The US Temperature Record 2: Twelve Data Sets

/ SEP 22, 2021

At the USHCN data store there are four measurements, or "elements," available:

- prcp: precipitation totals for each month
- tavg: the monthly average of the daily average temperatures [$(T_{min}+T_{max})/2$]
- tmax: the monthly average of the daily high temperatures
- tmin: the monthly average of the daily low temperatures

Question: why is the data stored by monthly average, when we have (at least I have) a weather station reporting the weather to national databases every five minutes? Because that's how it was reported until the 1990's. That's how the data was collected for most of the database existence, and they just kept it going. The data is far easier to gather now: no humans involved in reading, recording and resetting the thermometers each day, no emptying the rain trap, no missed days, no averages to calculate by hand, no reports lost in the mail or not sent. You can get an idea what these monthly reports look like from my weather station NOAA data page here.

And for each, three files, or "datasets," are available:

- raw
- tob
- FLs.52j

Twelve data sets in all, supposedly representing one set of measurements. The trouble is those three types of data. Let's see what they are:

raw:

raw data come directly from the reports received. Thermometer readings as they were reported each month.

tob:

Data which has been corrected for time of observation. If the thermometer was observed at 10 am, T_{max} represents yesterday's high, while T_{min} is that morning's low. The tob correction is supposed to correct for that, which should mean very little to the monthly average and almost nothing to the yearly averages as it shifts the days by one. Scientifically I have a big problem with using this data, as they have changed the primary data. You should never do that. You can adjust the model using the data, but data is the only truth we have in science, and holds a special, inviolable place.

FLs.52j:

This is a far more extreme correction, using what is called the "pairwise homogenization algorithm," the PHA. This is an attempt to level off any variation in the monthly temperature series by comparing each station's monthly averages to those of a nearby station. I have a real problem with this fiddling with the data. They are attempting to solve two problems with one correction: variability in the time series (caused by changes in the measuring equipment or housing), and variability in the spatial series (variability in the temperatures recorded by nearby stations the same day, caused by changes in land use around the station, new roads, even tree growth nearby). It's an attempt to remove variability in the data, which is done so it matches the models better. If there is variability in the data, the model should always reflect that variability; only a fool would change the data to make it match the model better. And this dataset is the tenth ("jth") iteration of the version 2.5 algorithm, meaning they got it wrong nine times in a row but still trust the PHA. Most of us walk away from a bad restaurant after one bout of food poisoning; these guys are eating at the same place ten days straight! Someone needs to explain that to me.

Revision #1
Created 7 April 2024 19:05:33 by bruce
Updated 7 April 2024 19:06:35 by bruce