Analyzing weather data and determining corrupted data

Zhixing Guo
Justin Kelm
Adriana Morales
Nikolas Wojtalewicz

PI4 Data Bootcamp UIUC

May 26, 2018

Overview

- Weather data collection
- Goals
- The process of data cleanup
- Data control tests
- Creating new variables
- Creating a model that helps us predict corrupted data to easily clean up
- Our Algorithm

The data

- Sensor networks are being used to monitor ecosystems and to record ecological phenomena
- Mainly weather conditions like: humidity, temperature, air pressure, winds, etc...
- Sensor networks are very useful, but they are susceptible to malfunctions that can result in lost or poor-quality data.
- Steps can be taken to minimize the risk of loss and to improve the overall quality of the data.

Team Goals

- Effectively determine corrupt data using data control tests
- Clean our data and create new variables
- Build an algorithm that can flag corrupted data

Steps taken to clean the data

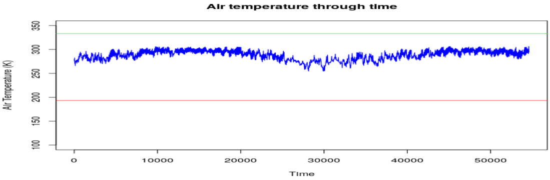
- We got rid of unnecessary variables i.e. ones that had same values
- Extracting data from certain variables

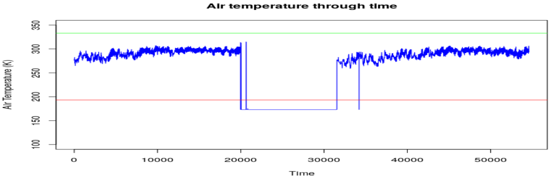
Quality Control Tests

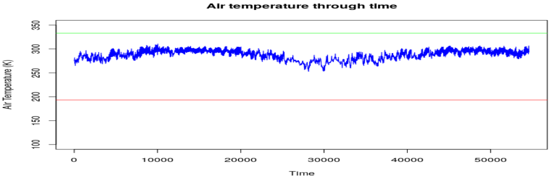
Date and time: Each data point has a date and time associated with it. Because streaming sensor networks collect data in chronological order, the date–time pairs should be sequential.	
Range: A range check ensures that the data fall within established upper and lower bounds. These bounds can be absolute, based on the characteristics of the sensor.	
Persistence: When the same value is recorded repeatedly, it may be indicative of a bad sensor or other system failure.	
Change in slope: A check for a change in slope tests whether the rate of change is realistic for the type of data collected. A sharp increase or d a very short time interval may indicate that the sensor was disturbed.	
Internal Consistency: Consistency checks determine if the data was collected under unsuitable conditions for a specific sensor.	

J. Campbell:
http://www.bioone.org/doi/full/10.1525/bio.2013.63.7.10

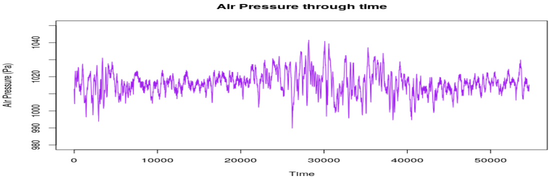
Initial Visualization of Data: Temperature

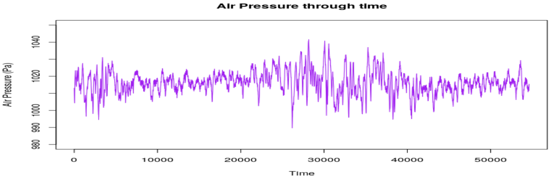


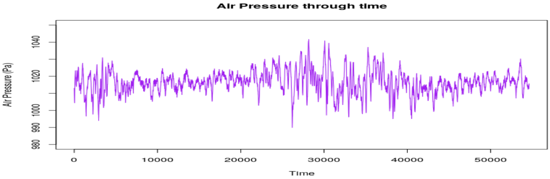




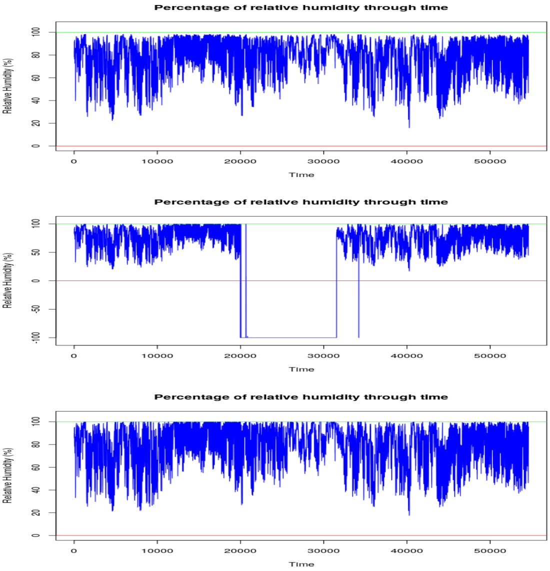
Initial Visualization of Data: Air Pressure







Initial Visualization of Data: Relative Humidity



Initial Algorithm

- The main goal is to create and algorithm that can help us easily identify corrupted data.
- Our approach is to use flags to indicate uncertainty in the value so that the user can decide what is suitable for the intended purpose.
- To start we created simple for loops that will flag the data that is not inside the range of the instrument's bounds
- Then we made these for loops a little more precise and divided the data into months.

Initial Algorithm

```
corrupt_humidity <- function (low_bound, up_bound){

relhumidity<- uiuc_weather2$relative_humidity

for(i in 1:length(relhumidity)){
  if((relhumidity[i]<low_bound)+(relhumidity[i]>up_bound))
  { relhumidity[i]<- 1
    }
    else{
      relhumidity[i]<- 0
    }
  }

  plot(relhumidity, col='blue')
}
```

Initial Algorithm

```
make_month_year <- function(DATAFRAME_TIME) {
  year <- c()
  month <- c()

  lol <- strsplit(toString(DATAFRAME_TIME),',')

  for(i in lol[[1]]) {
    tmp <- strsplit(i, '-')[[1]]
    tmp_year <- strsplit(tmp[1], ' ')[[1]]

    if(length(tmp_year) == 1) {
      year <- c(year, tmp_year[1])
    } else if(length(tmp_year) == 2) {
      year <- c(year, tmp_year[2])
    }
    month <- c(month,tmp[2])
  }
  return(data.frame(year, month))
}
```

Possible Flags Depending on the Data

Table 1. Examples of flags used to provide information about the data collected.

Type of flag	**Example**
Internal	
Missing value	No measured value available because of equipment failure or another reason
Low battery	Sensor battery dropped below a threshold
Calibration due	Sensor needs to be sent back to the manufacturer for calibration
Calibration expired	Value was collected with a sensor that is past due for calibration
Invalid chronology	One or more nonsequential date or time values
Persistent value	Repeated value for an extended period
Above range	Value above a specified upper limit
Below range	Value below a specified lower limit
Slope exceedance	Value much greater or lower than the previous value, resulting in an unrealistic slope
Spatial inconsistency	Value greatly differed from values collected from nearby sensors
Internal inconsistency	Value was inconsistent with another related measurement
Detection limit	Value was below the established detection limit of the sensor
External	
Pass	Value passed all quality control tests and is considered valid
Estimated	Estimated value from a model or other sources
Missing	Missing value
Uncertainty	Estimate of uncertainty of the value expressed as a percent

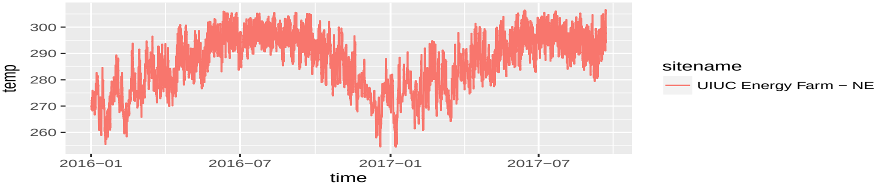
Note: Internal flags are for field technicians and data quality analysts; *external* flags are what the public sees.

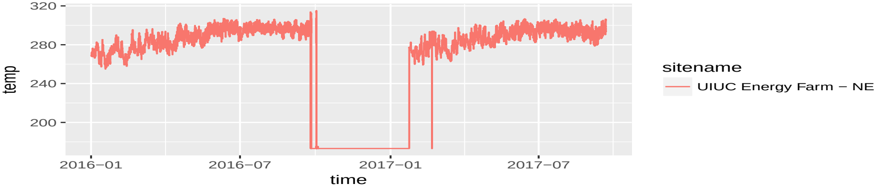
Figure 1: J. Campbell: http://www.bioone.org/doi/full/10.1525/bio.2013.63.7.10

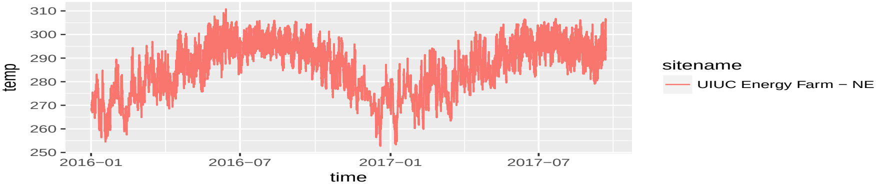
What can be done

- With more expertise and guidance a function can be made such that it flags any data point that provides erroneous information.
- The algorithm can be made such that it has the max and min of each day for the temperature, air pressure and relative humidity.
- Polish the code

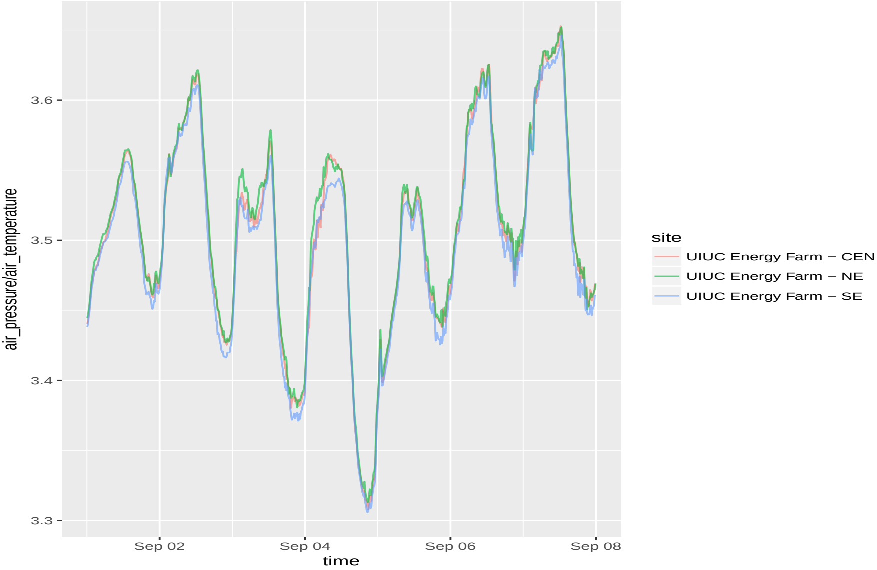
Initial Visualization of Data



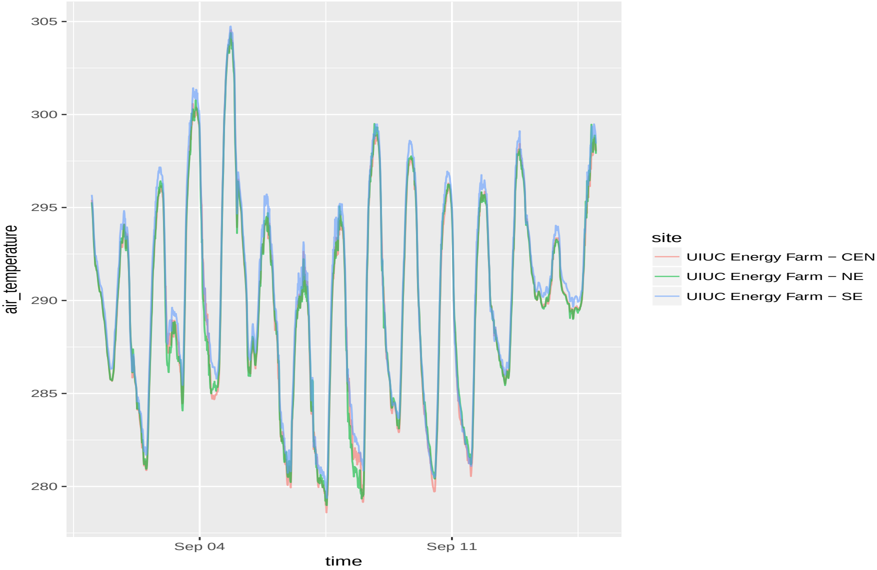




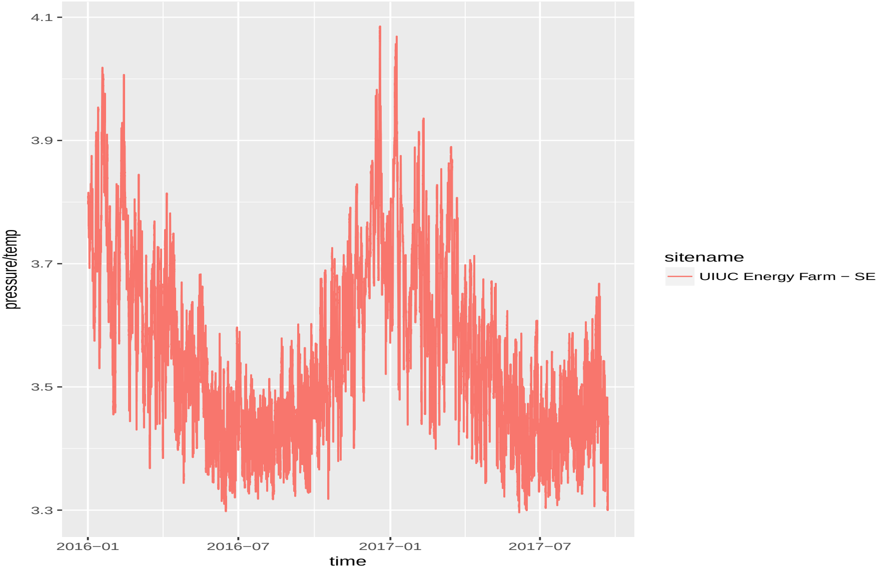
Initial Visualization of Data



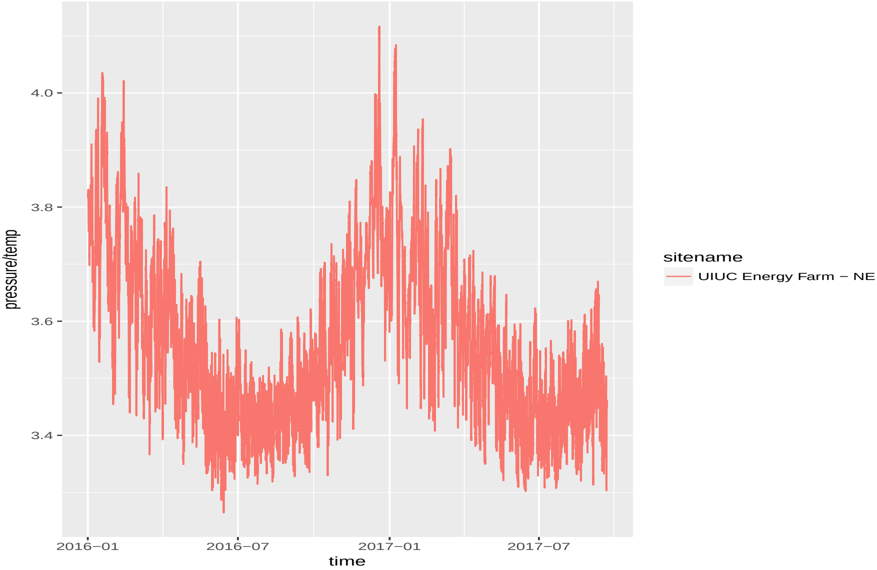
Initial Visualization of Data



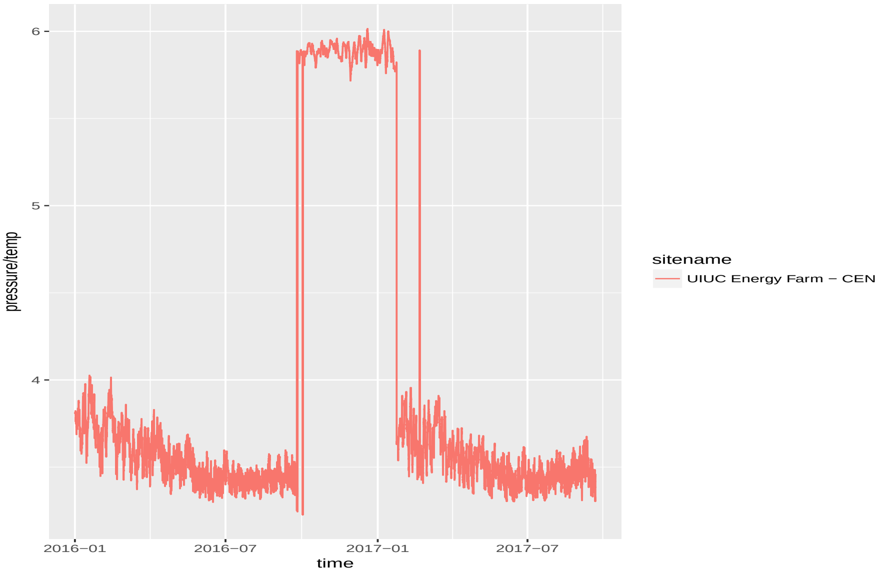
Initial Visualization of Data



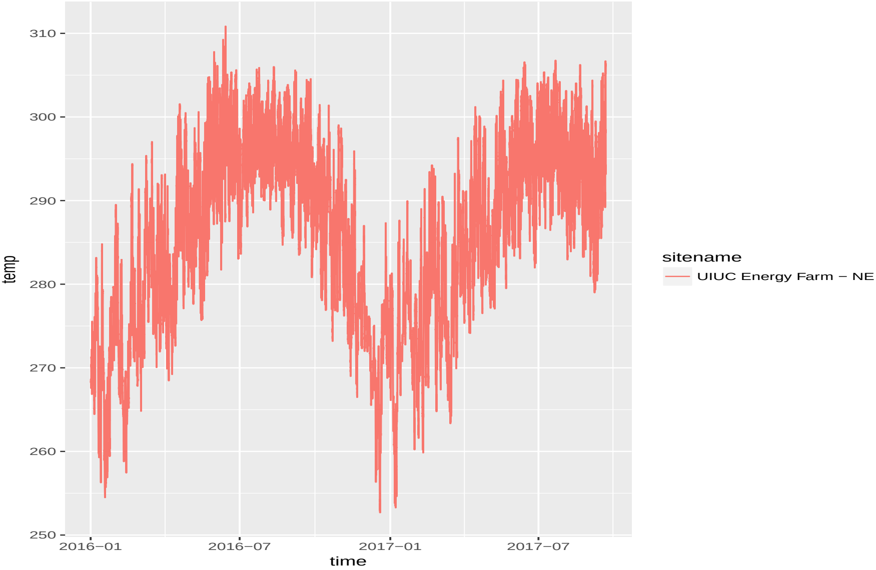
Initial Visualization of Data



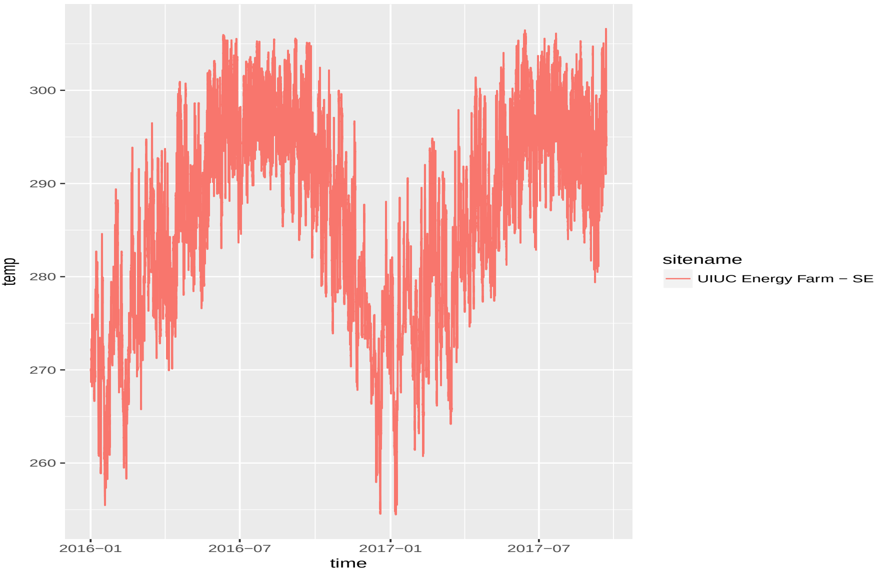
Initial Visualization of Data



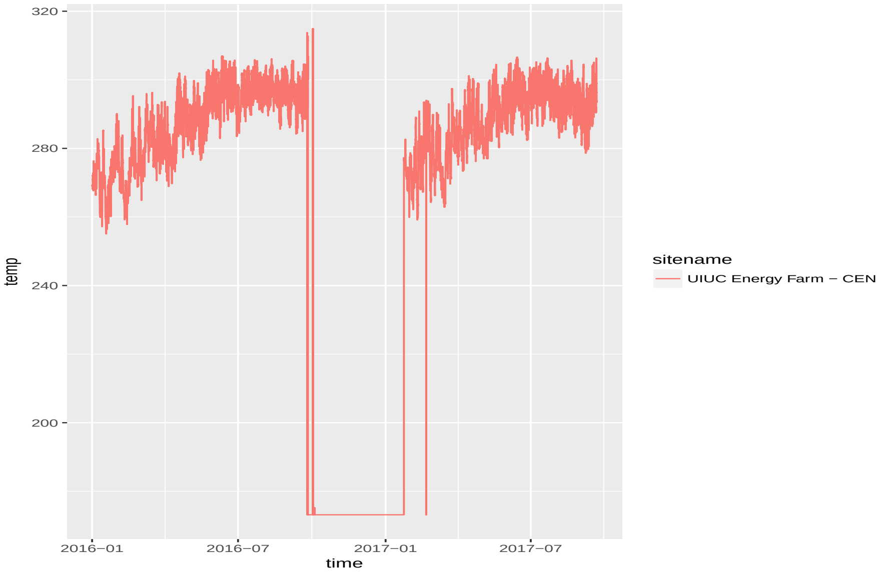
Initial Visualization of Data



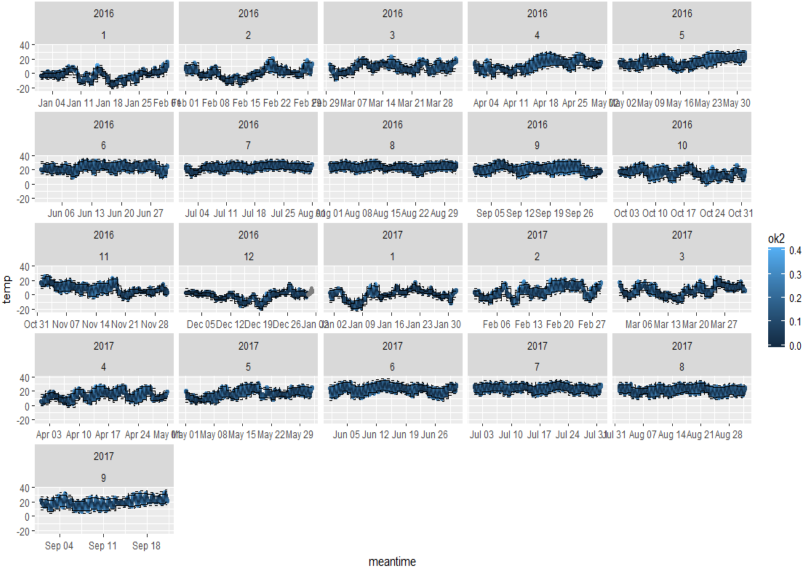
Initial Visualization of Data



Initial Visualization of Data



Visualization of Arizona Data Using Daymet Data



Anomalize

- R package used to automatically flag anomalies in time series data.
- http://www.business-science.io/code-tools/2018/04/08/introducing-anomalize.html

Bibliography

1. J. Campbell, L. Rustad, J. Porter, J. Taylor, E. Dereszynski, J. Shanley, C. Gries, D. Henshaw , M. Martin, W. Sheldon and E. Boose. *Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data*. BioScience, 63(7):574-585. 2013. URL: http://www.bioone.org/doi/full/10.1525/bio.2013.63.7.10
2. business-science.io. *"Anomalize: Tidy Anomaly Detection."* Business Science, 8 Apr. 2018, www.business-science.io/code-tools/2018/04/08/introducing-anomalize.html.
3. "Daymet V3." Daymet, daymet.ornl.gov/.