Building and updating the daily and monthly meteorological data sets

For the past 20 years or so, I have been maintaining a set of subbasin/lake data sets that are associated with a tech memo GLERL-83. The software used for that process has varied over the years as the source data repositories have evolved and changed. In the beginning, the data was obtained on magnetic tape from agencies like the National Climatic Data Center, USGS, and Environment Canada. As the internet became “a thing” the data transmission process became ftp-based (after requesting via email, etc) and is now largely a matter of using standardized data retrieval systems. My software and procedures have usually been documented to one degree or another (sometimes sketchy, sometimes detailed), but never put together into a comprehensive unit. As I am approaching retirement in a few years, I have decided that it is time to pull everything together into a single place so that my successor will have a clear set of instructions to carry on whatever part of this work is deemed relevant.

The reason I stipulate that it may not be deemed relevant, that most/all of this work was originally intended to only support our own internal use, but then was expanded into a public thing as a service to the community. The publication of these data has never actually been part of GLERL’s mission. Other entities either have a specific mandate or are more appropriate for this task due to their operational nature, in my opinion. So why did this become a GLERL thing? That is kind of a long story, and I don’t know all the details. Suffice it to say that there have been “turf battles” between agencies and individuals (who will get credit for the science), abdication of responsibility (GLERL is doing it so we will just use their numbers instead of doing it ourselves) and probably other reasons. In the beginning (pre-internet), pulling all of these monthly data together into a single place had the potential for great usefulness. Access to the raw station data was difficult and expensive, and methods for processing it were not so clear. But now, the raw station data are readily available from the source agencies, and software for aggregating the data are easily implemented (e.g. using GIS systems). The key thing is that there are MANY ways to aggregate the data (weighting methods, map delineations, station selection criteria, etc.) We essentially have our one method. But what I often run into when someone contacts me for help with the data is that they have calculated a similar data set (or seen someone else’s) and have different values from what we have published. Often the other data set has been produced using a fine-toothed comb to select the stations, root out bad data values, etc. Our data, on the other hand, is produced by a simpler and less labor-intensive process. We use all (or just about all) of the station data we can get our hands on (from the sources we use) and throw it all into the bucket. Simple range checks are done in an attempt to remove the worst of the data values, but we don’t dig deep. The result is a data set that is good, but not great.