2.8 DIGITIZATION OF HISTORICAL DAILY COOPERATIVE NETWORK DATA

Nathaniel B. Guttman *
National Climatic Data Center, Asheville, NC

1. INTRODUCTION

Daily data from the U.S. National Weather Service Cooperative Station Network are archived in digital form at the National Climatic Data Center (NCDC) in a format known as TD3200 (France, 1998). The period of record generally begins in 1948, but for a few states and a scattering of stations, the record begins in 1931 or earlier. In addition to the digitized data, the NCDC archives manuscript data that date to the beginning of the observation program at a station.

In 1998, through a congressionally directed economic development partnership, the NCDC began a program to digitize the daily data that was not included in TD3200. The NCDC sent data to a contractor in an economically underprivileged area of Appalachia, and the contractor keyed the data. The resulting digitized data are now included in a format called TD3206, COOP Summary Of Day - CDMP - Pre 1948.

The processing flow, quality assurance activities and information about the use of these data are described in the following sections.

2. KEYING PROCESS

The first step in the process was to decide what data needed keying. Inventories of the TD3200 digitized data were examined to determine the beginning of the period of record for each station. The availability of non-digitized data then needed to be determined. As part of a program to reduce the space needed to store records, in the 1970s the NCDC copied all of its manuscript daily data observation forms onto microfiche. These data on microfiche constituted the source for keying. Inventories of the microfiche were compared to the beginning of record of the TD3200 data, and periods of record to be keyed for each station were determined. The selected microfiche were then sent to the contractor.

The next step was to specify a keying format. A 180 column format was established that allowed for entries of station identifiers; hour of observation; date; latitude and longitude; temperatures at the time of observation and 7 a.m., 2 p.m. and 9 p.m.; daily maximum, minimum and mean temperatures; total precipitation amount, snowfall, and depth of snow on the ground; prevailing wind direction and total wind movement; evaporation; sky condition; and the occurrence of weather and obstructions to vision. Codes were developed for illegible data, monthly sums and means, blank entries for daily but not monthly data on the observation forms, and documentation on the observation forms indicating unreliable data or inconsistencies that cannot be resolved. The keying format was developed so that the digitized data could be easily read in a personal computer spreadsheet.

Coupling data entries with proper station identifiers was considered to be essential. The observation forms for each month of data contain a station name and sometimes a station number. Each microfiche generally contains five years of data and has a station name but no station number on its header. The NCDC also has station history information files that list station names, aliases, numbers, latitudes and longitudes, and periods of record. Unfortunately, in many instances there is disagreement among the sources of identifier information. Procedures were established for the NCDC to determine station numbers when selecting the microfiche and for the contractor to independently determine station numbers. When the contractor could not determine a station number or when the contractor and the NCDC disagreed on the proper number, the NCDC determined the number to be used.

The keying commenced with underlying instruction to “key what you see”. The data entry personnel are not trained in meteorology and therefore are not expected to interpret entries on the observation forms. Using high quality microfiche readers and optimal ambient lighting, the data were keyed twice by different key entry personnel to insure that the data on the form was correctly keyed (although well-trained, the key entry personnel were not expected to key the data without any mistakes). The two resulting data sets were compared, and the contractor resolved discrepancies between the two sets. The array of monthly data on an observation form was keyed column by column rather than row by row.

The keyed data were sent to the NCDC on magnetic tapes. The tapes were checked for read errors, the data records were stripped of extraneous characters at the end of each record, and the files were counted. Each file on a tape consists of data for one microfiche. “Inprocessing” management controls insure that what was returned from the contractor is what was sent to the contractor.

---

* *Corresponding author address*: Nathaniel B. Guttman, National Climatic Data Center, 151 Patton Ave., Asheville, NC 28801-5001; e-mail: ned.guttman@noaa.gov.

3. QUALITY ASSURANCE

The goals of the quality assurance program are to insure that station identifiers are correct and to eliminate egregious and systematic errors from the data. The check of the station identifiers revolves around inventory listings. For a given state, all files having data for the state are extracted from the complete data set (all states). The data for this state are then sorted by file number. The file numbers are consecutive and represent the order in which the data were sent to the contractor. With few exceptions, all the files for a station are numbered consecutively in chronological order. An inventory is produced which shows for every year the number of days by month with something keyed in at least one data field, the year, the station number, the file number, and the number of the tape that contains the file. The inventory also flags year-months with more counts than the number of days in the month, years that are out of chronological sequence, and changes of station numbers within a file.

The inventories are manually checked for proper sequencing of station numbers and chronology. Also, the station identifier on the inventory for a station's first year of data is manually compared to NCDC's station history files and to the name on the observation forms to insure that the identifier is correct. Any discrepancies or patterns that are unexpected are manually reviewed and, if necessary, corrected. This review consists of looking at the observation forms on microfiche, updated versions of station history information that were not available to the contractor, and if necessary, the complete manuscript station history files that are archived at the NCDC. Some of the keying problems encountered are station numbers keyed wrong, incorrect assignment of a station number, dates keyed wrong, duplicate keying, and different meteorological data for the same dates for a station. Some of the source data problems are unofficial names and incorrect dates on the observation forms, filming the wrong forms, observation forms for different stations on a microfiche, and incorrect chronologies on a microfiche.

The sort by sequential file number is a powerful tool for identifying most but not all station identifier problems. Once the problems found from these inventories are corrected, the data are checked with software for duplicates (duplicates are defined as consecutive records having the same identifier, year, month and day, but not necessarily duplicated data fields). Also, all data identified by the contractor as illegible or unreliable are manually reviewed to verify the contractor's assessment and appropriate corrections or changes are made to the digitized files. Another inventory is produced with the same counts but with a primary sort by station number and a secondary sort by year. This arrangement is a different way of looking at the same information, and flags identify duplicate keying that was missed in the review of the first inventory as well as remaining patterns of chronological inconsistencies such as periods of missing data.

The process is iterative, but once the final corrections are made, the data are considered to be properly identified. The next step is to automatically sum all the daily precipitation for a station-year-month and then compare this sum to the monthly total that was keyed. If they match, then all blank precipitation entries are filled with zeroes. Snowfall data are treated in the same manner. The intent of this step is to remove the doubt as to whether a blank entry is zero or missing. The assumption is that if the monthly total equals the sum of the non-zero daily amounts, then all other days must have zero precipitation amounts.

The data are then converted from the keying format into the NCDC's standard TD3200 format for the daily cooperative observing network data and processed through the ValHiDD (Validation of Historical Daily Data) quality control software. This automated program (Reek et al., 1992) was developed by the NCDC to identify, categorize, and eliminate gross digitization and observer errors from the cooperative observer network database. It applies only single-station checks and no spatial checks.

ValHiDD has been used for more than a decade at the NCDC and also was the basis for many of the checks incorporated in the development of the Midwest Climate Center Digitization Project database (Kunkel et al., 1998). Temperature checks include extremes, daily maximum temperatures less than minimum temperatures, spikes and steps in a time series of daily values, continuous runs of the same temperature, and excessive diurnal ranges. Precipitation checks include extremes of precipitation and snowfall, and inconsistencies among total precipitation, snowfall and snowdepth. The check for extremes compares appropriate data values to statewide period of record extremes in a given month of observed lowest and highest maximum and minimum temperatures, total precipitation and snowfall.

The software outputs error codes identifying data that failed a check as well as the offending value. The output is manually reviewed for systematic patterns such as chronological runs of one code or groups of codes and for unusual precipitation/snowfall values that may be indicative of a systematic problem. When a pattern is identified or suspected, the microfiche are reviewed, and appropriate corrections are made when necessary to the digitized database. The types of problems found and corrected include observer or keyed entries in the wrong column of the form, river stage data keyed as precipitation, the occurrence of snowfall without precipitation, the same data entered as two different elements, improper missing codes, digits not keyed, and data keyed with the wrong number of digits after a decimal point. The number of inconsistencies before the review and correction process ranges from several hundred to tens of thousands per state. After the corrections are

applied, the number of inconsistencies for a state lowered by about a third to a half. Although the number of flagged inconsistencies may seem large, it represents only about 0.4 percent before review and 0.2 percent after correction of the total number of days for which data were keyed.

ValHiDD also outputs, when possible, a replacement value for the offending datum. These replacement values are estimated by a predetermined set of rules. Examination of most of these replacement values showed that in many cases they could not be trusted; the reasons for this mistrust are discussed in the next section. Accordingly, replacement values are given only for three conditions: 1) the original data had a misplaced decimal point, 2) the sign of the original data was reversed, and 3) the original data value was wrong by 100 units. The original value that was replaced is also retained in the data set. For all other data that failed a check, no replacement value is given in TD3206, but data quality flags in the data set are set so that the offending elements are noted as having failed an internal consistency check.

4. IMPORTANT QUALITY ISSUES

Since proper station identification is imperative, great care has been placed in trying to assure that a given block of data is associated with the correct station name and number. However, the user should be aware that over the last century, names have changed, stations have moved, administrative procedures have changed, and validation programs have changed. Historical records of a station's identity are often incomplete, ambiguous and inconsistent as well as subject to error. Evaluation of the documentation also required some judgment and educated guesses. A very small amount of data that was keyed is not included in TD3206 because the station identifier could not be confirmed with any reasonable degree of confidence. The confidence in the validity of the station numbers for the data that are included in TD3206 is estimated to be very high for about 97 percent of the stations; the remaining 3 percent reflect a slightly lower confidence for the "best educated guesses."

The prevailing wind direction, sky condition, evaporation, temperature range, average temperature, temperatures at 7 a.m., 2 p.m. and 9 p.m., days with weather and obstructions to vision, total wind movement, and evaporation were not examined in the quality control process. In addition, the hour of observation was set to a missing code for all the data in TD3206. Reasonable verification of the historical times at which observations were made was not feasible from the station documentation and knowledge of observing practices (Karl et al., 1986).

ValHiDD only looks for inconsistencies among maximum and minimum temperature, total precipitation, snowfall and snowdepth. Since temperature range and average temperature were keyed when available, an attempt was made to incorporate these elements into temperature consistency checks. For the early years, the mean daily temperature could not be used because there is no place for an entry of this element on the observation form. Also, the temperature range has different meanings depending on whether the entry was made by the Weather Bureau validator or by the observer. When looking at various combinations of the four temperature elements, an obvious problem arose because of numerous arithmetic errors. Many observers and validators could not properly add, subtract or divide numbers. Rounding also caused consistency problems. The validity of the recorded temperature ranges and daily means are therefore questionable.

Another problem prevented the use of all the keyed temperature elements for determining inconsistencies. In an attempt to relate a 24-hour observation day to a calendar day, some of the keyed maximum temperatures have been date shifted either forward or backward. Sometimes the date shifting was done by the observer, and sometimes by the Weather Bureau validator. Compounding the problem are the inconsistent practices among states, stations, validators and keying operators over time. Because the spatial and temporal extent of the problem is not systematic, all automated attempts to identify and correct inconsistencies failed. A manual effort is needed to check the observation forms on the microfiche as well as other source data for annotations or other indications that the dates have been shifted. This labor-intensive effort is beyond the scope of the quality control process. Without the manual effort, all that can be said about the true date of occurrence of a daily maximum temperature is that it is + or – 1 day of the keyed date.

Extensive examination of existing software both at the NCDC and elsewhere indicated that ValHiDD is the best quality assurance software available. Most of its checks, including precipitation checks, are conditioned on daily temperature values. The date shifting problem in the keyed data causes problems for ValHiDD that result in an incomplete list of inconsistencies and improper determination of replacement values for the inconsistent data it finds. Each check in ValHiDD was therefore examined to determine its utility to the rescued daily data. The Fortran code was studied, modifications to error codes were made for easier identification of specific checks, and the output error codes and estimated "correct" values were compared to the original keyed data. New tables of extreme (high and low) statewide maximum and minimum temperatures were added to flag obvious outliers such as temperatures of several hundred degrees. The results indicated that for the rescued data, ValHiDD is an excellent tool for finding problems, but not a good tool for estimating replacement values.

The prevailing wind direction, sky condition and evaporation data were not assessed for quality for two reasons. First, the definition of the elements is not clear. It could not be determined what "prevailing"

meant to an observer, nor could it be determined how an observer decided to classify a 24-hour period as clear, partly cloudy or cloudy. Evaporation can be measured in several ways, and without an intensive manual review of the station history documentation, exactly what was measured could not, in most cases, be determined. Second, no other keyed elements provide consistent indicators that could be used in an automated evaluation of the quality of the data for these two elements.

Temperatures at 7 a.m., 2 p.m. and 9 p.m., days with weather and obstructions to vision, and total wind movement also were not examined in the quality control process. Although the temperatures at specific times could be related to an expected diurnal cycle or to maximum and minimum temperatures, they generally are not reported after the early 1900s. When they are reported, there is no automated method to determine if indeed they are measurements taken at the specified hours. Total wind movement and days with weather and obstructions to vision are generally independent elements that are not related to the other keyed elements, so internal consistency is not a problem.

5. USER CONCERNS

As discussed in the previous sections, the keyed data are subject to many sources of error. Many of the problems have been corrected, but the user is cautioned to remember that the quality assurance effort for the TD3206 data is limited to only eliminating obvious outliers and to rectifying obvious systematic internal inconsistencies. The quality flags should be considered in conjunction with the numeric data values. These flags identify some, but not all, suspect data. Because of the known date shifting problems, identified internal inconsistencies for temperature values should signal the user to further examine the data sequentially preceding and following the questionable values. Although an entire month or longer period may have been date shifted, the internal consistency check will only flag a maximum temperature that is less than the minimum temperature. Similarly, precipitation and/or snowfall data are flagged as inconsistent if they violate predetermined rules and may indicate the existence of a more systematic problem. The user should also read the Microsoft Word text documentation of all the changes made to keyed data that was received by the NCDC from the contractor. Scanning this file will give the user an appreciation for the types of problems that were noted.

Version 1.0 is known to be missing data for the beginning of 1948 for many stations. These data, as well as a small quantity of additional data for a few stations, will be added to the data set in the next version.

The TD3206 data and documentation can be obtained off-line from the NCDC by contacting customer service representatives (NCDC-DigOrds@noaa.gov; 828-271-4800) or by visiting the NCDC website (http//www.ncdc.noaa.gov). Data are available for the 50 states, Puerto Rico and the U.S. Virgin Islands. The NCDC plans to make the data available on-line from the NCDC website in the near future.

The TD3206 database for the rescued data is separate from the standard TD3200 cooperative network database. Although their formats are the same, the two databases have not been compiled in the same manner. Also, there are some overlaps data between the two that have not been examined. Plans call for merging the two databases into one comprehensive database in the future.

6. REFERENCES

France, L., 1998: Surface land daily cooperative summary of the day, TD-3200. National Climatic Data Center, Asheville, NC, 35 pp. [Available from National Climatic Data Center, Federal Building, 151 Patton Avenue, Asheville, NC 28801-5001].

Karl, T.R., Williams, C.N., Jr., Young, P. and Wendland, W.M., 1986: A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States. *J. Clim. Appl. Meteor.*, **25**, 145-160.

Kunkel, K. E., K. Andsager, G. Conner, W.L. Decker, H.J. Hillaker, Jr., P.N. Knox, F.V. Nurnberger, J.C. Rogers, K. Scheeringa, W.M. Wendland, J. Zandlo and J.R. Angel, 1998: An expanded digital daily database for climatic resources applications in the Midwestern United States. *Bull. Amer. Meteor. Soc.*, **79**, 1357-1366.

Reek, T., S.E. Doty and T.W. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. *Bull. Amer. Meteor. Soc.*, **73**, 753-762.