

# Guide 2: Cleaning and Pre-Processing Weather Station Data

Marc Los Huertos

February 6, 2024 (ver. 0.86)

## 1 Introduction

### 1.1 Goals

The goal of this guide is to provide a step-by-step process for cleaning and pre-processing weather station data. The data is from the Global Historical Climatology Network (GHCN) Daily dataset. The data is available from the National Oceanic and Atmospheric Administration (NOAA) and is available from the National Centers for Environmental Information (NCEI) at <https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatol>

### 1.2 Background

#### 1.2.1 What is GHCN-Daily?

The new dataset is referred to as GHCN-Daily, version 3. The new dataset has a number of improvements over the older dataset, including a more comprehensive and robust quality assurance process. The new dataset also includes more station records and is updated more frequently than the older dataset.

The new dataset is also available in a more user-friendly format, including a set of pre-packaged data files that are available for download from the NCEI website. However, developing instructions to interact with their website is way to time consuming! So, we'll use the R file transfer protocol, via https, to download the data.

In 2011, there was a major update to the daily updates to the GHCN dataset: ??IMMEDIATE – Changes to COOP Daily Form Access. It's a useful history.

### 1.3 Approach

We need to address several things that might get in the way of our analysis:

1. read station data into R
2. date format (convert to POSIXct, adding month and year as variables)

3. evaluate missing data
4. units conversion
5. outliers
6. Other stuff??

## 2 Cleaning and Pre-Processing Functions

### 2.1 Before Starting the Process

Before you begin, make sure you have the stations to read into R. Got to the “Files” tab in RStudio and make sure you see the station csv files in your data folder. If not, please Slack Marc and mentors, so we can troubleshoot the issue!

### 2.2 R Code with Custom Functions

From the Canvas page, go to the Guide1functions.R file and download the file to your computer. Then upload the file to Rstudio directory you are using for the project.

Open the file in Rstudio and run the code, using the “source”. button near the top of the editor window.

Run the [Guide2functions.R](#) code and the functions will be loaded into your environment automatically.

### 2.3 Function Descriptions and Use

The following custom functions are used to read the weather stations data into R, clean and pre-process data so we can analyze the data with the next guide.

### 2.4 Reading csv and Checking dataframe objects

**Read all csv files into R** The If the data are not in the R environment, you can read them in using the following but you have the uncompressed files, you can read them in using the following function:

```
## function (datafolder)
## {
##   my.stations = read.csv(paste0(datafolder, "my.inventory.csv"))
##   colnames <- c("ID", "DATE", "ELEMENT", "VALUE", "M-FLAG",
##                 "Q-FLAG", "S-FLAG", "OBS-TIME")
##   for (i in 1:nrow(my.stations)) {
##     assign(my.stations$ID[i], read.csv(paste0(datafolder,
##         noquote(my.stations$ID[i]), ".csv"), header = TRUE,
##         col.names = colnames), envir = parent.frame())
##   }
## }
```

```
## }
```

Example of how to use the function:

```
datafolder = "/home/mwl04747/RTricks/04_Regional_Climate_Trends/Data/SP24/"  
ReadStations2.fun(datafolder)
```

**Check the dataframes in the environment** You can do this by running the following code:

```
ls()  
  
## [1] "CleanUp.fun"          "ConvertUnits.fun"    "coverage.fun"  
## [4] "datafolder"           "fixdates.fun"        "MonthlyAnomalies.fun"  
## [7] "MonthlyNormals.fun"   "MonthlyValues.fun"   "QAQC.fun"  
## [10] "ReadStations.fun"     "ReadStations2.fun"   "USC00040693"  
## [13] "USC00041614"          "USC00042294"         "USC00043157"  
## [16] "USC00044412"          "USC00046074"         "USC00046136"  
## [19] "USC00046719"          "USC00046826"         "USC00047821"  
## [22] "USC00047902"          "USC00047916"         "USC00048351"  
## [25] "USC00049866"          "USW00023271"
```

**Check the structure of the dataframes** Using `str()`, make sure the datasets look right!

For example:

```
## 'data.frame': 225076 obs. of 8 variables:  
## $ ID : chr "USC00043157" "USC00043157" "USC00043157" "USC00043157" ...  
## $ DATE : int 18670601 18670602 18670603 18670604 18670605 18670606 18670607 18670608 ...  
## $ ELEMENT : chr "PRCP" "PRCP" "PRCP" "PRCP" ...  
## $ VALUE : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ M.FLAG : chr "" "" "" "" ...  
## $ Q.FLAG : chr "" "" "" "" ...  
## $ S.FLAG : chr "F" "F" "F" "F" ...  
## $ OBS.TIME: int NA NA NA NA NA NA NA NA NA NA ...
```

## 2.5 Clean Data

Next we'll "clean" the dataset by fixing the date format and preparing it for the analysis stages.

### Function to Fix Dates `## function (station)`

```
## {  
##   station$Ymd = as.Date(as.character(station$DATE), format = "%Y%m%d")  
##   station$MONTH = as.numeric(format(station$Ymd, "%m"))  
##   station$YEAR = as.numeric(format(station$Ymd, "%Y"))  
##   return(station)  
## }
```

Example of how to use the function (Note: the new dataframe name change):

```
USC00042294a <- fixdates.fun(USC00042294)
```

### Evaluation Data Coverage We need to know how much data we have for each station. This is important for the next steps in the process.

```
## function (station, element = "TMAX")  
## {  
##   Dates.all = data.frame(Ymd = seq.Date(from = min(station$Ymd),  
##     to = max(station$Ymd), by = "day"))  
##   station.full = merge(Dates.all, station, all = TRUE)  
##   station.coverage = sum(!is.na(station.full$VALUE[station.full$ELEMENT ==  
##     element]))/length(station.full$VALUE[station.full$ELEMENT ==  
##     element]) * 100  
##   return(round(station.coverage, 2))  
## }
```

Example of how to use the function:

```
coverage.fun(USC00042294a)
```

In general, we want something like 95% coverage for the period of record. If you don't have that, please let us know and we'll help you get additional stations with better coverage!

### Function to Convert Units According to the NOAA website, the csv.gz files have five core elements include the following units, which we will convert:

**PRCP** = Precipitation (tenths of mm) → mm

**SNOW** = Snowfall (mm) → cm

**SNWD** = Snow depth (mm) → cm

**TMAX** = Maximum temperature (tenths of degrees C) → degrees C

**TMIN** = Minimum temperature (tenths of degrees C) → degrees C

```
## function (station)
## {
##     station$VALUE = station$VALUE/10
##     return(station)
## }
```

Example of how to use the function (Note: the new dataframe name change):

```
USC00042294b <- ConvertUnits.fun(USC00042294a)
```

**Checking for Outliers** NOAA website conducts a very rudimentary data quality check, but every year, we find stations with wacky numbers. Hopefully this function will find them. But if you do have some, let Marc and mentors know so we can figure out how to address them!

Here are some stuff we'll look for:

**Extreme Values** Plot values with time, is the scale crazy with just a few observations at the extreme?

**Sudden Shift** ???

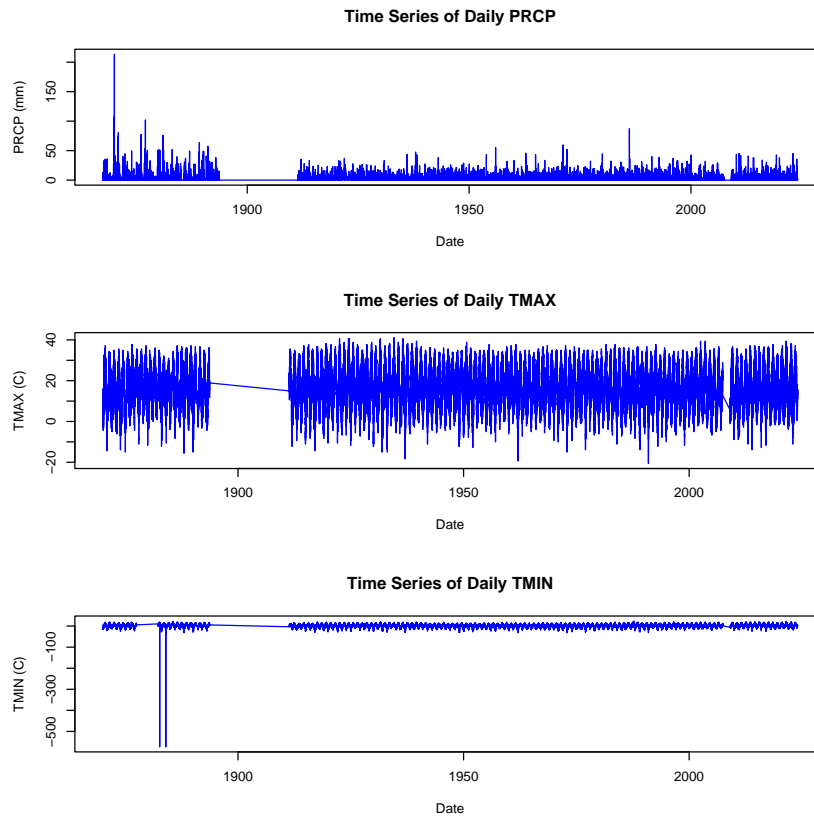
**Note QA/QC Flags** TBD

Yes, I am still working on this. QA/QC is a big deal and I want to make I can explain it AND make sure the code works.

```
## function (station)
## {
##     par(mfrow = c(3, 1))
##     plot(VALUE ~ Ymd, data = subset(station, subset = ELEMENT ==
##         "PRCP"), type = "l", col = "blue", main = "Time Series of Daily PRCP",
##         xlab = "Date", ylab = "PRCP (mm)")
##     plot(VALUE ~ Ymd, data = subset(station, subset = ELEMENT ==
##         "TMAX"), type = "l", col = "blue", main = "Time Series of Daily TMAX",
##         xlab = "Date", ylab = "TMAX (C)")
##     plot(VALUE ~ Ymd, data = subset(station, subset = ELEMENT ==
##         "TMIN"), type = "l", col = "blue", main = "Time Series of Daily TMIN",
##         xlab = "Date", ylab = "TMIN (C)")
##     station = subset(station, Q.FLAG != "")
##     station = subset(station, M.FLAG != "")
##     station = subset(station, S.FLAG != "")
##     return(station)
## }
```

Example of how to use the function:

```
QAQC.fun(USC00042294b)
```



If you have any hints of data problems, let's sort them out together!

**Function to Create Monthly Values and Normals** For TMAX and TMIN, we want monthly means, for rainfall, we'll want monthly totals.<sup>1</sup>

Here's the function for Monthly Values:

```
## function (x)
## {
##   x.TMAX.monthly = aggregate(VALUE ~ MONTH + YEAR, data = subset(x,
##     ELEMENT == "TMAX"), mean)
##   names(x.TMAX.monthly) <- c("MONTH", "YEAR", "TMAX")
##   x.TMIN.monthly = aggregate(VALUE ~ MONTH + YEAR, data = subset(x,
##     ELEMENT == "TMIN"), mean)
```

<sup>1</sup>Brody/Evlyn: We need code to exlue months with missing data, these might not be reprenesative of the month if missing, especically for PRCP!

```
##      names(x.TMIN.monthly) <- c("MONTH", "YEAR", "TMIN")
##      x.PRCP.monthly = aggregate(VALUE ~ MONTH + YEAR, data = subset(x,
##        ELEMENT == "PRCP"), sum)
##      names(x.PRCP.monthly) <- c("MONTH", "YEAR", "PRCP")
##      return(list(x.TMAX.monthly, x.TMIN.monthly, x.PRCP.monthly))
##    }
```

Here's the function for Monthly Normals:

```
## function (x)
## {
##   x.normals = subset(x, Ymd >= "1961-01-01" & Ymd <= "1990-12-31")
##   x.TMAX.normals.monthly = aggregate(VALUE ~ MONTH, data = subset(x.normals,
##     ELEMENT == "TMAX"), mean)
##   names(x.TMAX.normals.monthly) <- c("MONTH", "NORMALS")
##   x.TMIN.normals.monthly = aggregate(VALUE ~ MONTH, data = subset(x.normals,
##     ELEMENT == "TMIN"), mean)
##   names(x.TMIN.normals.monthly) <- c("MONTH", "NORMALS")
##   x.PRCP.normals.month.year = aggregate(VALUE ~ MONTH + YEAR,
##     data = subset(x.normals, ELEMENT == "PRCP"), sum)
##   x.PRCP.normals.monthly = aggregate(VALUE ~ MONTH, data = subset(x.PRCP.normals.m
##     mean)
##   names(x.PRCP.normals.monthly) <- c("MONTH", "NORMALS")
##   return(list(x.TMAX.normals.monthly, x.TMIN.normals.monthly,
##     x.PRCP.normals.monthly))
## }
```

Example of how to use the functions:

```
USC00042294.monthly <- MonthlyValues.fun(USC00042294b)
USC00042294.normals <- MonthlyNormals.fun(USC00042294b)
```

**Function to Create Anomalies** Here's the function:

```
## function (station.monthly, station.normals)
## {
##   for (i in seq_along(station.monthly)) {
##     TMAX <- merge(station.monthly[[1]], station.normals[[1]],
##       by = "MONTH")
##     TMAX$TMAX.a = TMAX$TMAX - TMAX$NORMALS
##     TMAX$Ymd = as.Date(paste(TMAX$YEAR, TMAX$MONTH, "01",
##       sep = "-"))
##     TMIN <- merge(station.monthly[[2]], station.normals[[2]],
##       by = "MONTH")
```

```
##      TMIN$TMIN.a = TMIN$TMIN - TMIN$NORMALS
##      TMIN$Ymd = as.Date(paste(TMIN$YEAR, TMIN$MONTH, "01",
##                               sep = "-"))
##      PRCP <- merge(station.monthly[[3]], station.normals[[3]],
##                   by = "MONTH")
##      PRCP$PRCP.a = PRCP$PRCP - PRCP$NORMALS
##      PRCP$Ymd = as.Date(paste(PRCP$YEAR, PRCP$MONTH, "01",
##                               sep = "-"))
##      return(list(TMAX = TMAX, TMIN = TMIN, PRCP = PRCP))
##    }
## }
```

Example of how to use the function:

```
USC00042294.anomalies <- MonthlyAnomalies.fun(USC00042294.monthly, USC00042294.normals)
```

## 2.6 Checking on the Results

Getting into the data is a bit tricky. The datasets is a list of dataframes. Each dataframe is a different variable, where 1 is TMAX, 2 is TMIN, and 3 is PRCP.

The get access to each you use the following code:

- `USC00042294.anomalies[[1]]` for TMAX
- `USC00042294.anomalies[[2]]` for TMIN
- `USC00042294.anomalies[[3]]` for PRCP

## 2.7 Clean Up R Environment

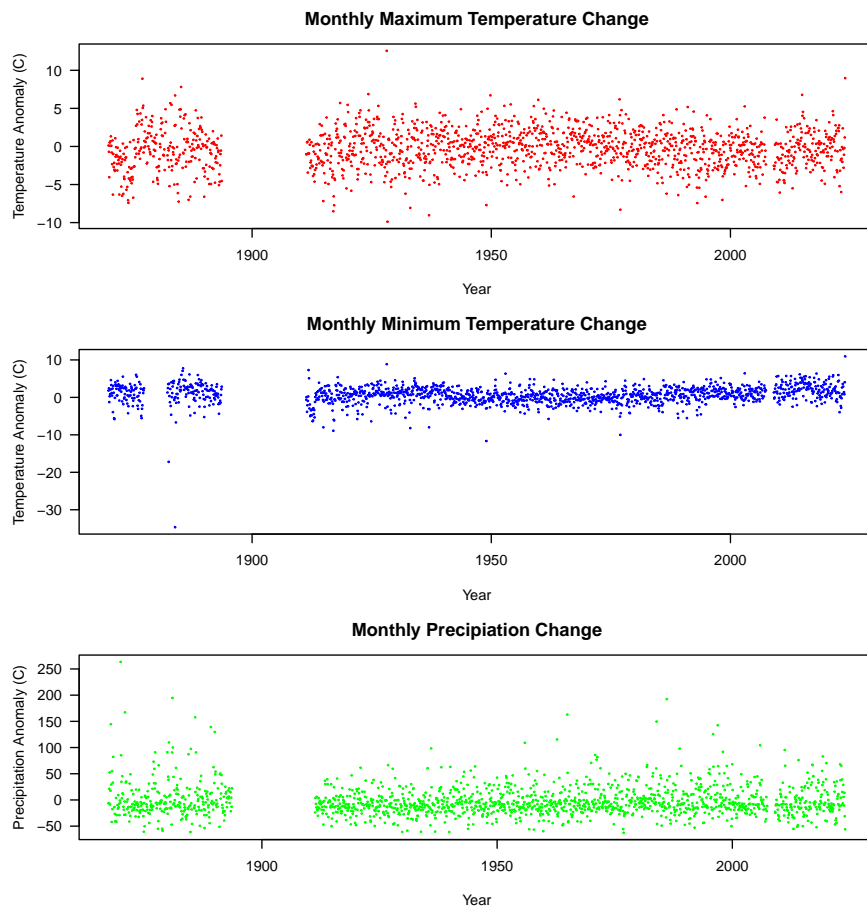
I tend to avoid having lots of objects in my R environment. I like to clean up after myself.

Also, I like to move the final products into csv files. Here's the function to do that. However, I noticed that it's deleting needed files. YIKES! It's also clunky, but not that important at this point, nevertheless, I'll work on it later. So, I suggest you don't run this code until I fix it.

```
CleanUp.fun(datafolder, USC00042294.anomalies, "USC00042294")
```

Your are done with this guide, now take a break, a walk, and enjoy some downtime without looking at a computer. Next, go to Guide 3!





### **3 Explaining the Code, TBD!**

#### **3.1 QA/QC: Troubleshooting**

I will be getting all the guides working before working on this! But if there are errors with the custom function, this is where workarounds will be described! Please Slack me and mentors if you have any problems!

##### **3.1.1 Coverage Problems (<95%)**

If you have too many stations that don't have enough data, we'll need to download additional stations. I can show you a way to decipher that ahead of time if you'd like to know. Otherwise, I suggest you double the number of stations selected in the code that generated my.inventory. I have increase the number of stations per state to 15, so perhaps that will give you enough to work with.

##### **3.1.2 Missing Data**

Similar to the problem above, missing data can be a problem. The real issue that I can see is that rainfall is so central because we use the data to calculate monthly totals, but if the month is missing data, then we have a severe bias. Same issue in temperature, but because we are looking at averages, it's less likely to be a major source of bias. But we should should check!

##### **3.1.3 Plot Anomaly**

Graphic has lots of issues. more next time! But here's a start.

```

options(scipen=5)
par(mar=c(4,6,2,5))

plot(ANOMALY ~ YEAR, data = subset(station1.TMAX, MONTH == 1),
     las=1, pch=19, col = "blue", cex=.5, #xlab = "Year",
     ylab = "Maximum Temp Anomaly (C)",
     main="January Maximum Temp Anomaly")
mtext("Maximum Temp Anomaly (C)", side = 2, line = 3)
temp.lm = lm(ANOMALY ~ YEAR, data = subset(station1.TMAX, MONTH =
abline(coef(temp.lm), col = "red")

```

We can see huge periods of time where no data was collected. Yikes! I don't think I can use this station.

My custom functions are probably sensitive to missing values, need to work on that!

### 3.2 Work Arounds

TBD!

## 4 Next Steps

### 4.1 Apply Function to All Stations

So far, I have only run function for 1 station, but I suspect you can figure out how to run it for each one!

This is all we need to do so far. Next week, we'll look at different way to visualize the data!

### 4.2 Clean Up R Environment

I'll save all the station data into csv files, then use them in the next guide to clean, process, and visualize data. I

don't think the function is all that useful, so I can show you better ways of doing this in class.

```
station1.clean=cleandataframe.fun(station1)
```