CliMates precipitation data dashboard
An exploratory tool for analyzing U.S. precipitation data

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 contributed equally to this work.

Abstract

We created a Shiny application to explore precipitation, found at https://jimmylovestea.shinyapps.io/datadash/. This application is based on the CESM-LENS and ERA-Interim 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 application is intended as an exploratory data analysis (EDA) tool to investigate both local and regional precipitation patterns, and allows the user to explore the data directly without the overhead of handling large files with complex types.

Introduction

Data exploration is what connects an abundance of information with concise questions to address and analyze. Without exploration of the patterns and complexities in a large dataset, it is extremely difficult for a researcher to target their area of interest and investigate an in-depth question. In turn, the resulting inquiry 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."

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 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). 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 [2]. In essence, the original dataset was generated by integrating climatological 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.

The CESM-LENS dataset 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 [3]. The dataset consists of an ensemble model with 40 members where 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) [4]. The goal of the CESM Large Ensemble Community Project is to elucidate the differences between climate change and internal climate variability. 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.

Shiny web applications: interactivity and optimizing for deployment

Shiny [5] is an R [6] package that allows for easy building and deployment of interactive web applications. Our application is hosted on shinyapps.io, a platform provided by RStudio for hosting Shiny applications. Our application 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 "Model Variability", "Decadal Cumulative Precipitation", "Total, Accumulation, Variation Time Series", and "Seasonal Precipitation Deviation". The first two tabs make use of the CESM-LENS data, while the third and fourth utilize the ERA-Interim data.

Climate data are 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 we decided to aggregate by timescale. The first two tabs are pre-aggregated by decade, the third by year, and the fourth by season within a year. Even with this aggregation, the app still takes several seconds to open initially, but once loaded, the application empowers users to rapidly interact with the application.

Subsetting and interactive components for areas of interest

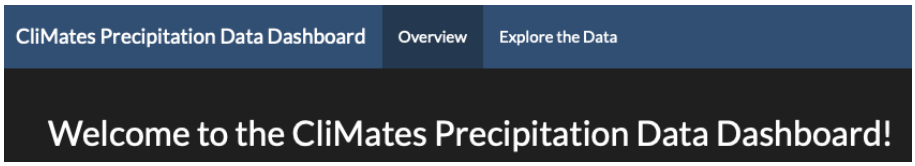
A quick solution to the difficulty of working with big datasets is to subset as early and as much as possible. This was able to be efficiently done using tidync [7] and the

tidyverse [8]. However, in order to create an interactive visualization such as a Shiny app, the subsetting needs to be left to the user.

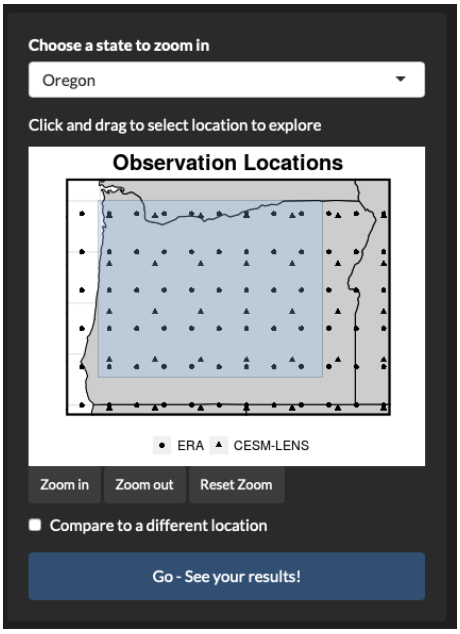
We employed maps with detailed borders for selection of the area of interest, accessed via the built-in map functions [9] rather than through a specific shapefile. This accommodates easy recognition of visual landmarks such as states. The tradeoff is that it doesn't allow for the flexibility that a user provided shapefile would allow, and we suggest this as a future feature that would be beneficial for a more generalized use case. For these purposes this level of exactness is ideal. 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 the original format of a latitude and longitude bounding box – just significantly more user friendly. This combination provides an understanding of what data have been selected while ensuring a speedy delivery.

Initially we started by simply picking a maximum and minimum latitude and longitude and using them to create a selected area. Under the assumption that an average user would not know the latitude and longitude of their location of interest, we considered intersecting shapefiles in a "cookie cutter" fashion to subset. Though the precision of this 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. Combining a point-and-click selection method with the built-in map functions provide a balance between ease of use and computational efficiency.

Results and discussion



The landing page of our app starts with a brief overview of the application, authors, and the materials used to make it. The user then navigates to the "Explore the Data" tab to start.



There the user decides what location they want to explore and the spatial scale at which to explore it. 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 CESM-LENS dataset. We can see here that the ERA-Interim data is at a much finer spatial scale. Different tabs rely on different datasets.

Now the user selects an area in the map to explore 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.The user also has the option to compare to a different region by checking the “Compare to a different location” checkbox. Another map appears, and the user repeats the selection process described above.

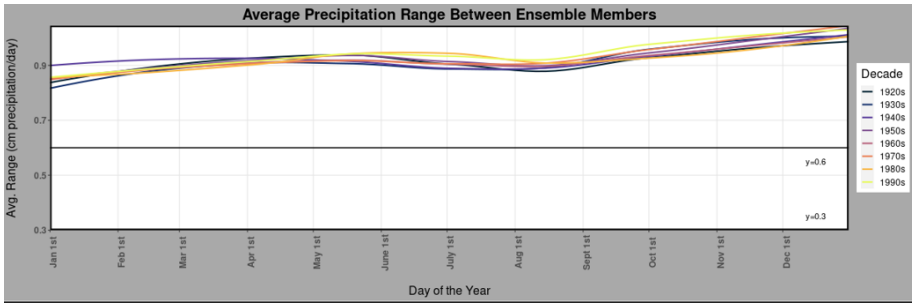
Model variability

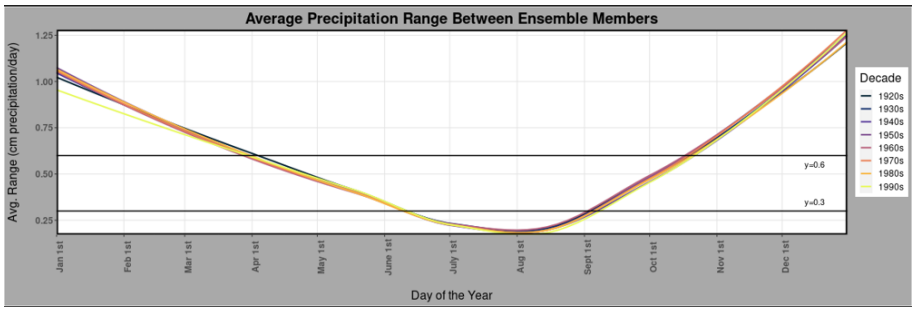
Since the goal of the CESM-LENS data is to be able to distinguish between climate change effects and internal climate variability, a topic of interest is the variation between different members of the ensemble model. Members of an ensemble model can be thought of as iterations of the same model, each with slightly different starting points. With the plots on this tab 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 for each unique combination of location, decade, and ensemble member. for each unique combination of location, decade, and ensemble member. Next, the range of these daily average precipitations over all of the members 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 depends on user input.

When the user selects a region, these ranges are averaged again to get one value over the 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. With the smoothed plot the user can investigate seasonal changes in variability, while the boxplot facilitates comparison of total decadal variability. Calculations for the curve were done with LOESS [10]. These values of average range are treated as a proxy for variability between members over time, but the intent is to develop intuition about this variability rather than strictly trying to quantify it.

See the figures below for an example of two states (Washington on top, Maine on bottom). Washington has a noticeable seasonal pattern in variability with low variability in summer. Conversely, Maine has very little change in variability over the year.

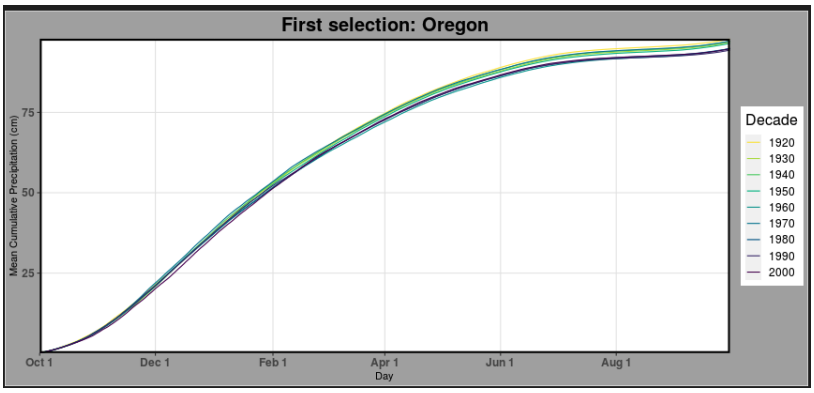




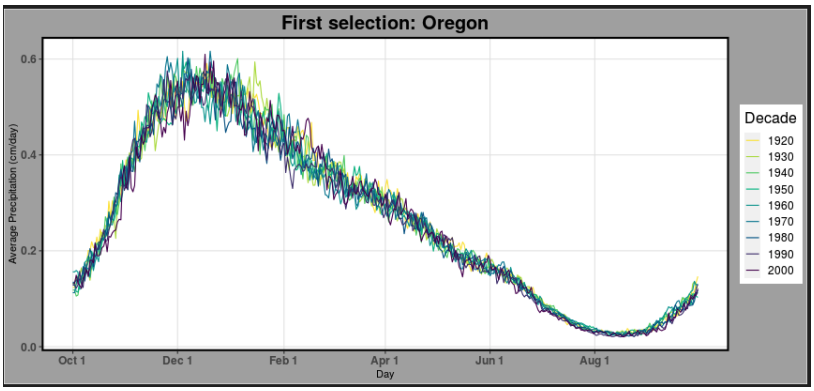
Decadal cumulative precipitation

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 calculates the cumulative sum, then takes the mean across all the years in a given decade. Taking the average across a decade reduces the variability of the estimator, but sacrifices some of its specificity. The cumulative sum resets to zero on October 1st of each year, which was chosen since it's the beginning of a water year [11]. The first plot combines these variance reduction strategies to show the average cumulative precipitation for each decade. The user can zoom in by clicking and dragging to select an area of the plot, then double clicking to zoom in. This facilitates investigation into parts of the plot that would otherwise be difficult to read.

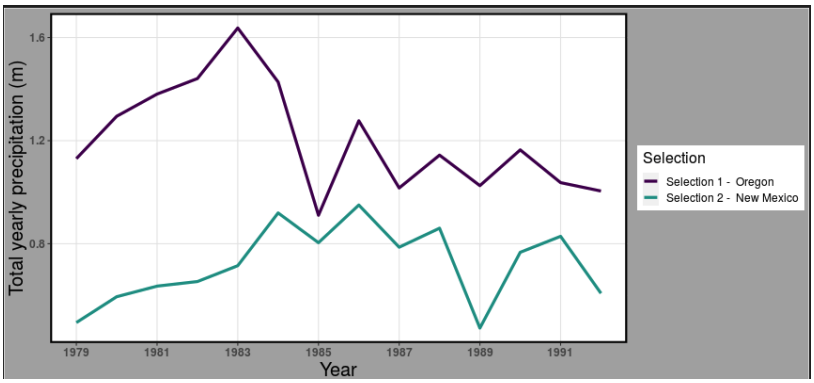


The second plot is the numeric derivative of the first plot. This shows what days of the year had the most rain. By comparing the two plots, it becomes clear how taking a cumulative sum reduces the noise. Calculating this as a numeric derivative rather than calculating the mean precipitation cuts down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



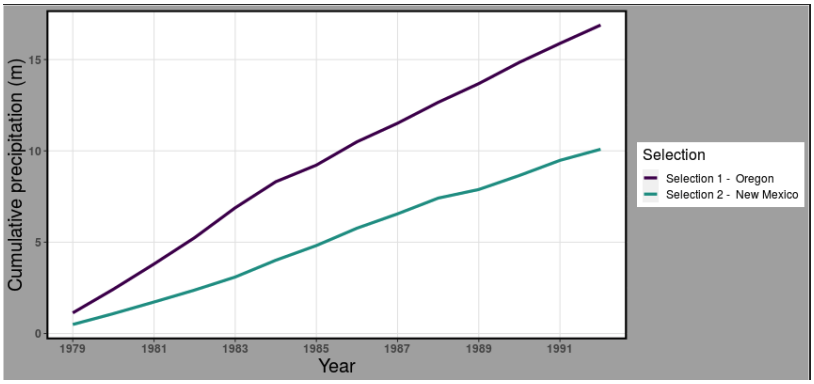
Yearly total, cumulative, and by-month variability

The tab labeled "Yearly Total, Cumulative, and By-Month Variability" uses the ERA-Interim data set to provide an opportunity 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 universal sidebar choice of location(s) of interest. The first plot depicts the yearly total precipitation averaged by the location points in the chosen boundary(ies). With this plot, the user can see how total precipitation is changing over time in their location(s) of interest.



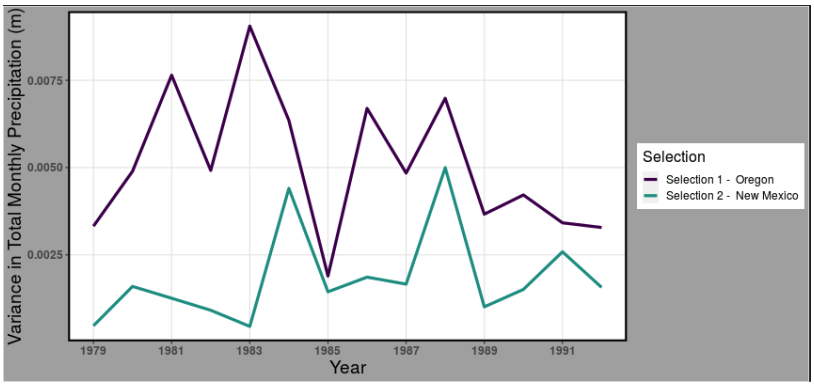
We can see that Oregon's total precipitation is higher overall than that of Arizona (as one would expect). Interestingly, the behavior of total precipitation in these two locations has an inverse relationship from 2013 to 2017 while in prior years, the behaviors are similar.

The second plot displays the cumulative precipitation for the location and range of years chosen; Similarly to the above plot, cumulative precipitation for all points in the location boundary are averaged and plotted by year, allowing the user to see the behavior of the cumulative precipitation over a wide span of time.



The plot of cumulative precipitation is helpful as the user can see slopes of the cumulation lines compared and detect any jumps fairly easily.

The third plot reveals the within-year variability of precipitation in the location of interest. Monthly totals of precipitation are taken similarly to the above plots and then the variance amongst those month totals for a given year are plotted. This allows the user to see how a year's worth of precipitation values varies over time.



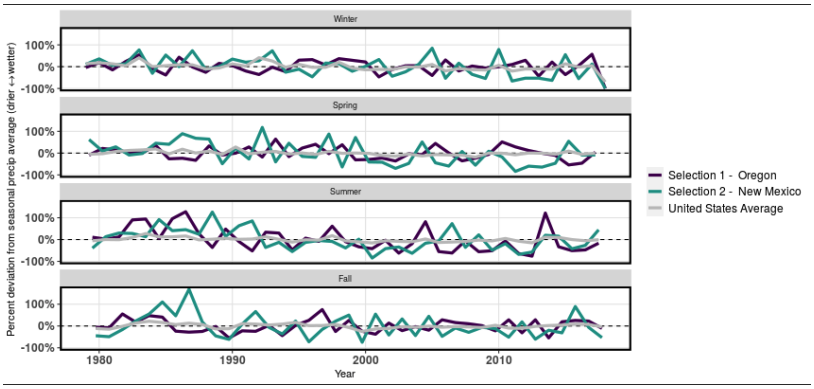
Above, for example, we can see that locations with higher total precipitation tend to have higher variability in monthly total precipitation values.

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

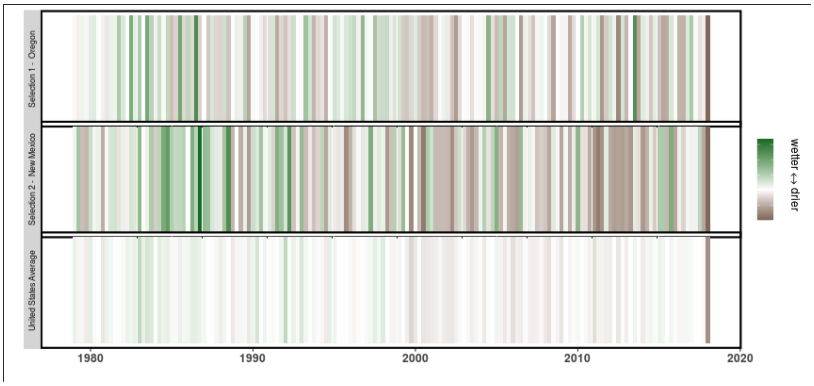
The first plot shows a line chart breaking down these deviations by season. The overall United States average is included as a comparison (and the user can click the checkbox to remove this if wanted). If the user selected another region to compare there will be a third line on this chart.

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, New Mexico for instance, shows a much larger negative percent deviation in recent years, the most noticeably in summer months.



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 [12]. Each strip represents the percent deviation from average precipitation values (described above). Brown stripes represent drier seasons, green

stripes represent wetter seasons. 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 and wet happen at the same time for different locations across the US.



Conclusion

We were able to see interesting regional aspects of the data that were not able to be revealed when investigating the entire spatial scale (all of the contiguous United States). Throughout our plots depicting variability in precipitation, we found that time ranges and locations with greater precipitation had more variance. The inverse of this also held; areas and times with lesser precipitation had a tendency to vary much less than that of rainier areas. Even when selecting minimal locations using this EDA application, numerous questions were raised about precipitation in different locations by both the creators and beta users of the application . We hope that this application serves as a helpful starting place for researchers, students, or analysts interested in examining regional precipitation trends and also suggest it as a case study that could be expanded upon to other datasets and circumstances.

Acknowledgments

We'd like to thank our faculty advisors for their support in this endeavor: Lisa Ganio PhD., James Molyneux PhD., and Charlotte Wickham PhD. Additionally, thank you to Jupiter Intelligence, the CESM Large Ensemble Community Project, and the European Centre for Medium-Range Weather Forecasts as well as the ASA ENVR Section for aggregating the data resources used in this effort. And, thank you to CJ Keist at OSU CoSINe IT Services for setting up and overseeing the RStudio server. Thank you to the ASA Environmental Section for hosting this competition. The combination of R, Shiny, GitHub, and large climate datasets used in this competition provided fun and challenging means for us to learn important tools in our field.

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. Kotu V, Deshpande B. In: 2. Morgan Kaufmann Publishers; 2019. p. 19–37.

5. 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`.

6. R Core Team. R: A Language and Environment for Statistical Computing; 2019. Available from: `https://www.R-project.org/`.

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

8. 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.

9. code by Richard A Becker OS, version by Ray Brownrigg Enhancements by Thomas P Minka ARWR, Deckmyn A. maps: Draw Geographical Maps; 2018. Available from: `https://CRAN.R-project.org/package=maps`.

10. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American statistical association. 1988;83(403):596–610.

11. Explanations for the National Water Conditions;. Available from: `http://water.usgs.gov/nwc/explain_data.html`.

12. Show Your Stripes;. Available from: `https://showyourstripes.info/`.