Merging Methodology

Author: Jared Rennie

Contributors: Peter Thorne, Jay Lawrimore

10/4/12

This document covers, at a high level, the process in which individual stage two sources are combined to form a comprehensive stage three dataset. Because many sources may contain records for the same station it is necessary to create a process for identifying and removing duplicate stations, merging some sources to produce a longer station record, and in other cases for determining when a station should be brought in as a new and distinct record.

The beta release of the databank includes a recommended version of the stage three dataset, along with seven variants. The majority of this document will highlight the steps towards creating the recommended version. A description of the variants will be provided in Part 6.

This document will be superseded, when published, by a fuller exposition of the databank merge methodology currently under preparation.

Part 1: Source Hierarchy

Before a merge is performed, a hierarchy of all the source datasets within the databank is created. Sources with higher priority will take precedence over lower priority sources when more than one record for the same station and same period of time exists. The priority that one source may have over another is based on a number of criteria. Sources that have better data provenance, have extensive metadata, come from a national or international holding, or have long and consistent periods of record are in the upper tier of the hierarchy.

Because of the emphasis the International Surface Temperature Initiative places on data provenance, the stage three databank holdings are envisaged to constitute as close to the raw data as possible. Ideally data should be tracked as far back as the original hard copy. In addition, data rescued in the recent past when the importance of such provenance has been explicitly recognized is given higher preference. Other data given higher preference during merging include monthly mean maximum and minimum temperature. This is preferred over monthly mean temperature because they can be directly used to calculate the monthly mean and because there is compelling evidence that many data artifacts affect max and min differently.

With this framework in mind, 45 stage two sources have been prioritized using the previously defined structure. The raw version of GHCN-D, priority number one, is considered the highest, or target dataset. This provides a backbone of maximum/minimum values that is analyzed regularly and curated carefully on an ongoing basis with regular updates. As GHCN-D is assumed to be the Stage 3 daily source in the medium term this also enforces vertical coherency between the daily and monthly holdings enabling investigators to delve back to a finer temporal data resolution where the daily data exists to investigate individual stations or values in further detail.

Part 2: General Description of Merge

The merge process occurs in an iterative fashion, starting from the highest priority data source (target) and running progressively through the other sources (candidate). The merge process is designed to be broadly-speaking Bayesian in approach and based upon metadata matching and data equivalence criteria. The highest deck is read in and compared to the source of next lower priority. Every candidate station makes comparisons with the target stations, and one of three possible decisions is made. First, a station match is found and data should be merged. Second, the candidate station is considered unique and added to the target dataset. Third, there is not enough / conflicting / ambiguous information and the candidate station is withheld. After all stations in the two sources are tested and combined into a new merged source dataset, the process then applies the same tests using the next source in succession.

All of the sources are run through, looking only at stations which have TMAX and TMIN. Afterwards, the target dataset generates TAVG (simply the average of the two), and the sources are checked a second time, only looking at stations which have TAVG. Once this has been completed, target stations with less than 12 months of data are removed and the result is the final, merged stage three dataset.

Part 3: Metadata Comparisons

Each candidate station runs through all the target stations and calculates three metadata criteria as the first test to identify matching stations. This is necessary because the same station may have different precision / values for longitude, latitude and elevation between decks and the name may also differ, particularly for countries which were once colonial and have subsequently gained independence.

Using the stations latitude and longitude, the geographical distance between the two stations is computed. The distance is then fitted to an exponential decay function (which decays to zero at 100Km distance), and a probability that the two stations are the same is determined. Next, the same approach is performed using the height difference between two stations (here the exponential decays to zero at 500m height difference). Third, the similarity of the station name is considered. This is done using the Jaccard

Index (JI), which is defined as the intersection divided by the union of two sample sets, A and B:

$$JI = \frac{|A \bigcap B|}{|A \bigcup B|}$$

In other words, JI will look for cases in which certain letters exist in both station names, as well as the number of times letters occur in one name, but not in the other. Once the ratio is known, a probability is calculated. One caveat to JI is that it does not take into account the position of the character within the word. Therefore anagrams (i.e. TOKYO and KYOTO) would have a perfect JI of 1.

The three metadata metrics based on location, elevation, and station name each have a prior probability from 0 to 1. Using a Bayesian approach, the probabilities are combined to form a posterior probability of possible station match, known simply as the *metadata probability*:

$$metadata\ probability = \frac{(9 * dist) + (1 * height) + (5 * JI)}{15}$$

Weights are given to each metric. Since the latitude and longitude are one of the premier methods of determining station match, it is given the highest weight. The height of the station can sometimes be misleading or inaccurate, so it is given the lowest weight. If this probability surpasses a threshold of 0.50, an evaluation based on data comparisons is then made. The threshold is set relatively low to account for possible errors in the metadata. If the elevation probability is missing, the equation is readjusted for only distance and JI.

If none of the comparisons between the candidate station and all the target stations pass the metadata threshold, it then checks the validity of each individual metadata metric. If two of the three metrics (dist, height or JI) are greater than 0.90, then there is the possibility that incorrect or missing metadata within the candidate station has altered the overall metadata probability and it is withheld. If this is not the case, it is determined that the candidate station is unique and it is added to the target dataset without any further tests being performed.

Part 4: Data Comparisons

If any of the target stations passes the metadata threshold, a data comparison is made between that target station and candidate station. In order to have a direct and significant data comparison, there is an overlap threshold that must exist between the two stations. The default is 5 years, or 60 months. If this threshold is passed, then the data comparison is made using the Index of Agreement (Willmott 1981).

The Index of Agreement is a “goodness-of-fit” measure and is defined as the ratio between the mean square error and the potential error. It was designed to overcome the insensitivity of correlation measures such as the coefficient of determination. Legates and McCabe (1999) argued that the sensitivity of outliers would lead to high values due to the squaring of the difference terms. Their modified Index of Agreement removed the squared term, and is the equation used in the data comparison:

$$IA = 1.0 - \frac{\sum_{i=1}^{n} |T_i - C_i|}{\sum_{i=1}^{n}(|C_i - \bar{T}| + |T_i - \bar{T}|)}$$

Between a candidate station (C) and a target station (T), IA is applied twice, one to the overlapping TMAX period and the other TMIN, and resulting values range between 0 and 1. While these are considered probabilities of station match, there is no way of taking into account how many months of overlap occur. The minimum requirement is 5 years, but there could be as many as 25 years, or even 50 years. This may lead to a bias, giving preference to longer periods of overlap.

To account for this a lookup table was generated to provide a probability of station match (H1), as well as station uniqueness (H2). Bootstrapping was applied by changing the shifts in mean and variance of certain criteria, and running IA 1,000 times. For station match, shifts were applied using a station with a long period of record. For our purposes, the station from De Bilt, Netherlands was used, since data has been continuous since 1706 for TAVG (1901 for TMAX and TMIN). For station uniqueness, statistics derived from stations that were within 50Km of the candidate station in GHCN-D were used to derive reasonable expectations of how distinct nearby stations may be expected to differ on a month-to-month basis. Using these results, a cumulative distribution function is fit for each contingency (same station and unique station) and stratified by the overlap period. The higher the overlap period, the more perfect IA needs to be in order to be considered a station match.

This data comparison is applied to all the target stations that could match with the candidate station according to the metadata test. If no data comparison was made, then there was insufficient overlap period between the candidate and target stations. In this case the final decision is based solely upon the *metadata probability*. Because of this the metadata comparisons need to be near perfect, so the *metadata probability* threshold is increased from 0.50 to 0.85. If the highest metadata comparison with a target station received a *metadata probability* larger than this new threshold, then the candidate station merges with that station. Otherwise it is withheld.

There are also cases where data comparisons were made, but the *metadata probability* of a non-overlap case was higher than any of the overlap cases. If this is found to be true, then that target station is merged with the candidate station. Otherwise there are five resulting probabilities, one *metadata probability*, and four data probabilities (tests for station match and uniqueness, for both TMAX and TMIN). These prior probabilities are then recombined to form two new posterior probabilities, one of station match, and one of station uniqueness. The unique equation was structured so it favors a lower *metadata probability* (near 0.50), and because it is not weighted, this value can range between 0.50 and 2.50.

$$posterior\ probability\ same_{TMAX/TMIN} = \frac{metadata\ probability * H1_{tmax} * H1_{tmin}}{3}$$

$$posterior\ probability\ unique_{TMAX/TMIN} = (1 - metadata\ probability) + H2_{tmax} + H2_{tmin}$$

Once these posterior probabilities are made for all possible comparisons between a candidate station and its target stations, thresholds are set for station match and uniqueness (0.50 and 1.30 respectively) to determine the final fate of the candidate station. If any of the *posterior probability same* probabilities exceed the threshold, then the candidate station is merged with the target station with the highest *posterior probability same*. If none of the stations exceed that threshold, but one of the *posterior probability unique* values exceeds the unique threshold, then the candidate station becomes unique and is added to the target dataset. If no probabilities pass either threshold, then the station is withheld.

If merging of data is performed, only data from the candidate station that are not already in the target station record are added to create the new merged record. If data occurs for both the candidate station and the target station, preference is always given to the target, since it contains data that were higher in the prioritized list. The merging appends data from the candidate to the target to create a single record. No candidate data are inserted into the middle of the target series unless they could fill a string of at least 5 consecutive years of missing data. Data segments can be added to a single station from multiple sources through the iterations across source decks.

Data comparisons of TAVG are similar to those of TMAX and TMIN, with the exception of the final posterior probabilities. This is because of only one temperature variable (TAVG) instead of two (TMAX and TMIN):

$$posterior\ probability\ same_{TAVG} = \frac{metadata\ probability * H1_{tavg}}{2}$$

$$posterior\ probability\ unique_{TAVG} = (1 - metadata\ probability) + H2_{tavg}$$

Part 5: Validation

All of the decisions made that created the recommended version of the merge were tested against an independent dataset. The dataset is a subset of GHCN-Daily, and contains known time of observation biases. Some of these stations have been corrected for this bias and added to GHCN-Daily, but not all. Because GHCN-Daily is the first priority in the source hierarchy, this pseudo-source is considered the candidate source and is tested against GHCN-Daily. This is performed to prevent type I and type II errors, in which stations that should be merging are not, or to ascertain if stations are merging incorrectly. Out of the 945 stations that should be in GHCN-D, 776 (82.12%) were selected to merge, 85 (8.99%) became unique, and 84 (8.89%) were withheld. Out of the 776 that were chosen to merge, 729 (93.94%) merged with the correct station.

Part 6: Variants of merge program

The following are 8 thresholds that can be defined by the user in the program:

metadata_threshold: the first metadata threshold that takes into account the distance, height, and jaccard probabilities (default is 0.50)

- Increasing this value will tend to pull more through as unique stations
- Decreasing this value will lead to more data comparisons

metadata_threshold2: the second metadata threshold used if there is no overlap period between the target and candidate station (higher than the first metadata threshold) (default is 0.85)

- Increasing this value will tend to withhold more stations
- Decreasing this value will lead to more merging of stations

posterior_threshold_same_txn: threshold where TMAX/TMIN candidate station has to exceed in order to merge with the target station (default is 0.50)

- Increasing this value will tend to make stations either unique or withheld
- Decreasing this value will lead to more merging of stations

posterior_threshold_unique_txn: threshold where TMAX/TMIN candidate station has to exceed in order to be considered a unique station (default is 1.30)

- Increasing this value will tend to withhold more stations
- Decreasing this value will lead to more unique stations

posterior_threshold_same_txn: threshold where TAVG candidate station has to exceed in order to merge with the target station (default is 0.50)

- Increasing this value will tend to make stations either unique or withheld
- Decreasing this value will lead to more merging of stations

posterior_threshold_unique_txn: threshold where TAVG candidate station has to exceed in order to be considered a unique station (default is 0.90)

- Increasing this value will tend to withhold more stations
- Decreasing this value will lead to more unique stations

overlap_threshold: overlap period that must exist between the target and candidate station in order to calculate a data comparison via the Index of Agreement (default is 60 months)

- Increasing this value will tend to pull more through as unique stations
- Decreasing this value will lead to more data comparisons

gap_threshold: gap period that must exist when merging a candidate station with the target station (default is 60 months)

- Increasing this value will lower the number of merges
- Decreasing this value will increase the number of merges

Changing these thresholds can significantly alter the overall result of the program. The same can be said when the source hierarchy is changed. In order to characterize the uncertainty of the program, seven different variants of the stage three product are made available alongside the recommended. A description of each variant is below:

Variant One (colin)

In this variant, the source deck is shifted to prioritize sources that originated from their respective National Meteorological Agencies (NMA's). This way, the most up to date locally compiled data is favored over consolidated repositories, which may or may not be up to date. In addition, sources that are either raw or quality controlled are favored over homogenized sources.

Variant Two (david)

Here, NMA's are favored, having TMAX, TMIN, and comprehensive metadata as the highest priority. The overlap threshold is lowered from 60 months to 24 months, in order for more data comparisons to be made.

Variant Three (peter)

The source deck is changed under the following considerations. No TAVG source (or data from mixed sources) is ingested into the merge. This is because there is uncertainty in the calculation of TAVG (ie, it is not always TMAX+TMIN/2). TAVG in the final product is only generated from its respective TMAX and TMIN value. For the remaining sources, GHCN-D is the highest priority, and the rest are ranked by order of longest station record present within the source deck, from longest to shortest. The metadata equation is changed to give weighting to the distance probability (10) over the height (1) and Jaccard (1) probabilities (default is 9, 1, and 5, respectively). Finally the thresholds to merge and unique the station are lowered and favored to merge more stations.

Variant Four (jay)

Within the algorithm, the data comparison test results in three distinct possibilities. The station is merged, unique, or withheld. In this variant, this is altered so the candidate station is either merged or unique.

Variant Five (matt)

All homogenized sources are removed. Nothing else is altered compared to the recommended merge.

Variant Six (more-unique)

Thresholds are adjusted to make more candidate stations unique, thus increasing the overall station count.

Variant Seven (more-merged)

Thresholds are adjusted to make more candidate stations merge with target stations, thus decreasing the overall station count.

References

Legates, D. R., and G. J. McCabe, Jr., 1999: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model evaluation. *Water Resour. Res.*, **35,** 233-241.

Willmott, C. J., 1981: On the validation of models. *Phys. Geogr*., **2,** 184-194.