Computational Statistics (2023) 38:1135–1155
https://doi.org/10.1007/s00180-023-01339-3

ORIGINAL PAPER

Let's talk about the weather: a cluster-based approach to weather forecast accuracy

Jill F. Lundell[1] · **Brennan Bean**[1] · **Jürgen Symanzik**[1]

Received: 15 April 2019 / Accepted: 22 February 2023 / Published online: 30 March 2023

Abstract
Improved understanding of characteristics related to weather forecast accuracy in the United States may help meteorologists develop more accurate predictions and may help Americans better interpret their daily weather forecasts. This article examines how spatio-temporal characteristics across the United States relate to forecast accuracy. We cluster the United States into six weather regions based on weather and geographic characteristics and analyze the patterns in forecast accuracy within each weather region. We then explore the relationship between climate characteristics and forecast accuracy within these weather regions. We conclude that patterns in forecast errors are closely related to the unique climates that characterize each region.

Keywords Climate · Clustering · Data Expo 2018 · Glyph plots · Random forests · Visualization

1 Introduction

From the icy, wet winters along the Great Lakes, to the hot and dry summers in the Southwest, the United States (U.S.) experiences a wide range of climatic extremes. These extremes create unique challenges when forecasting the weather. Understanding forecast errors across such a diverse landscape is equally challenging, requiring multi-dimensional visualizations across space, time, and climate measurements. Better understanding of the nature and patterns in forecast errors across the U.S. helps meteorologists as they strive to improve weather forecasts. It can also help

✉ Jill F. Lundell
jflundell@gmail.com

Brennan Bean
brennan.bean@usu.edu

Jürgen Symanzik
Juergen.Symanzik@usu.edu

[1] Department of Mathematics and Statistics Utah State University, Logan, USA

everyday Americans know how much faith to put in the weather forecast on the day of an important event.

The 2018 Data Expo of the Sections on Statistical Computing and Statistical Graphics of the American Statistical Association (ASA) provided an opportunity to explore and compare weather forecast errors across the U.S. Our analysis focused on the question:

How do weather forecast errors differ across regions of the U.S.?

This motivating question prompted the subsequent questions:

- Do U.S. weather stations cluster into regions based on weather characteristics?
- How do error variables correlate and do these correlations change by region?
- How do forecast errors change by region and by season?
- Where are the best and worst forecast accuracies?
- Which variables are important in determining forecast errors?

Preliminary results of our analysis are published in the proceedings for the 2018 Joint Statistical Meetings (Lundell et al. 2018).

This article is devoted to answering these questions. We use ensemble graphics to create an overall picture of weather forecast errors across different regions of the U.S. (Unwin and Valero-Mora 2018). Ensemble graphics enhance traditional analyses by connecting several visualizations of the data with adjoining text. This presentation is able to tell a cohesive story of the data more effectively than would be possible with a few disjointed graphics. In Sect. 2, we summarize the data and then show that the U.S. can be clustered into six well-defined weather regions using the provided climate measurements, elevation, and distance to coast. These clusters, or weather regions, form the basis of our comparison of forecast accuracy across the U.S. through a series of multi-dimensional plots and variable importance analyses described in Sect. 3. In Sect. 4, we introduce the interactive application we created to enhance our data explorations. We conclude in Sect. 5 that the climate differences that distinguish the weather regions of the U.S. also create region-specific patterns and differences in forecast accuracy. Two appendixes are included at the end of this paper to explain data cleaning and how to create the glyphs used in this article.

2 Weather regions

The data contain measurements and forecasts for 113 U.S. weather stations from July 2014 to September 2017. Information about the data and other analyses done with the data can be found in Cetinkaya-Rundel and Martinez (2023). These data can be obtained from our supplemental materials or at the following URL:

https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2018.

Daily measurements for eight different weather metrics were recorded for each location including temperature, precipitation, dew point, humidity, sea level pressure, wind speed, cloud cover, and visibility. Many notable weather events are also

textually recorded such as thunderstorms and fog. Daily measurements of the minimum, maximum, and mean were recorded for each metric. Weather characteristics used in this article are listed in Table 1. Data were supplemented with some geographic information and carefully examined and cleaned. Details on data cleaning, obtaining additional data, and the justification behind our final variable selection are found in Appendix A.

2.1 Developing weather clusters

The U.S. has been divided into regions based on environmental characteristics such as watersheds and climate (Commission for Environmental Cooperation 1997, Briggs et al. 2003). We examined the set of existing environmental regions and were unable to find one that made sense in terms of weather in the context of this analysis. We created our own weather regions by clustering the weather stations based on the metrics in Table 1. Thus, clusters are defined by weather characteristics observed at each station. We use these clusters to determine how weather forecast error patterns are related to the unique climate measurements of a particular region. A review of existing weather regions and how they correspond to our weather regions is discussed in Sect. 2.2. Data were aggregated across each weather station by taking the mean and standard deviation of each variable in Table 1 for each of the 113 weather stations over the period of record.

Hierarchical clustering (Hastie et al. 2001, pp. 520-526) with Euclidean distance and Ward's minimum variance clustering method (Murtagh and Legendre 2014) was used to identify clusters. The clusters were examined spatially to determine the performance of the clustering method and select the final number of clusters. We wanted to ensure the weather station clusters were of a sufficient size to be practical. Five clusters resulted in one cluster that included all of the stations from the Midwest to the East Coast which we think is too large because of the differences in coastal and inland climates. Seven clusters produced a cluster that contained only

Table 1 Weather variables included in our analysis. All observations outside the indicated ranges were removed prior to our analysis

Variable	Unit	Abbreviation	Range
Min/Max temperature	Fahrenheit	$^{\circ}F$	$[-37, 127]$
Precipitation	Inches	in	$[0, 12.95]$
Min/Max dew point	Fahrenheit	$^{\circ}F$	$[-50, 90]$
Min/Max humidity	Percent	%	$(0, 100]$
Min/Max sea level pressure	Inch of mercury	inHg	$[28.2, 31.2]$
Mean/Max wind speed	Miles per Hour	mph	$[0, 70]$
Min visibility	Miles	mi	$[0, 10]$
Cloud cover	Number of Eighths of the sky covered	okta	$\{0, 1, \ldots, 8\}$
Distance to coast	Miles	mi	$[0, 807]$
Elevation	Feet	ft	$[3, 7422]$

five weather stations which is too small. Thus, we chose six clusters to divide the U.S. into weather regions.

Figures 1 and 2 show the results of the cluster analysis. Figure 3 shows a parallel coordinate plot of the characteristics for each weather region. The Z-score for mean and standard deviation for each of the variables in Table 1 was computed and plotted on the parallel coordinate plot. It is difficult to distinguish the six weather regions from each other so an interactive app was created that provides a better view of the features of each cluster. The app is discussed in Sect. 4.

The names and characteristics of each weather cluster are as follows:

- *Cali-Florida* (13 stations): Warm and humid with high dew point and pressure. Low variability in almost all measurements.
- *Southeast* (22 stations): Warm and humid with lots of rain. High variability in precipitation and low variability in temperature.
- *Northeast* (39 stations): Cold, humid, and low visibility. High variability in temperature, dew point, and pressure.
- *Intermountain West* (19 stations): Cold and dry, with high variability in temperature, wind speed, and pressure. Low variability in precipitation and dew point.
- *Midwest* (13 stations): Landlocked with high wind speed and high variability in temperature, pressure, and wind speed.
- *Southwest* (7 stations): Warm, sunny, and dry with little variation in temperature or precipitation. High variability in wind speed and humidity.

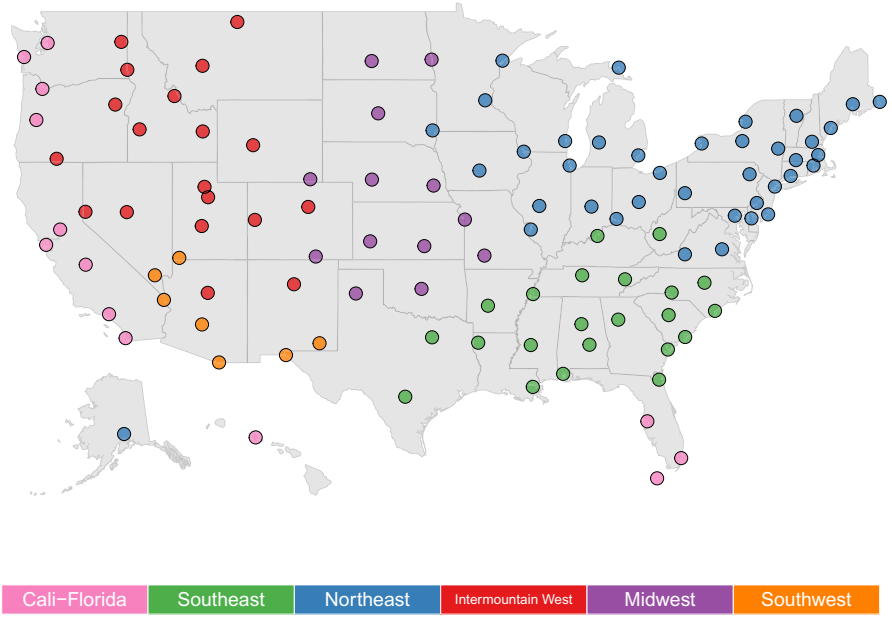

Fig. 1 Map of the six weather regions. The color band at the bottom identifies each region by name and color

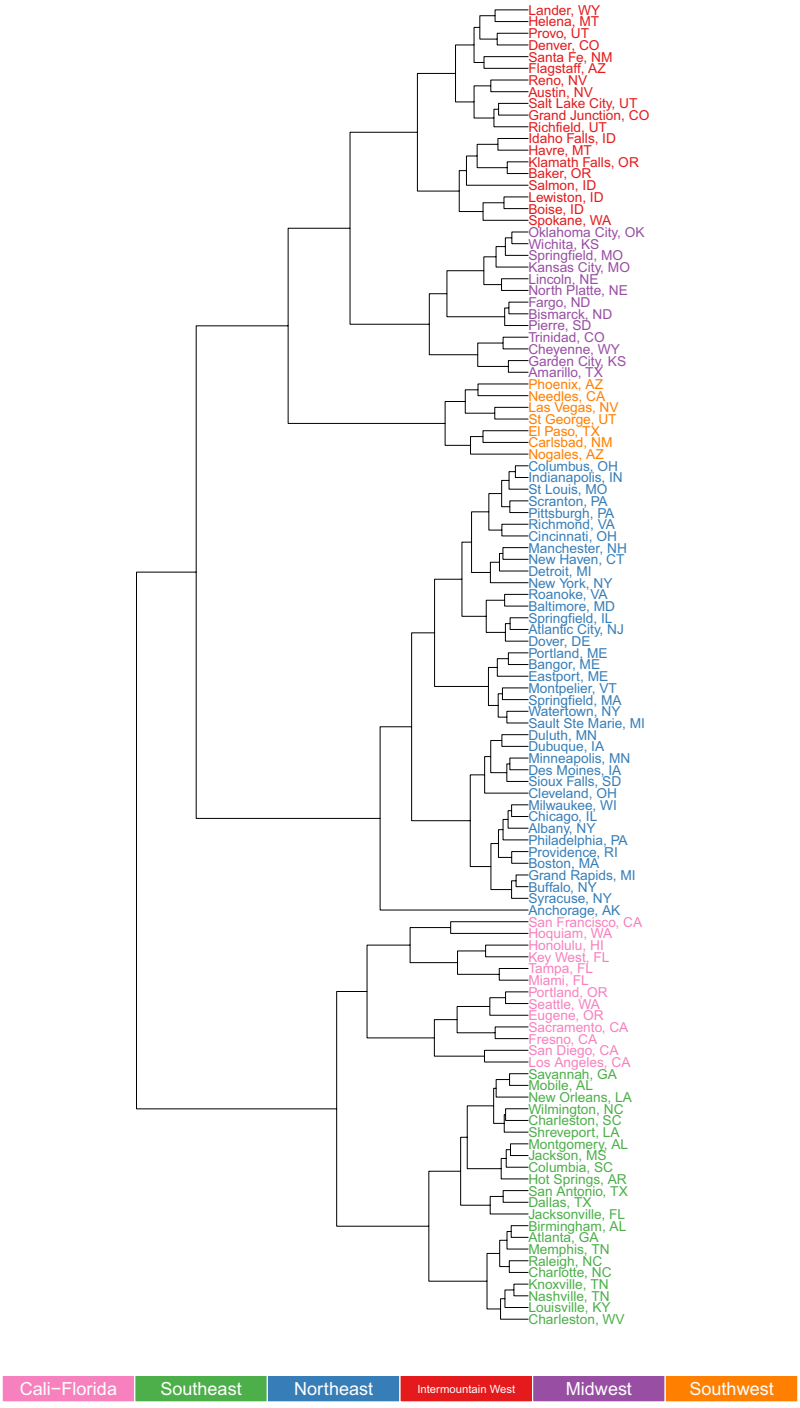

Fig. 2 Dendrogram of weather clusters identified in Fig. 1

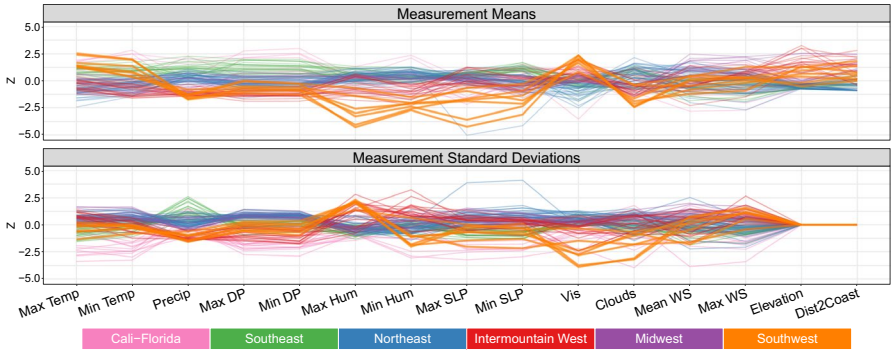

Fig. 3 Parallel coordinate plot of the means and standard deviations of the weather variables listed in Table 1. Each line in the plot represents one of the 113 weather stations. The color of the lines match the weather region to which the station belongs. An interactive app is available that allows for better identification of regional trends.The Southwest region is highlighted in this graph to emphasize its weather characteristics

2.2 Comparison to existing climate regions

Ecological and climate regions have been developed for the U.S. in other studies. Many of these studies focused on smaller regions in the U.S., but a few have looked at the U.S. as a whole. Clustering methods and the variables used to identify clusters differ from study to study. The ecological regions of North America defined by the Commission for Environmental Cooperation (Commission for Environmental Cooperation 1997) used ecosystems to develop regions. Air, water, land, and biota, including humans, were used to create the ecoregions. These ecoregions show a strong longitudinal trend that corresponds well with the longitudinal trends in our clusters. Clusters were not determined by statistical clustering methods, but by careful assessment of ecological properties across North America.

The National Oceanic and Atmospheric Administration (NOAA) developed climate regions that incorporate seasonal temperature and precipitation information (Karl and Koss 1984). These regions differ substantially from the North American ecological regions as they also have a lateral trend in addition to the longitudinal trend and are constrained by state boundaries. Spectral curves assessing drought and wet spells were used to define the NOAA regions (Diaz 1983). The NOAA regions correspond roughly to our general weather regions despite region borders being defined by state boundaries. The north/south division in the eastern U.S. closely aligns with our cluster division in that area. The major east/west division in our clusters is in a similar location to the NOAA clusters as well.

The International Energy Conservation Code (IECC) climate clustering of the U.S. (Briggs et al. 2003) and subsequent reclassification by Hathaway et al. (2013) divided the U.S. into fourteen regions based on temperature, dew point, wind speed, and radiation. Cluster methods included K-means clustering and Monte-Carlo

sifting. Monte-Carlo sifting is a method developed by Hathaway et al. (2013) that identifies a candidate and dropout reference cell by iteratively searching through a set of Monte-Carlo runs. Both sets of regions show a strong lateral trend in the Eastern U.S. These regions also show distinct separation of the West coast and Southwest deserts from the rest of the Western U.S. Similar trends are also seen in our clusters. The lateral trend in the Eastern U.S. is not as strong in our clusters, but this is likely because we chose a smaller number of weather clusters. The inclusion of additional variables insensitive to lateral trends such as distance to coast, elevation, and humidity, all serve to reduce the lateral separation in our clusters.

One key difference between our weather regions and the regions seen in other studies is that we combine Florida and the Pacific coast into a single weather region. This is likely a result of our choice to omit geographic proximity of weather stations in the cluster analysis calculations and consider only similarities in weather patterns. Both Florida and the Pacific coast experience less seasonality in their weather patterns than the rest of the country. This results in smaller than average standard deviations for many of the climate variables in both of these regions. These small standard deviations create a measure of closeness between Florida and the Pacific coast, which likely explains why these two geographic areas fall into a single cluster when working with six or fewer clusters. The Florida and Pacific stations split into separate clusters when using seven clusters with exception of two stations from the Pacific Coast that cluster with the Florida stations. Hawaii and Alaska are either ignored in the literature or placed in their own regions. Because we did not use spatial proximity as a clustering variable and we assigned all weather stations to one of our six weather clusters, Hawaii and Alaska are clustered with Cali-Florida and the Northeast respectively. Our clusters show that weather patterns typically have strong spatial correlations, with temperate coastal regions being a notable exception.

3 Forecast error explorations

Given the clear separation of the country into distinct weather regions, we seek to determine if there are clear differences in forecast error patterns among the regions. Forecasts were restricted to minimum temperature, maximum temperature, and the probability of precipitation. The forecast error for minimum and maximum temperature is calculated as the absolute difference between forecast and measurement. The forecast error for precipitation is measured using the Brier Skill Score (BSS), a well-known measure of probabilistic forecast accuracy (Weigel et al. 2007). It is defined for a particular weather station as

$$\mathrm{BSS} = 1 - \frac{\sum_{i=1}^{N}\sum_{j=0}^{M}\left(Y_{ij} - O_i\right)^2}{\sum_{i=1}^{N}\sum_{j=0}^{M}\left(P - O_i\right)^2} \tag{1}$$

where $Y_{ij} \in [0, 1]$ is the predicted probability of rain on day i with forecast lag j; $O_i \in \{0, 1\}$ is a binary variable with value 1 if *any* precipitation fell during the day and 0 otherwise. We define a precipitation event as a positive precipitation

measurement or the inclusion of the words “rain" or “snow" in the event information; $P \in [0, 1]$ is the average daily chance of precipitation over the period of interest, defined as $P = \frac{1}{N} \sum_{i=1}^{N} O_i$; N denotes the number of days of recorded precipitation in the period of record and $M \in \{0, \ldots, 5\}$ denotes the number of forecast lags.

Note that the BSS $\in (-\infty, 1]$, with 1 indicating a perfect forecast skill and movement towards $-\infty$ indicating worse forecasts. We chose to use $1 -$ BSS so all three error variables are consistent in orientation. The following subsections explore differences in forecast errors both between and within the previously defined weather regions visualized in Fig. 1. Forecast errors are averaged over lag and in some cases averaged over month in each graph. The visualizations in the following subsections confirm our hypothesis that different weather regions experience distinctly different weather forecast error patterns.

3.1 Error correlations

Are the forecast errors for the three different measurements (i.e., minimum temperature, maximum temperature, and precipitation) correlated with each other? How do these relationships change between the different weather regions? We explore such correlations through the use of correlation ellipses (Murdoch and Chow 1996) superimposed on a map of the U.S. in Fig. 4. We calculated Spearman correlations between each pair of measurements for the locations within each cluster. The sign of the correlation coefficient is denoted by the slope of the ellipse and the strength of correlation is denoted by the width of the ellipse.

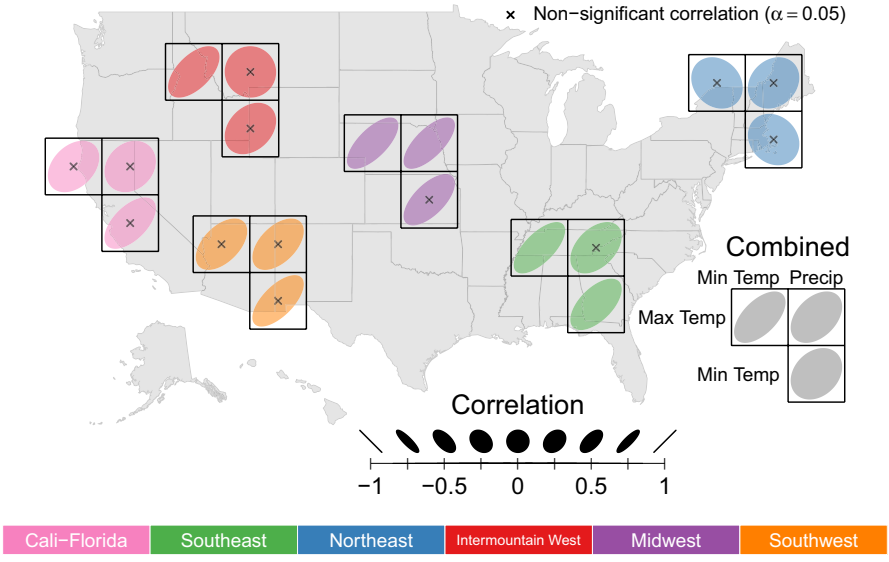

Fig. 4 Spearman correlations between forecast error variables represented as ellipses superimposed on a map of the United States. The p value for each correlation is compared against a 0.05 level of significance

All of the correlations between error variables are positive except for correlations between minimum temperature and the other two variables in the Northeast. The strongest relationships are seen in the Midwest, the South and the Southwest. The weakest relationships are found in the Northeast. Only a few cluster-specific correlations are significant. This is likely due to the small number of stations in many of the weather regions. However, the overall correlations for the 113 weather stations are all positive and significant. This indicates that areas with good predictions for one forecast variable have generally good predictions for the other forecast variables as well. The weakest correlations are between minimum temperature and precipitation predictions. Although there are relationships between the three weather forecast variables, those relationships are not particularly strong and the strength differs within each region. The observations made using this correlation ellipse map illustrate how this plot style facilitates multi-dimensional comparisons across space. Information on the calculations and implementation of the correlation glyphs can be found in Appendix B.

3.2 Error scatterplots

Scatterplots reveal outliers and overall trends within weather regions and across forecast lag. Forecast lag is defined as the number of days between the day of forecast and the day being forecast. Thus, same day forecasts would have a lag of 0, one day prior forecasts a lag of 1, and so on. Because we are comparing three variables spatially and temporally across the U.S., static graphs are not optimal for assessing all relationships of interest. We constructed an interactive scatterplot app using Shiny (Chang et al. 2019) that facilitates examination of trends between the three forecast error variables aggregated across all forecast lags or for individual forecast lags. Figure 5(a–c) shows examples of plots from the interactive app. The figure shows the scatterplot for the data aggregated over all forecast lags, as well as the scatterplots for lags of 5, 3, and 1, to illustrate how forecast accuracy changes over forecast lag.

Figure 5a compares minimum temperature forecast accuracy with precipitation accuracy. Weather stations with the worst predictions of minimum temperature are located in New England and the Intermountain West. New England is known for extreme winter weather and the frequency of extreme weather events seems to be increasing (Cohen et al. 2018). This likely contributes to the struggle these stations have predicting minimum temperature. The worst predictor of minimum temperature is Austin, Nevada. This location is addressed further in Fig. 5c. Cali-Florida uniformly has the best predictions of minimum temperature. However, Cali-Florida also has some of the greatest variability in precipitation prediction accuracy when examining individual lags.

Figure 5b compares maximum temperature prediction accuracy with precipitation accuracy. Four weather stations in the Great Lakes region have the worst precipitation predictions in the dataset. Poor precipitation forecast accuracy in this region illustrates the difficulty in forecasting lake-effect snow. This phenomenon is discussed in greater depth in Sect. 3.3. Precipitation forecast accuracy for the Great

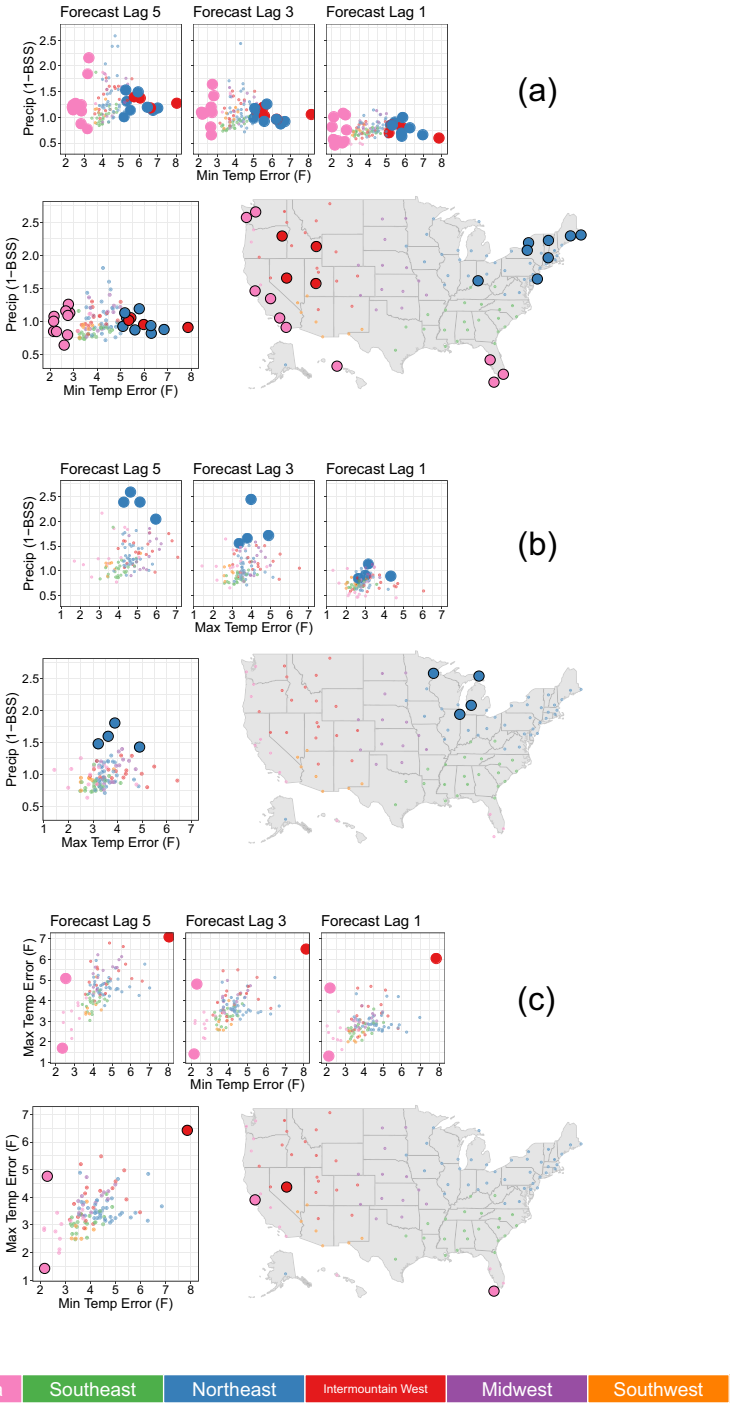

Fig. 5 Scatterplots comparing the three forecast error variables. The scatterplot to the left of the map is aggregated over all forecast lags. Points of interest discussed in the text are highlighted in the plots

Lakes region improves substantially as the forecast lag decreases and forecasts with lag 1 are as accurate as the rest of the nation.

Figure 5c shows the relationship between minimum and maximum temperature forecast accuracy. Three outliers stand out in these scatterplots, namely Key West, Florida, Austin, Nevada, and San Francisco, California. Key West predicts both minimum and maximum temperature more accurately then any other weather station. Key West also ranks in the top five for lowest variability in eight of the weather variables, which likely explains the accurate forecasts. Austin is the poorest predictor of both measures. Seventy miles along the "loneliest highway in America"(The Greater Austin Chamber of Commerce 2018) separate Austin from its weather measurements which were collected in Eureka, Nevada. The poor predictions for maximum and minimum temperature can be explained by the change in climate over such a large distance. This is reflected in a negative prediction bias of around 5°F for maximum temperature and a positive bias of around 7°F for minimum temperature. San Francisco has good predictions of minimum temperature and poor predictions for maximum temperature. This phenomenon is further explained in Sect. 3.3.

The interactive app developed in conjunction with this project allows for further investigation of forecast accuracy trends. The app is discussed in Sect. 4.

3.3 Seasonal trends

The position of the U.S. in the northern hemisphere makes most of the country subject to distinct weather seasons. Seasons are most pronounced in the northern U.S. We hypothesize that the forecast error behavior is inextricably linked to this seasonality. We explore this through a series of space-time graphs. Modeling space and time simultaneously creates a three-dimensional problem usually visualized as small multiples. Small multiples are "a series of graphics, showing the same combination of variables [e.g., latitude and longitude], indexed by changes in another variable [e.g., time]" (Tufte 2002, p. 170). The issue with this approach is that it becomes difficult to visually comprehend all but the most drastic changes from graph to graph. One alternative that allows simultaneous visualizations of both space and time is through the use of glyphs, or symbols, that allow for multi-dimensional visualizations in a spatial context (Carr et al. 1992,Wickham et al. 2012).

Figure 6 shows glyph plots of seasonal forecast errors throughout time. The forecast error is visualized as the scaled distance from a center point to the edge of a polygon with twelve observations starting with January at the 12:00 position and proceeding clockwise. The asymmetry of the glyphs about their center points illustrates how forecast errors change across time and across space. For example, locations in the Northeast are worse at forecasting precipitation in the winter than in the summer, while locations in the Southeast forecast precipitation equally well throughout the year.

In addition to highlighting forecasting asymmetries, Fig. 6 reveals location-specific anomalies. For example, San Francisco, California, predicts minimum temperatures well all year, but only predicts maximum temperatures well in the winter months. This is likely due to chilling coastal fogs known to frequent the region

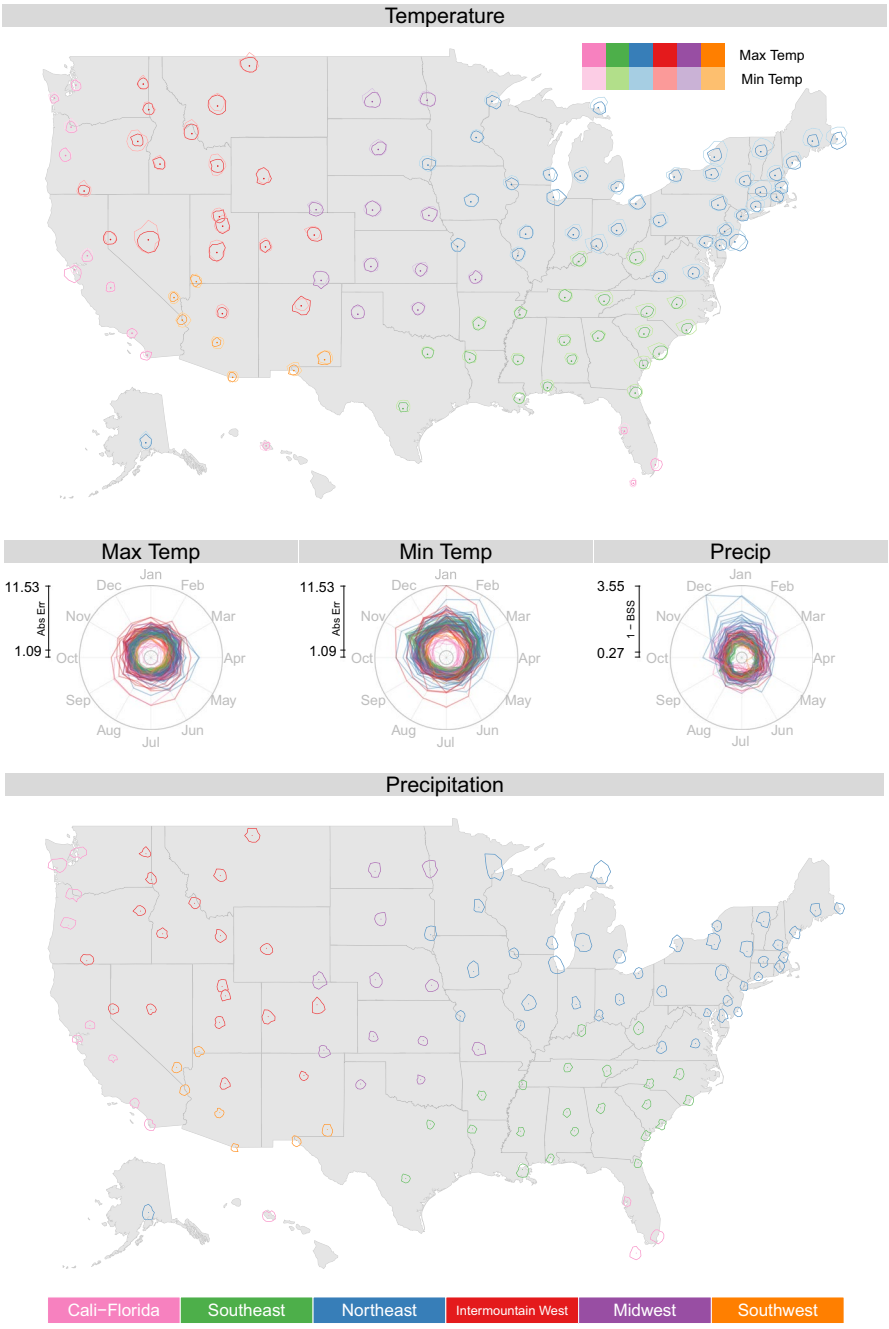

Fig. 6 Glyph plots of weather forecast accuracy averaged by month. The error is represented as the scaled distance from a center point to the edge of a polygon beginning with January at the 12:00 position and proceeding clockwise

throughout the year that can create sharp temperature differences over short distances (Nolte 2016). The struggle to predict temperature seems reasonable in light of these facts as this measurement location is more than 11 miles inland from the forecast location. The issue is likely less pronounced in the winter because the contrast between inland and coastal temperatures is reduced.

Maximum temperature predictions are particularly poor in the summer months in Austin, Nevada. It is unclear why predictions are worse in the summer than in the winter.

Another location-specific anomaly of note is the drastic seasonality of precipitation forecasts for locations surrounding the Great Lakes, as observed in Fig. 6. The error scatterplots in Fig. 5b show that precipitation accuracy is poor in this region, but the seasonality of the predictions cannot be observed in the scatterplots. The unusually bad forecasting in the winter is likely due to lake-effect snow which is prevalent in the region. Up to 100% more snow falls downwind of Lake Superior in the winter than would be expected without the lake-effect (Scott and Huff 1996). This area has been previously identified as having the most unpredictable precipitation patterns in the nation (Silver and Fischer-Baum 2014). The above examples demonstrate the ease with which comparisons can be made across space and time with these glyph-based plots. Information about how to generate the glyphs is included in Appendix B.

3.4 Variable importance

The differences in forecast error patterns across regions prompt identification of the most important climate measurements for predicting forecast error. We used random forests (Breiman 2001) to determine which weather variables had the greatest impact on the forecast errors. The data were aggregated over forecast lag and month. Three random forest models were generated for each weather region using the forecast error variables as the response. The means and standard deviations for each of the weather variables listed in Table 1 and the forecast lag were the predictor variables. Figure 7 contains three parallel coordinate plots that show the variable importance measures in each region for each forecast error variable. The importance measures obtained from random forests were recentered by subtracting the minimum importance measure and then rescaled to the interval (0, 100) by dividing by the maximum importance measure of the recentered values for each weather cluster and forecast error variable combination, and finally multiplying by 100. Thus, the most important variable within each weather region has a value of 100 and the least important has a value of 0 for each error measure. This allows direct comparisons of importance between weather regions and across error measures.

Figure 7 shows that the most important variable for the precipitation error is forecast lag regardless of weather region. None of the other variables are very important relative to lag. The Southeast shows minimum dew point (DP) and the standard deviation of maximum dew point as being somewhat important. Cloud cover is important for the precipitation error in the Northeast.

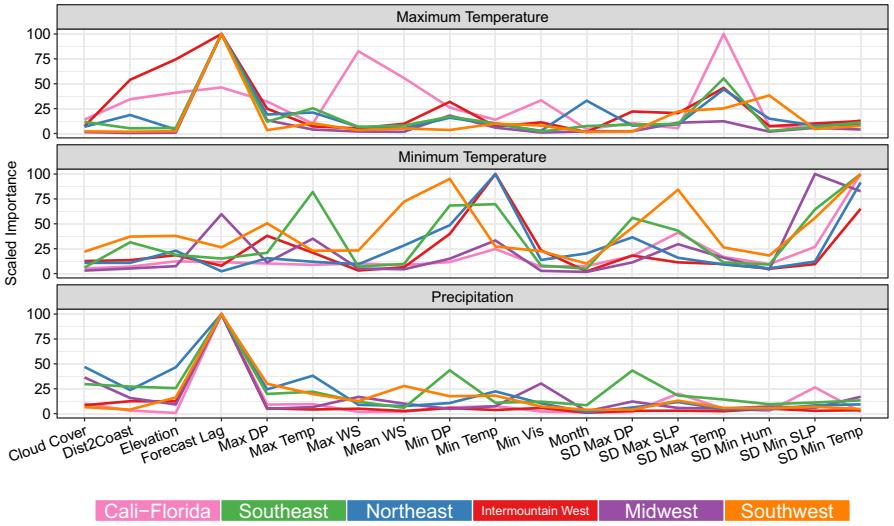

Fig. 7 Variable importance for each of the three forecast accuracy measurements. Variable importance measures have been rescaled to make the measures directly comparable between weather regions and accuracy measures

Forecast lag is also the most important variable for the maximum temperature error for all weather regions except Cali-Florida. The standard deviation of maximum temperature and maximum wind speed (WS) are more important than lag in Cali-Florida. The variability in maximum temperature is also important for the Southeast, Northeast, and the Intermountain West. Distance to coast (Dist2Coast) and elevation are important for the maximum temperature error in the Intermountain West.

Variables that are important for the minimum temperature error varied substantially across weather regions. The variability in minimum temperatures is important for all regions, but other important variables differ widely from region to region. Minimum temperature is the most important for the Northeast and Intermountain West, but maximum temperature is important for the Southeast. Minimum dew point and the variability in the maximum sea level pressure (SLP) are important in the Southwest while variability in minimum sea level pressure is the most important for the Midwest, Southeast, and Southwest. Forecast lag is not particularly important for any of the regions except for the Midwest.

4 Interactive application

It is difficult to identify the patterns in climate measurements and forecast errors for all weather regions with static visualizations. We developed an interactive Shiny app to enhance our weather data explorations. This app can be accessed at

https://jilllundell.shinyapps.io/finaldataexpoapp/.

The first tab of the app is an interactive version of the parallel coordinate plot introduced in Fig. 3. The app allows the user to select a weather region which is highlighted on the graph. Characteristics of the selected region can be easily seen and compared to all other observations.

The second tab of the app is an interactive scatterplot. Figure 5(a–c) shows examples of the graphs generated in this tab. The user can select up to two of the three forecast error variables to be on the axes. The forecast lag can also be selected. Points on the scatterplot can be brushed or clicked and the selected points show up on a map of the U.S. Information about selected stations is listed in a table under the graph. The idea of linked brushing between scatterplots and maps was first introduced in Monmonier (1989). This app allows for a more complete exploration of outliers and trends in the data across forecast lags and between error variables than a static graph.

5 Conclusions

Climate patterns in the United States cleanly separate into six recognizable regions through a cluster analysis using the means and standard deviations of the weather variables provided in Table 1. We explored the relationship between the three weather forecast variables (i.e., minimum temperature, maximum temperature, and precipitation) using correlation ellipses shown in Fig. 4. We found that all clusters show signs of positive correlations among the error variables with the exception of the Northeast cluster.

We visualized the pairwise relationship between forecast errors through a series of scatterplots across all forecast lags in Fig. 5. These plots highlight the superiority of locations in the Cali-Florida region for predicting minimum temperature across all lags, and also show that the poor precipitation predictions of the Great Lakes region are mostly confined to forecasts greater than lag 2. Lastly, the abnormally high errors in Austin, Nevada, are likely a product of the large distance between forecast and measurement locations.

We explored seasonal differences of forecast errors in Fig. 6 and observed that seasonal differences in forecast errors tend to be more pronounced in northern, inland clusters than southern clusters. We also showed that location specific anomalies, such as the asymmetry in seasonal maximum temperature forecast errors in San Francisco and the precipitation forecast errors near the Great Lakes, have plausible explanations in the literature.

Next, we compared the important variables in determining forecast errors across clusters using scaled random forest variable importance measures in Fig. 7. These measures demonstrate that forecast lag is most important in determining the maximum temperature and the precipitation forecast errors, but not important in predicting the minimum temperature forecast errors. Many clusters place similar importance on a few variables, but there are some variables that are important only in a single cluster, such as the importance of maximum wind speed in predicting the maximum temperature forecast error in Cali-Florida.

For further insight regarding the nature of forecast errors across these six clusters, we refer readers to our R Shiny app described in the previous section. A current version of the app can be found at the following URL:

https://jilllundell.shinyapps.io/finaldataexpoapp/.

This app, in conjunction with the visualizations presented in this article, reinforces the idea that the U.S. cleanly clusters into well defined weather regions and patterns in forecast errors are closely related to the unique climates that characterize each region.

The visualizations in this paper, both interactive and static, were designed to be scalable for larger weather datasets. We anticipate illustrating this capability on an expanded set of stations in the future. An expanded analyses will also serve to validate the regional patterns observed and described in this paper. In addition, we anticipate adapting several of the static glyph plots presented in this paper for interactive use. Greater interactivity will allow for more detailed explorations of weather patterns in the United States across both time and space.

Appendix A: Methods for data cleaning

We primarily used the dataset provided by the Data Expo to perform the analyses described in the article. We supplemented the provided location information with elevation and distance to the nearest major coast. Elevation information was obtained for each location through Google's API server (Google 2018) via the rgbif R package (Chamberlain 2017). Distance to coast was calculated as the closest geographical distance between each measurement location and one of the vertices in the U.S. Medium Shoreline dataset (National Oceanic 2018b), which includes all ocean and Great Lakes coasts for the contiguous 48 states. Because this dataset does not include the coastlines of Alaska and Hawaii, distance to coast calculations for these locations used manually extracted shorelines from NOAA's Shoreline Data Explorer (National Geodetic Survey 2018). We acknowledge there are limitations to this method of distance calculation, as distances for some locations, such as Arizona (Flagstaff, Nogales, and Phoenix), are slightly longer than they would be had we used shoreline information for Mexico's Gulf of California. Nevertheless, these measurements effectively separate inland weather stations from coastal stations.

Table 1 shows the weather variables included in our final analysis. We excluded mean daily measurements for temperature, precipitation, dew point, humidity and sea level pressure as these measurements were near perfect linear combinations of their corresponding minimum and maximum measurements. We also excluded maximum visibility from the analysis as this measurement was equal to 10 miles for more than 97% of all recorded measurements. Lastly, we combined the information provided by maximum wind speed and maximum wind gust by retaining only the lower of the two measurements after removing outliers. The decision to combine the information from these two wind variables was motivated by the fact that 13% of all maximum wind gust values were missing. In addition, it is difficult to separate unusually high, yet valid, maximum wind gust and wind speed measurements from true outliers.

Some stations did not record relevant climate variables. When possible, these missing observations were replaced with corresponding measurements obtained from the nearest National Weather Service (NWS) first order station as obtained through the National Climatic Data Center (NCDC) (National Oceanic 2018). Missing values include wind speed in Baltimore, Maryland, precipitation in Denver, Colorado, and replacements of outlier precipitation measurements at multiple locations. When replacements were not readily obtained through the NCDC, systematic missing observations were replaced with corresponding observations from the nearest geographical neighbor within the dataset, as was the case for visibility and cloud cover in Baltimore, Maryland (replaced with Dover, Delaware, measurements) and Austin, Nevada (replaced with Reno, Nevada, measurements).

Table 1 also shows the observation ranges for each of the included variables. These measurement ranges are either definitional, such as the bounds for humidity, or simply practical, such as the bounds for temperature. All measurements falling outside the bounds shown in Table 1 were removed prior to our analysis. Several individual outliers were also removed or replaced based on location-specific inconsistencies including

- Removal of one unusually low minimum temperature measurement in Honolulu, Hawaii, (< 10°F) and two in San Francisco, California (< 20°F);
- Replacement of the following unusually high precipitation readings with precipitation readings at nearby weather stations (National Oceanic 2018a):
  - Oklahoma City, Oklahoma, on 8/10/2017 (38.33 in $\rightarrow$ 0.8in)
  - Salmon, Idaho, on 4/21/2015, 5/2/2016, and 5/3-4/2017 (10.02 in $\rightarrow$ 0in)
  - Flagstaff, Arizona, on 12/24/2016 (7.48 in $\rightarrow$ 0.97in)
  - Indianapolis, Indiana, on 7/15/2015 (9.99 in $\rightarrow$ 0in);
- Removal of one unusually low minimum dew point measurement in Honolulu, Hawaii (< 40°F), two in Hoquiam, Washington (< 0°F), four in Las Vegas, Nevada (< -15°F), and two in Denver, Colorado (< -20°F).

Forecast variables were restricted to minimum temperature, maximum temperature, and the probability of precipitation. We found no obvious outliers in the weather forecasts. This is reasonable due to the fact that forecasts are not subject to inevitable sensor technology failures that occur when taking an actual measurement. Rather, the forecast data were replete with duplicate values for minimum temperature and precipitation. We retained the lowest forecast of minimum temperature and the highest forecast of precipitation probability for each forecast.

Forecast lags of six or seven days contained a large number of missing values. We removed all forecasts past lag 5. We also removed all forecasts containing negative lags (i.e., a forecast made *after* the actual observation).

Appendix B: Polar coordinate considerations for geographic maps

The glyph plots in Figs. 4 and 6 rely on proper conversions from polar to geographic or Cartesian coordinates. This allows the glyphs to be plotted directly on the underlying map, rather than embedding polar coordinate subplots in the image. Avoiding subplots allows for greater precision in the placement of the glyphs and avoids the computational burden of creating and embedding multiple figures. This direct plotting approach requires special considerations for geographical maps, as polar coordinate glyphs become distorted when projecting geographical coordinates to a Cartesian plane. For example, a perfect circle in geographical coordinates will appear elongated in the vertical direction when the circle is projected in the northern hemisphere. One solution to this issue is to project all geographical coordinates to a Cartesian plane prior to the glyph construction. This can be conveniently accomplished using the mapproject() function in the mapproj R package (McIlroy et al. 2017).

Polar coordinates are defined in terms of radius r and angle θ. Figure 6 defines $r \in [0, 1]$ as the scaled average absolute error between predicted and actual temperature and $\theta = \frac{(4-m)\pi}{6}$ where m represents the numeric month. We center each glyph at 0 with Cartesian coordinates

$$(x, y) = (r\cos\theta, r\sin\theta)$$

Let $(\mathbf{x}_i, \mathbf{y}_i)$ represent the set of Cartesian coordinates centered at the origin that create the glyph associated with location i. These coordinates are defined using the same units as the underlying map projection. The final coordinates of the rendered glyph are defined as

$$\alpha \cdot (\mathbf{x}_i + u_x, \mathbf{y}_i + u_y)$$

where (u_x, u_y) represents coordinates of location i and α represents a global scaling parameter used to adjust the size of the rendered glyphs on the map. A point is drawn at location (u_x, u_y) to serve as a reference for the glyph. Asymmetry about the point (u_x, u_y) reveals seasonal patterns in the forecast errors.

We construct the correlation ellipses of Fig. 4 with foci F_1, F_2 located along the semi-major axis $y = x$ $(\theta = \frac{\pi}{4})$ for positive correlations and $y = -x$ $(\theta = -\frac{\pi}{4})$ for negative correlations. We fix F_1 at the origin and denote r as the radius extending from F_1 to the edge of the ellipse, as illustrated in Fig. 8. This approach to ellipse creation is outlined in Knisley and Shirley (2001) and adapted here where we define r for $\theta \in [0, 2\pi]$ as

$$r = \frac{(1 - |\rho|)^2}{1 - \sqrt{|\rho|(2 - |\rho|)}\cos(\theta - \frac{\text{sign}\,(\rho)\pi}{4})},$$

where $\rho \in (-1, 1)\backslash\{0\}$ represents the desired correlation between forecast errors. In the event that $\rho = -1, 1,$ or 0, we use $\rho \pm \epsilon$ $(\epsilon > 0)$ when creating the ellipse to avoid numerical precision errors. The ellipse is then converted to Cartesian coordinates and centered at the origin as

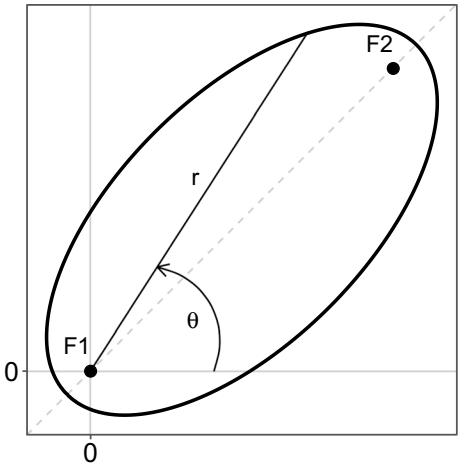

Fig. 8 Sample ellipse with F_1 located at the origin

$$\left(r\cos(\theta) - \frac{|\rho|(2-|\rho|)}{\sqrt{2}}, r\sin(\theta) - \text{ sign }(\rho)\frac{|\rho|(2-|\rho|)}{\sqrt{2}} \right).$$

Each ellipse is scaled to be circumscribed in the $[-0.5, 0.5] \times [-0.5, 0.5]$ square. This scaling makes it possible to create a matrix of ellipses using a common grid size. It also reduces the difference in areas between ellipses which facilitates comparisons of shape. This scaling is defined as

$$(\mathbf{x}'_i, \mathbf{y}'_i) = \left(\frac{\mathbf{x}_i}{2 \cdot \max(|\mathbf{x}_i|)}, \frac{\mathbf{y}_i}{2 \cdot \max(|\mathbf{y}_i|)} \right).$$

Note that there are three ellipses for each location. We define a matrix of ellipses centered at the shared vertex of the lattice denoted by (u_x, u_y). Let $(\mathbf{x}_i, \mathbf{y}_i)$ represent the coordinates of one of the three ellipses centered at this location. Each ellipse is centered and scaled on the map as

$$\alpha \cdot (\mathbf{x}'_i + u_x + o_1, \mathbf{y}'_i + u_y + o_2)$$

where o_1 and o_2 represent offset terms used to separate the centers of the three ellipses in the matrix defined for each location.

This direct plotting approach of the ellipses eases plot customization, as there is no need to reconcile formatting differences between independently created subplots. This approach can also be generalized to plot other geometric shapes on a geographic map. It is also helpful for interactive applications that require fast renderings of images in response to dynamic inputs.

Acknowledgements The authors would like to thank the Sections on Statistical Computing and Statistical Graphics of the ASA for providing the data used in this analysis. The primary analytical tool for this analysis was R, a free software environment for statistical computing and graphics (R Core Team 2018). Additional data information regarding specific measurement locations were provided in the weatherData R package (Narasimhan 2017). Distance and spatial calculations made use of the fields (Nychka

et al. 2015), geosphere (Hijmans 2016), mapproj (McIlroy et al. 2017), rgdal (Bivand et al. 2018), and sp (Bivand et al. 2013) R packages. Other data manipulations and visualizations made use of the tidyverse (Wickham 2017), as well as the ggforce (Pedersen 2018), latex2exp (Meschiari 2015), RColorBrewer (Neuwirth 2014), and reshape2 (Wickham 2007) R packages. Variable importance models made use of the randomForest (Liaw and Wiener 2002) R package.

References

Bivand R, Keitt T, Rowlingson B (2018) rgdal: Bindings for the 'geospatial' data abstraction library. https://cran.R-project.org/package=rgdal, R package version 1.2-18

Bivand RS, Pebesma E, Gomez-Rubio V (2013) Applied spatial data analysis with R. http://www.asdar-book.org/

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Briggs RS, Lucas RG, Taylor ZT (2003) Climate classification for building energy codes and standards: part 2-zone definitions, maps, and comparisons. ASHRAE Trans 109:122

Carr DB, Olsen AR, White D (1992) Hexagon mosaic maps for display of univariate and bivariate geographical data. Cartogr Geogr Inf Syst 19(4):228–236

Cetinkaya-Rundel M, Martinez W (2023) The 2018 data challenge expo of the American Statistical Association. Comput Stat

Chamberlain S (2017) rgbif: Interface to the global 'biodiversity' information facility API. https://cran.R-project.org/package=rgbif, R package version 0.9.9

Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2019) shiny: Web application framework for R. https://CRAN.R-project.org/package=shiny, R package version 1.4.0

Cohen J, Pfeiffer K, Francis JA (2018) Warm Arctic episodes linked with increased frequency of extreme winter weather in the United States. Nat Commun 9(1):869

Commission for Environmental Cooperation (1997) Ecological regions of North America: toward a common perspective. http://www3.cec.org/islandora/en/item/1701-ecological-regions-north-america-toward-common-perspective-en.pdf

Diaz HF (1983) Drought in the United States. J Climate Appl Meteorol 22(1):3–16

Google (2018) Get started. https://developers.google.com/maps/documentation/elevation/start

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning, vol 1. Springer, New York

Hathaway J, Pulsipher T, Rounds J, Dirks J (2013) Statistical methods for defining climate-similar regions around weather stations using NLDAS-2 forcing data. Tech. rep., PNNL-SA-98705. Richland, WA, USA: Pacific Northwest National Laboratory

Hijmans RJ (2016) geosphere: Spherical trigonometry. https://cran.R-project.org/package=geosphere, R package version 1.5-5

Karl T, Koss WJ (1984) Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983. https://repository.library.noaa.gov/view/noaa/10238

Knisley J, Shirley K (2001) Multivariable calculus online. http://math.etsu.edu/multicalc/prealpha/

Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22

Lundell J, Bean B, Symanzik J (2018) Let's talk about the weather. JSM Proceedings. Statistical Computing Section, American Statistical Association, pp 1944–1958

McIlroy D, Brownrigg R, Minka TP, Bivand R (2017) mapproj: Map projections. https://cran.R-project.org/package=mapproj, R package version 1.2-5

Meschiari S (2015) latex2exp: Use LaTeX expressions in plots. https://cran.R-project.org/package=latex2exp, R package version 0.4.0

Monmonier M (1989) Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. Geogr Anal 21(1):81–84

Murdoch D, Chow E (1996) A graphical display of large correlation matrices. Am Stat 50(2):178–180

Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J Classif 31(3):274–295

Narasimhan R (2017) weatherData: get weather data from the web. https://cran.R-project.org/package=weatherData, R package version 0.5.0

National Geodetic Survey (2018) NOAA shoreline data explorer. https://www.ngs.noaa.gov/CUSP/

National Oceanic and Atmospheric Administration (2018a) Climate data online. https://www.ncdc.noaa.gov/cdo-web/
National Oceanic and Atmospheric Administration (2018b) NOAA medium resolution shoreline. https://shoreline.noaa.gov/data/datasheets/medres.html
Neuwirth E (2014) Rcolorbrewer: ColorBrewer palettes. https://cran.R-project.org/package=RColorBrewer, R package version 1.1-2
Nolte C (2016) The story of the San Francisco summer is a bit foggy. https://www.sfchronicle.com/bayarea/nativeson/article/Summer-fog-can-be-here-today-gone-tomorrow-9141028.php
Nychka D, Furrer R, Paige J, Sain S (2015) fields: Tools for spatial data. https://cran.R-project.org/package=fields, R package version 9.0
Pedersen TL (2018) ggforce: Accelerating 'ggplot2'. https://cran.R-project.org/package=ggforce, R package version 0.1.2
R Core Team (2018) R: A language and environment for statistical computing. https://www.R-project.org/
Scott RW, Huff FA (1996) Impacts of the Great Lakes on regional climate conditions. J Great Lakes Res 22(4):845–863
Silver N, Fischer-Baum R (2014) Which city has the most unpredictable weather? https://fivethirtyeight.com/features/which-city-has-the-most-unpredictable-weather/
The Greater Austin Chamber of Commerce (2018) Austin, Nevada: So much to do. http://austinnevada.com/visitor-information/chamber-information/
Tufte ER (2002) The visual display of quantitative information. Graphics Press, Cheshire
Unwin A, Valero-Mora P (2018) Ensemble graphics. J Comput Graph Stat 27(1):157–165
Weigel AP, Liniger MA, Appenzeller C (2007) The discrete Brier and ranked probability skill scores. Mon Weather Rev 135(1):118–124
Wickham H (2007) Reshaping data with the reshape package. J Stat Softw 21(12):1–20
Wickham H (2017) tidyverse: Easily install and load the 'tidyverse'. https://cran.R-project.org/package=tidyverse, R package version 1.2.1
Wickham H, Hofmann H, Wickham C, Cook D (2012) Glyph-maps for visually exploring temporal patterns in climate data and models. Environmetrics 23(5):382–393