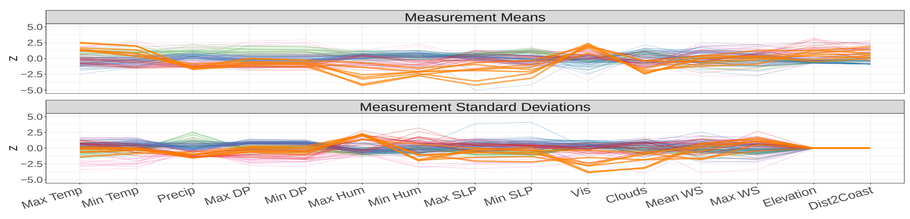


Midwest

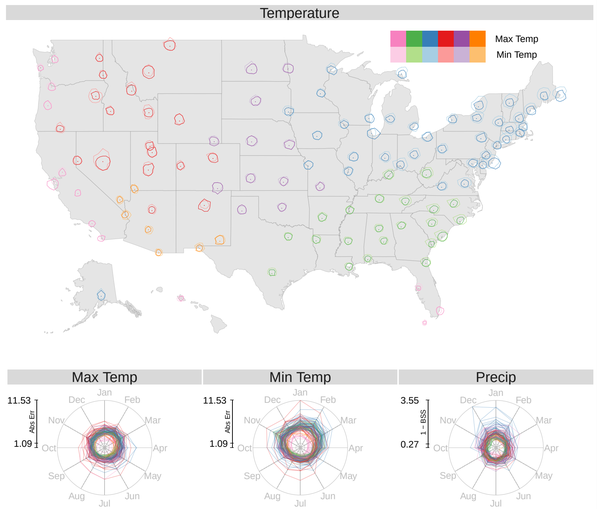
Northeast

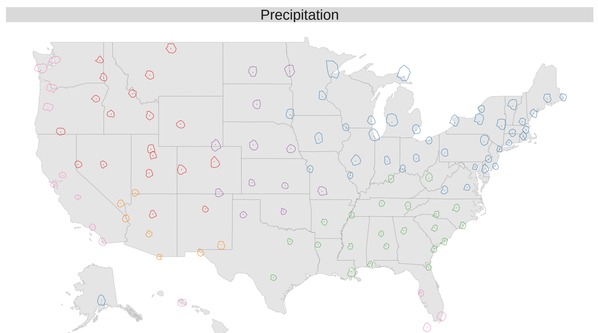
Southwest



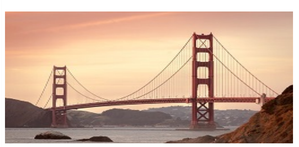


Seasonality





San Francisco, CA

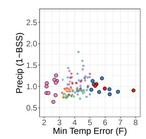


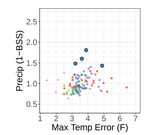
Great Lakes

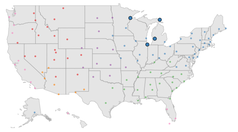
Trends & Outliers

Cali-Florida

New England

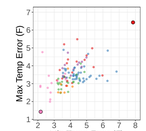




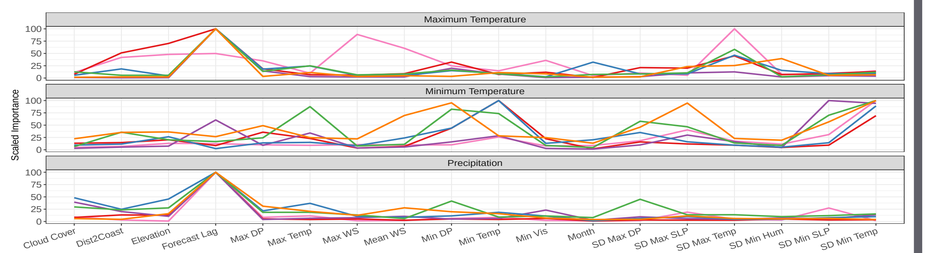


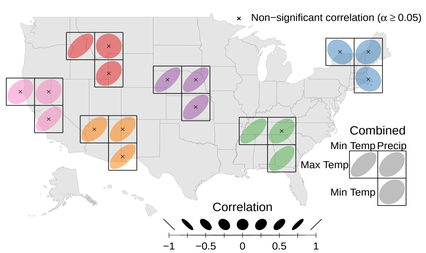
Key West, FL

Austin, NV



Importance & Correlations





Conclusions

References

R Tools

Can we cluster U.S. weather stations into regions based on their measurements?

Cali-Florida

Warm and humid with high dew point and pressure. Low variability in almost all measurements.

Intermountain West

Cold and dry, with high variability in temperature, wind speed, and pressure. Low variability in precipitation and dew point.

Southeast

Warm and humid with lots of rain. High variability in precipitation and low variability in temperature.

Midwest

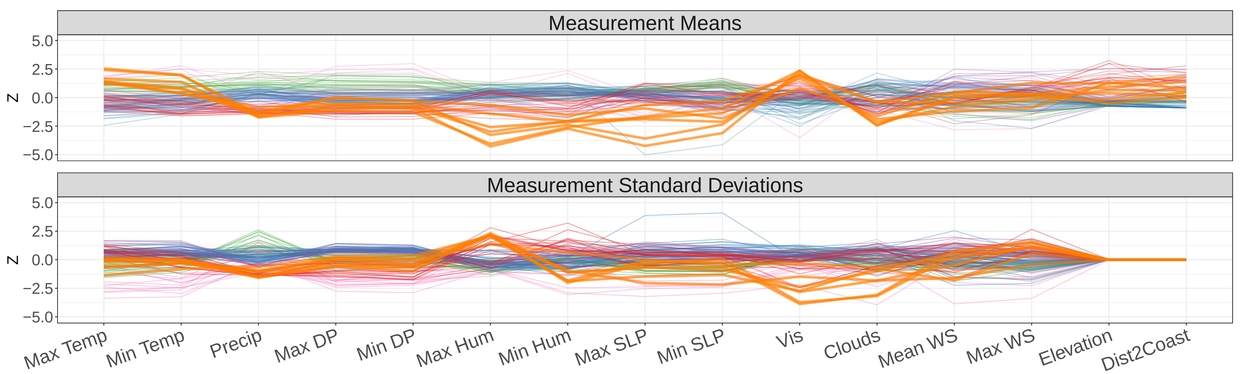
Landlocked with high wind speed and high variability in temperature, pressure, and wind speed.

Northeast

Cold, humid, and low visibility. High variability in temperature, dew point, and pressure.

Southwest

Warm, sunny, and dry with little variation. High variability in wind speed and humidity.

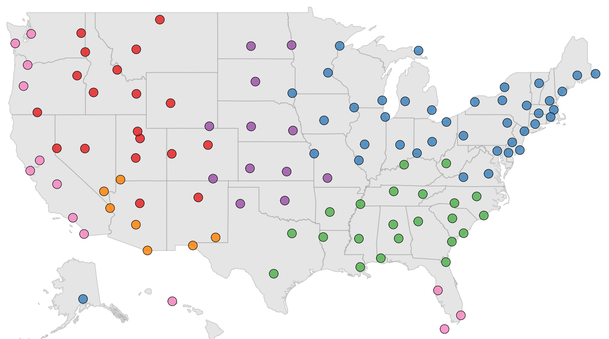


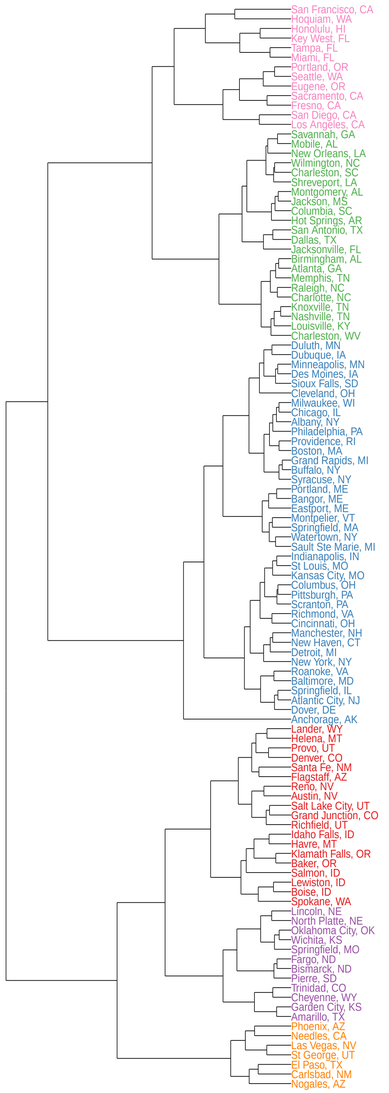
These parallel coordinate plots highlight the normalized table values for all locations in the **Southwest** cluster, or region. Please **see the app** (motivated by [1]) to explore other regions.

Data

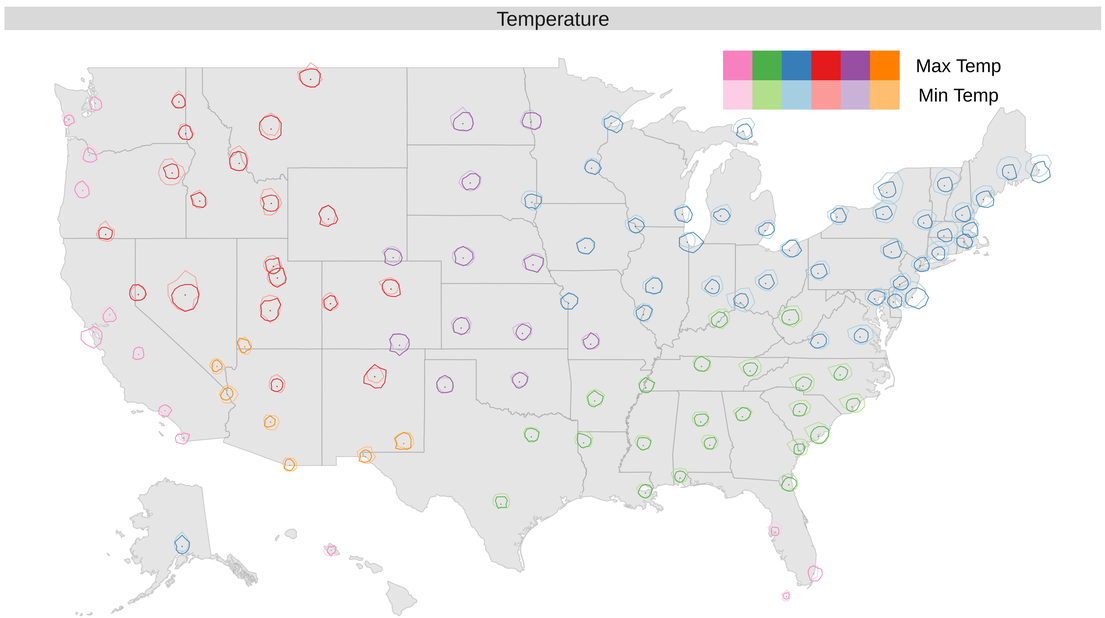
The data contain measurements for 113 United States (U.S.) weather stations from July 2014 to September 2017. We compute the mean and standard deviation of all daily measurements at each location for each variable in the table below. We create clusters using **Ward's method** after calculating the **Euclidean distance** between the z-scores of each measurement.

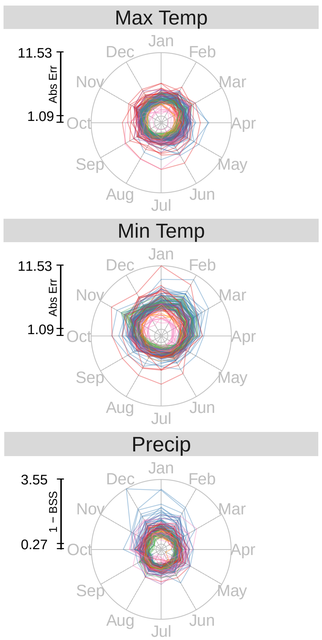
Weather Variables		
Max Temperature (Temp)	$°F$	Min Temperature
Max Dew Point (DP)	$°F$	Min Dew Point
Max Humidity (Hum)	%	Min Humidity
Max Sea Level Pressure (SLP)	inHg	Min Sea Level Pressure
Max Wind Speed (WS)	mph	Mean Wind Speed
Max Visibility (Vis)	mi	Distance to Coast* (Dist2Coast)
Precipitation (Precip)	in/ft	Elevation*
Cloud Cover (Clouds)	0-8	* variables downloaded manually

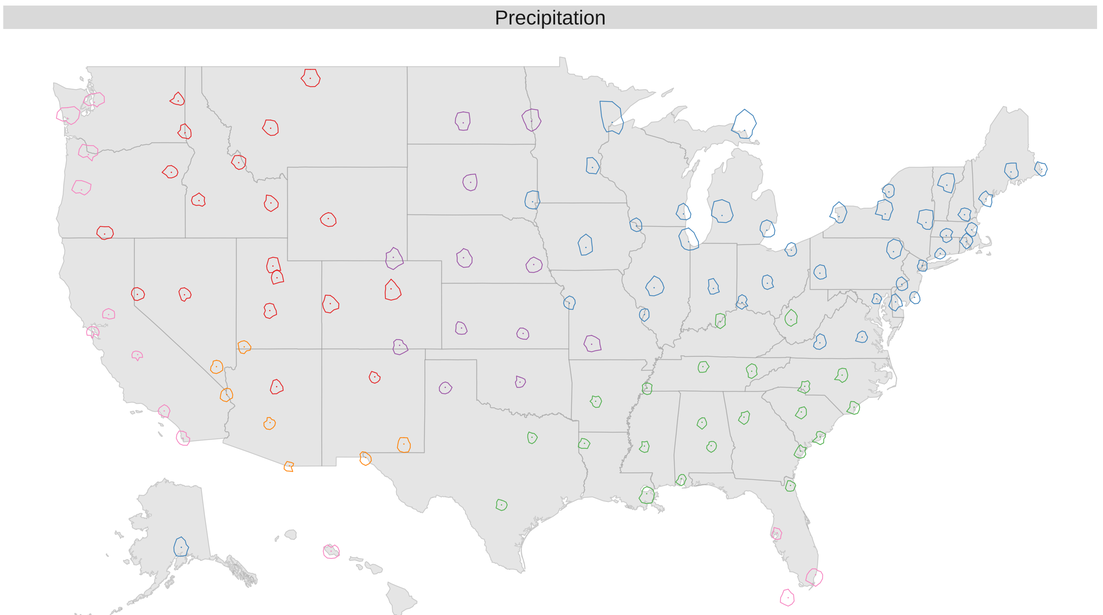




How does forecast error change by cluster and by season?







Glyph Plots

These **glyph plots** [2] show forecast error (see Forecast and Error Variables) averaged over lag and month as the scaled distance from point to edge. The asymmetry of the glyphs reveal seasonality in the forecast errors, something prevalent in the `Min Temp` and `Precip` forecasts of the Northeast, where forecast errors are worse in the winter. In contrast, the entire Southeast region predicts `Precip` consistently well throughout the year.

San Francisco, CA

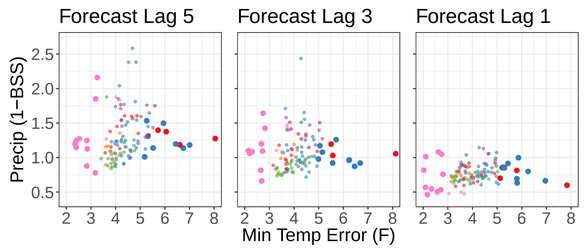
San Francisco, CA, does well predicting `Min Temp`, like other cities in the region. However, the city is also known for chilling coastal fogs in the summer that create distinct micro-climates over very short distances [3]. This likely explains why San Francisco, CA, struggles to predict `Max Temp` during the summer months.

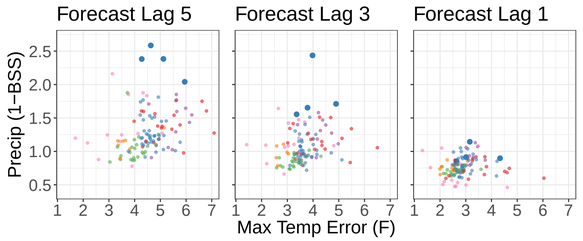
Great Lakes

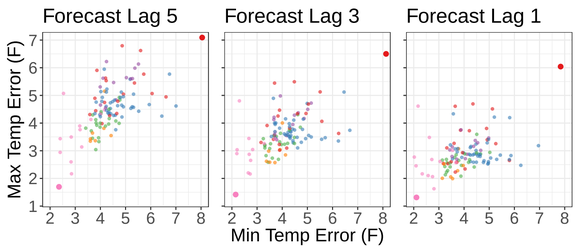
The poor `Precip` forecast accuracy of this region in the winter illustrates the difficulty in forecasting lake-effect snow. Up to 100% more snow falls downwind of Lake Superior in the winter than would be expected without the lake-effect [4]. This area has also been previously labeled as having the most unpredictable precipitation patterns in the nation [5].

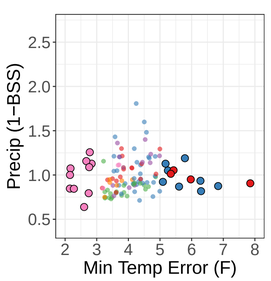
Cali-Florida | Southeast | Northeast | Intermountain West | Midwest | Southwest

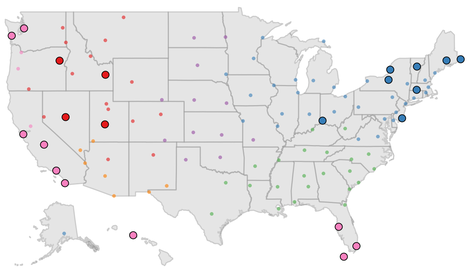
Who are the winners and losers in terms of overall forecast accuracy?

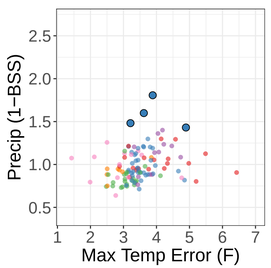


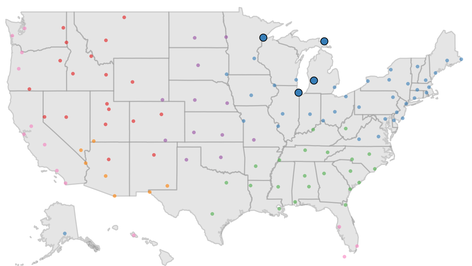


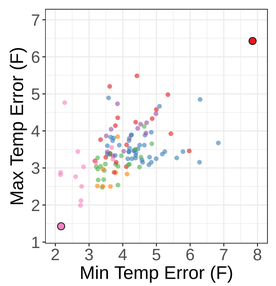


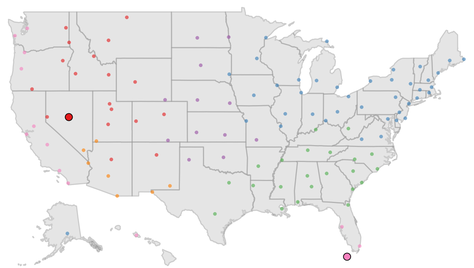












Scatterplot Error

These plots highlight locations with the best and worst forecasts across all three dimensions of error (see **Forecast and Error Variables**). Scatterplots show overall error, as well as error for three selected forecast lags. Interesting points brushed in the scatterplots are highlighted on the maps. **See the app** to interact with the data for all forecast lags.

Cali-Florida

Low variability in daily temperatures perhaps explains why the lowest **`Min Temp`** forecast errors are composed entirely of observations from this region.

Key West, FL, has the lowest overall forecast error for both **`Max Temp`** and **`Min Temp`**. Not surprisingly, this location ranks in the top five for lowest variability in 8 of the **Weather Variables**.

Northeast

New England is known for extreme winter weather and the frequency of extreme weather events seems to be increasing [6]. This likely contributes to the struggle these stations have predicting **`Min Temp`**.

Poor **`Precip`** predictions in the **Great Lakes** region are highlighted again in these scatterplots (see **Seasonality**), but improve rapidly as forecast lag decreases.

Austin, NV

Seventy miles along the "loneliest highway in America" [7] separate this city from its weather measurements in Eureka, NV. The poor predictions for **`Max`** and **`Min Temp`** can be explained by the change in climate over such a large distance, as reflected in a negative prediction bias of around 5°F for **`Max Temp`** and a positive bias of around 7°F for **`Min Temp`**.

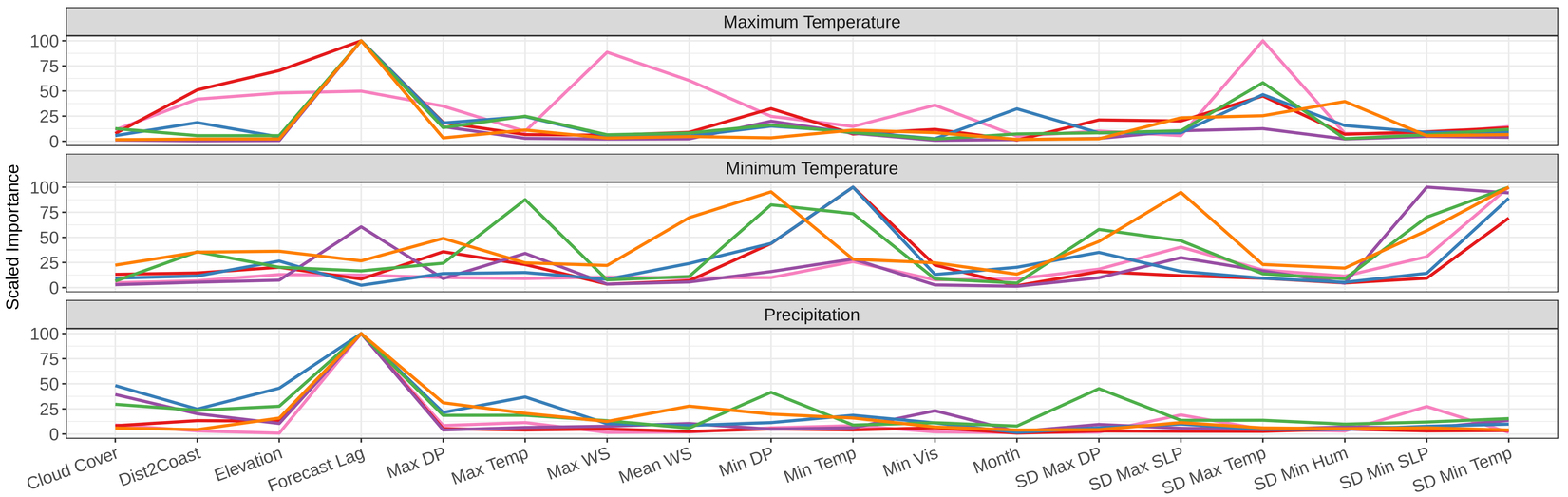
Which variables are important in determining forecast error? How do error variables correlate?

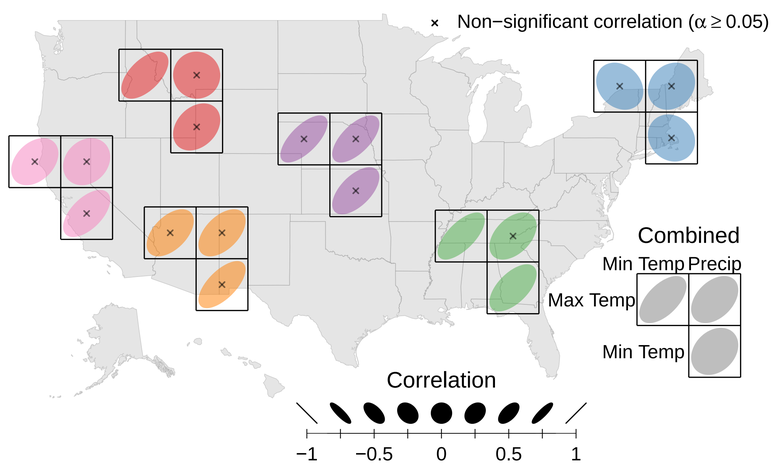
Variable Importance

Random forests variable importance measures the percent increase in mean square error (%incMSE) that results when the information from a variable is removed from the model. We normalize this measure of importance for each of the Weather Variables and use each forecast error as the response. This parallel coordinates plot displays all variables that ranked top three in importance for at least one region and forecast combination.

Forecast lag is the most important variable when predicting **`Precip`** forecast error and most important when predicting **`Max Temp`** forecast error except in Cali-Florida. In contrast, **`Min Temp`** and **`SD Min Temp`** are often most important in predicting **`Min Temp`** forecast error. Other variables such as **`Max Temp`**, **`SD Max DP`**, and **`SD Max SLP`** are only important in the Southwest and Southeast.

Forecast and Error Variables		
Max Temp	$^\circ F$	Absolute Error
Min Temp	$^\circ F$	Absolute Error
Forecast Precip	%	1 − Brier Skill Score (BSS) [8]
Lag		Days out from forecast





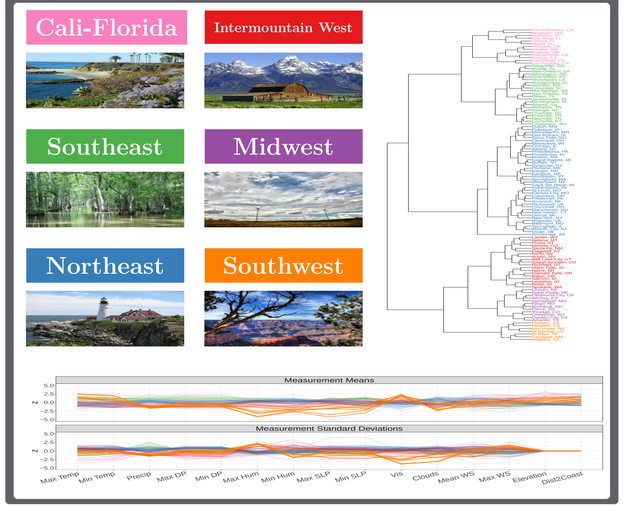
Correlation

We compute the **Spearman correlations** among each of the three forecast measurements within each cluster. The strength and significance of these correlations are visualized as ellipses [9] on the map. These ellipses show a general positive correlation for each forecast error except in the Northeast. Significant correlations are only found in the Southeast and Intermountain West.

What did we learn?

The United States cleanly clusters into well defined weather regions. Patterns in forecast errors are closely related to the unique climates that characterize each region.

Clusters



Climate patterns in the United States cleanly separate into at least **six recognizable regions** through a cluster analysis using the means and standard deviations of the Weather Variables.

Random forests variable importance identify the variables most important in predicting forecast error. While many important variables are common across all clusters, certain variables are only important within specific regions.

Strong **correlations** among the error variables in the **Southeast** are not found in the **Northeast**. Overall, the error variables are positively correlated.

Importance & Correlations

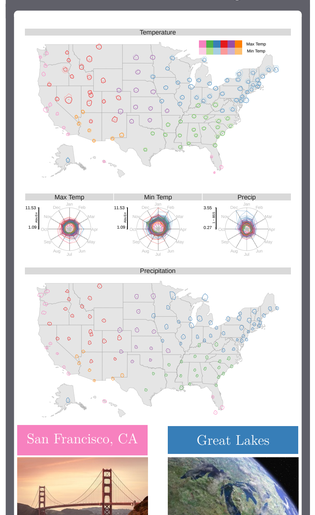
R Tools

- fields (D. Nychka et al. 2015)
- fiftystater (W. Murphy 2016)
- geosphere (R. Hijmans, 2016)
- ggforce (T. Pedersen 2018)
- gridExtra (A. Baptiste, 2017)
- latex2exp (S. Meschiari, 2015)
- mapproj (D. McIlroy et al. 2017)
- RColorBrewer (W. Neuwirth, 2014)
- randomForest (A. Liaw and M. Wiener, 2002)
- reshape2 (H. Wickham, 2007)
- rgbif (S. Chamberlain, 2017)
- rgdal (R. Bivand et al. 2018)
- sp (E. Pebesma and R. Bivand, 2013)
- tidyverse (H. Wickham, 2017)
- weatherData (R. Narasimhan, 2017)

Trends & Outliers

Cali-Florida excels in predicting **Max** and **Min Temp**, likely due to low temperature variability throughout the year. **New England** struggles to predict **Min Temp** in the winter. The worst predictions are in **Austin, NV**, and can be attributed to the large distance between forecast and measurement locations.

Seasonality



Glyph plots visualize the differences in forecast error seasonality across regions. This seasonality is best seen in the **Northeast**, where the **Great Lakes** struggle predicting **Precip** in the winter. Conversely, **San Francisco, CA**, struggles to predict **Max Temp** in the summer.

References

[1] A. Unwin, "Requirements for interactive graphics software for exploratory data analysis," *Computational Statistics*, vol. 14, no. 1, pp. 7–22, 1999.

[2] H. Wickham, H. Hofmann, C. Wickham, and D. Cook, "Glyph-maps for visually exploring temporal patterns in climate data and models," *Environmetrics*, vol. 23, no. 5, pp. 382–393, 2012.

[3] C. Nolte, "The story of the San Francisco summer is a bit foggy." https://www.sfchronicle.com, August 2016.

[4] R. W. Scott and F. A. Huff, "Impacts of the Great Lakes on regional climate conditions," *Journal of Great Lakes Research*, vol. 22, no. 4, pp. 845–863, 1996.

[5] N. Silver and R. Fischer-Baum, "Which city has the most unpredictable weather?." https://fivethirtyeight.com, December 2014.

[6] J. Cohen, K. Pfeiffer, and J. A. Francis, "Warm Arctic episodes linked with increased frequency of extreme winter weather in the United States," *Nature Communications*, vol. 9, no. 1, p. 869, 2018.

[7] "Austin, Nevada: So much to do.." http://austinnevada.com.

[8] A. P. Weigel, M. A. Liniger, and C. Appenzeller, "The discrete Brier and ranked probability skill scores," *Monthly Weather Review*, vol. 135, no. 1, pp. 118–124, 2007.

[9] D. Murdoch and E. Chow, "A graphical display of large correlation matrices," *The American Statistician*, vol. 50, no. 2, pp. 178–180, 1996.