Title of submission to PLOS journals

Emily M.M. Palmer[1‡], Katherine R. Pulham[1‡], Jessica R. Robinson[1‡], Ericka B. Smith[1‡]

1 Statistics Department, Oregon State University, Corvallis, Oregon, United States

‡These authors also contributed equally to this work.

Abstract

For our submission the authors created a Shiny app to explore precipitation, found at (insert link here). This app is based on the CESM-LENS and ERA global reanalysis datasets, and includes four tabs, which explore ensemble member precipitation variability in CESM-LENS, decadal cumulative precipitation, trends in yearly total and cumulative precipitation and variability in precipitation, and seasonal trends in percent deviation from average precipitation. This app is intended as an EDA tool for researchers to explore both local and regional precipitation patterns.

Introduction

Data exploration is what connects an abundance of information with concise questions to address and analyze. Without exploration of the patterns or behaviors in a large dataset, it is extremely difficult for a statistician or analyst to target their interest and answer an in-depth question. In turn, the resulting research is likely to fail in providing insight and sound reason for change to the public. Stephen Few, in his book *Now you see it* [1], says "[Visualization] provides a powerful means to net the prize fish from the vast schools of data that swim the information ocean."

The use of exploratory data analysis (EDA) in the expansive and growing bulk of available climate data is a quintessential example of the utility and scope of this task.A quintessential example of this is the use of exploratory data analysis (EDA) in the expansive and growing bulk of available climate data. Climate datasets are notoriously large due to their inclusion of many unique measurements over expansive ranges of time and location. With advancements in technology and modelling techniques, information about the earth's climate is growing rapidly- a double-edged sword for climate scientists and statisticians and a perfect reminder of the importance of data exploration and visualization.

Though imperative to a worthwhile analysis effort, exploration and visualization can be time-consuming when created by hand in a computer programming language such as R. This is especially true of large datasets where it is difficult to cohesively explore without extensive manipulation and multiple plots.

This is the the motivation for the CliMates Precipitation Data Dashboard- a multifaceted point-and-click explorative tool created with the goal of understanding the behaviors and patterns of precipitation in the United States.

Materials and methods

Data Used

The ERA Interim dataset comes from a climate data reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF) [2] . ERA Interim is a global atmospheric reanalysis that tracks a large number of variables, including maximum temperature, minimum temperature, and precipitation from 1979 to 2017. Dee, et al. In essence, the original dataset was generated by integratingclimatological measurements, such as from satellites and weather stations and constraining the variables by their physical properties, to generate as accurate a model as possible for global climate resulting in measurements for points across a global raster grid. For this competition, the spatial scope was limited to the continental United States, with a small rectangular buffer region extending into the ocean. In essence, the original dataset was generated by constraining the variables by their physical properties, as well as in situ measurements made globally, to generate as accurate a model as possible for global climate. For this competition, the spatial scope was limited to the continental United States, with a small rectangular buffer region extending into the ocean. This dataset integrates climatological measurements, such as from satellites and weather stations into daily precipitation amounts in mm across a 80km grid.

The CESM-LENS dataset used in this work is a set of climate model simulations created by the Community Earth System Model Large Ensemble Community Project and supercomputing resources provided by NSF/CISL/Yellowstone, and led by Dr. Clara Deser and Dr. Jennifer Kay. (Kay et al. 2005) [3]. The data are an ensemble model with 40 members. Each member has a slightly different starting point (initial atmospheric state), but follows the same "rules" (uses the same model and undergoes the same radiative forcing scenario). (Kotu 2019). The data provided for the competition includes a rectangle around the contiguous United States, with a small buffer zone. Within that space we focused on the historical time period (1920-2005) and precipitation data. The goal of the project was to elucidate the differences between climate change and internal climate variability.

shiny

Shiny [4] is an R package that allows for easy building of interactive plots as well as deployment. Our app is hosted on shinyapps.io, a platform that allows for hosting of shiny applications. Our app consists of two main pages: the first an explanation of the competition, the second the interactive precipitation EDA tool. The EDA tool consists of four tabs: 1234. The first two tabs make use of the CESM-LENS data, and the third and fourth make use of the ERA-Interim data.

Climate data is large. Even limited to the United States, these files take up multiple gigabytes of storage. Even simple calculations using these datasets take seconds to minutes to compute which is not ideal in an interactive application where many users expect immediate results. Thus a usable application requires these datasets to be pre-aggregated in some way to make computation time conducive to user interactivity. Since we were interested in keeping the original spatial resolution, as much spatial resolution we decided to aggregate by timescale.

Explain how we each pre-aggregated our data - First tab? Second tab data were aggregated by decade and averaged across members Third tab data were aggregated by year

The fourth tab is based on a pre aggregated dataset that is aggregated down to the seasonal level. Thus, for every grid point in the ERA dataset, there are four values per year.

Even with this aggregation, the app still takes several seconds to open initially.

Filtering

A quick solution to the difficulty of working with big datasets is to subset as early and as much as possible. This was able to be efficiently done using tidync [5] and the other packages from the tidyverse [6] (primarily purrr and dplyr). However, in order to create an interactive visualization such as a shiny app, the subsetting needs to be left to the user.

We explored a number of different methods for subsetting. Perhaps the most intuitive way of subsetting when doing an initial data exploration is picking two latitudes and two longitudes to put directly into the console that creates a boundary box. This is how we started. For these data raw boundary boxes were effective but not particularly useful, because we did not want to presume our average user would be knowledgeable about the latitude and longitude coordinates of locations throughout the contiguous United States, so we moved away from this initial choice. Seeing this we explored the option of intersecting our data with shapefiles in a "cookie-cutter" fashion. This was also effective but computationally inefficient, which is not coherent with our goal of making the data accessible. The precision available via the shapefile method had been attractive in theory. In practice, these data are too coarse for the difference to be worth the cost in computing resources and time. This can be understood clearly by viewing the observation locations on our map.

In the end we found a middle ground. We employed maps with detailed borders for selection of the area of interest, but they were accessed via the mapdata function rather than through a specific shapefile. This precision allows the user to quickly recognize visual landmarks for what area of the world they're looking at, and to choose different states. The tradeoff is that it doesn't allow for the flexibility that a user provided shapefile would allow. We suggest this as a future feature that would be beneficial for a more generalized use case, as it would allow users interested in geographic features, cities, vegetation types, and more to view the data in subsets specific to what they're interested in. We believe our setup to be more than sufficient for the purposes of this work though. The user can select areas directly from the map and change the size and placement easily as they see fit. The underlying selection mechanism for slicing the data is still in that original format of a latitude and longitude bounding box. Combining strategies like this allows for understanding of what data have been selected while ensuring a speedy delivery.

CESM-LENS Member Plots

Since the goal of the CESM-LENS data is to be able to distinguish between climate change effects and internal climate variability, a topic of interest is the variation between different members of the ensemble model. Recall that the "mMembers" of an ensemble model can be thought of as iterations of the same model, each with slightly different starting points. With these plots we explore that internal climate variability. During the preprocessing stage each pair of latitude and longitude is treated individually. Data were reduced by calculating average precipitation for each day of the year in groupings of location, decade, and member. At this point, decade 9 (2000-2005) was eliminated due to not being a full decade. Next, the range (from all of the members) of these daily average precipitations was calculated for each location and decade. The reasoning here is that these ranges can be used as a proxy for the variability between members. This marks the end of preprocessing and the next step of filtering is chosen by the user. When the user selects a location, these ranges are averaged again to get one value over that entire spatial area of the bounding box. Distinction between decades is retained

throughout. Then, the boxplot displays the distribution of these averaged range values over all of the days of the year for each decade. The smoothed plot retains the day of the year information and displays these values by day of the year for each decade. Calculations for the curve were done via the geomsmooth() function within ggplot2 and with method = 'loess' and formula 'y x'. These values of average range are described as equating to variability between members over time, but the intent is to develop intuition about this variability rather than strictly trying to quantify it.

Results

The landing page of our app starts with a brief overview of the project and data used (much of which is repeated in this report). The user then navigates to the "Explore The Data" tab to start.

(insert overview screenshot here)

The user can then decide what location they want to explore, and the spatial scale at which to explore it. Since the authors are based in Oregon, we'll start there. The user selects the state of interest if they want to zoom in to a finer scale. Minor zoom adjustments can be made at this point. There are two sets of points: circles denote the grid points for the ERA-Interim dataset, while triangles denote the grid points for the the other for the CESM-LENS dataset. We can see here that the ERA data is at a much finer spatial scale. Different tabs rely on different datasets.

Now the user selectshas to select what area in the map to explore. This is done by clicking and dragging a rectangular selection of interest. Once the point selection is finalized, the user clicks "Go" and the selected plot is rendered for the subsetted data.Make sure to click "Go" after a location is selected!

(insert brushedpoints oregon screenshot)

The user also has the option to compare to a different region. This is done by checking the "Compare to a different location" checkbox. Another map pops up, and the user repeats the selection process described above. The authors enjoyed comparing eastern/western Oregon where the rain shadow effect plays a role in the amount of rainfall, and comparing Oregon to New Mexico, where long term trends in cumulative precipitation are notably different.

Member Variability

Decadal Cumulative Precipitation

The tab labeled "Decadal Cumulative Precipitation" uses CESM-LENS to investigate trends in cumulative precipitation from 1921 to 2005, and to compare seasonality of precipitation between different locations. One of the challenges of doing an exploratory data analysis of precipitation is that it tends to be very noisy data. Precipitation can go from zero one day to an unusually high measurement the next day, back down to zero and then to a smaller non-zero measurement. This presents a challenge when attempting to get a rough idea of "when does it rain" and "what are the long term trends in precipitation" for a particular geographic area. Two ideas for working around this issue came to mind: calculating mean precipitation for a period of time, and looking instead at cumulative precipitation. To reduce the variability of precipitation, this plot does two things: taking the mean across all the years in a given decade, and calculating a cumulative sum. Taking the average across a decade reduces the variability of the estimator, but sacrifices some of its specificity. The second technique employed was calculating a cumulative sum which resets to zero on October 1st of each year, which was chosen since it's the beginning of a water year [USGS]. The first plot

combines these variance reduction strategies to show the average cumulative precipitation for each decade.

(insert cumulative precip graph)

The second plot is the numeric derivative of the first plot. This shows what days of the year had the most rain. By comparing the two plots, it becomes clear how taking a cumulative sum reduces the noise. Calculating this as a numeric derivative rather than calculating the mean precipitation cuts down on the files that need to be stored and does not take up enough computation time to interfere with the user experience too much.

(insert derivative average precip graph)

Yearly

The tab labeled "Yearly" uses the ERA-Interim data set to provide an opportunityoutlet for the user to explore yearly behavior of precipitation with plots of three measurements. On this tab, the user is given the choice of a range of years in addition to the sidebar choice of location of interest. The first plot depicts meantotal rainfall aggregated by year; all precipitation measured within the selected location boundary is added for the years chosen.

(Insert plot)

The second plot portrays the cumulative precipitation for the location and range of years chosen; total yearly precipitation is cumulatively summed from the start to end year.

(Insert plot)

The third plot reveals the within-yearwithin year variability of precipitation in the location of interest. This is calculated by finding the mean total monthly precipitation for the location for all twelve months of in the year and then calculating how these total values vary within the year. This allows the user to see how a year's worth of precipitation values varies as a function of yeartime.

(Insert plot)

Seasonal Precipitation Deviation

This tab allows users to explore trends at a seasonal level (winter/spring/summer/fall). It explores the percent deviation in rainfall, calculated as the average seasonal precipitation for a given year and selected location, minus the average overall seasonal precipitation for the selected location, divided by the average overall seasonal precipitation at the selected location.

The first plot shows a line chart breaking down these deviations by season. The overall United States average is included as a comparison (and the user can click the checkbox to remove this if wanted). If the user selected another region to compare there will be a third line on this chart.

The results of this tab are widely different depending on the selected location. Overall, for the United States average, we see that precipitation has been lower than average for the past 15 years or so. This trend is noisier and less pronounced when looking at the selected Oregon points. Other selected locations - South Dakota for instance, shows a much larger negative percent deviation in recent years, the most noticeably in Summer months.

(include seasonal line plot)

The second plot on this tab shows precipitation strips - inspired in part by the Warming/Climate stripes created by Ed Hawkins, Professor for Climate Science at the University of Reading. Each strip represents the percent deviation from average precipitation values (described above). Brown stripes represent drier seasons, green stripes represent wetter. . This graphic shows that often drier periods last longer than a

season, similarly for wetter periods. Some even last multiple years. Each row of this graphic is a different selected location, including the US average baseline. Often, periods of dry/wet happen at the same time for different locations/across the US.

(include strips plot)

Discussion

We were able to see interesting regional aspects of the data that were not able to be seen when aggregated across the entire spatial scale (all of the United States).

Throughout our plots, we sawwere able to see that times/locations with more rainfall were more variable in said rainfall.

Conclusion

Even from selecting just a few locations using this EDA app, the authors came up with numerous questions from the observations in this app. We hope that this app is a helpful starting place for researchers interested in examining regional precipitation trends.

Acknowledgments

Thanks to our faculty mentors James, Charlotte and Lisa Shiny, git, and big data were all relatively new topics for the authors, and this competition provided a fun and educational time for us to learn these tools important in our field. Special thanks to CJ Keist for helping us get set up with an RStudio Server instance.

[4]

References

1. Few S. Now you see it: Simple Visualization Techniques for Quantitative Analysis. Analytics Press; 2009.

2. Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society. 2011;137(656):553–597. doi:10.1002/qj.828.

3. Kay JE, Deser C, Phillips A, Mai A, Hannay C, Strand G, et al. The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. Bulletin of the American Meteorological Society. 2015;96(8):1333–1349.

4. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R; 2019. Available from: `https://CRAN.R-project.org/package=shiny`.

5. Sumner M. tidync: A Tidy Approach to 'NetCDF' Data Exploration and Extraction; 2019. Available from: `https://CRAN.R-project.org/package=tidync`.

6. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686. doi:10.21105/joss.01686.