7.5 DATA QUALITY ASSESSEMENT: WHY WE NEED TO SHARE

Nathaniel B. Guttman
National Climatic Data Center, Asheville, North Carolina

1. INTRODUCTION

Climate data are used to solve practical problems in societal endeavors. To reduce uncertainties in the observation and prediction of climate, we must institute a program of long-term information management. It is not until, and unless, all the observations are combined and analyzed in the context of one another that the complete picture of climate variability can be viewed. Information synthesis ensures the integrity and preservation of the most accurate climate observation record. Also, enhancement of the basic information technology infrastructure allows data sets to be provided to the widest possible array of users in the most cost effective and useful manner.

2. INTEGRATED APPROACH

Data quality assessment is dependent on the kind of information inherent in the observation as well as on the network generating the data. Similar kinds of data (e.g., daily data observed at coop, first order, etc. sites) should be treated together with the same rules and algorithms in an integrated manner. Not only should data be integrated, but assessment techniques and procedures should also be integrated. Algorithms developed by the many entities assessing data should be linked into one unified system so that all basic climate data that are distributed to the public by various agencies are treated in a consistent manner.

The benefits of an integrated approach to data quality assessment include:

a) Reduction in quality assessment development and maintenance costs would be realized because fewer systems would be built. Presently, each of various agencies has its own quality assessment program thereby duplicating common elements of the systems. Even within an agency such as the National Climatic Data Center (NCDC), quality assessment of, for example, daily data is performed independently for each source data set (COOP, first order, etc.).

b) Data quality assessment would become more consistent. Presently, with independent systems, there is rarely a rigorous attempt to insure that algorithms that should be the same in the multiple systems are indeed the same. Also, sometimes algorithms that should be common to all systems are missing in one or more systems. By integrating the assessment into one system, the consistency problems of maintaining multiple systems would be eliminated.

c) Baseline data provided to users would be the same no matter which agency services the customer request. Now, there is no guarantee that values provided by the NCDC and those provided by Regional Climate Centers (RCC) are the same.

d) Integrating data reduces chances for errors and inconsistencies among data sets that span multiple observing networks and platforms. With one system and a distributed network, mirrored copies of data and documentation would insure that all agencies use the same information.

e) Standardized products could be more easily developed for servicing software, data summarization, visualization, climate monitoring, etc. Presently, the myriad source data sets, processing software, application software, and visualization tools cause many interface problems when trying to link parts of the many systems into a new product or application.

f) The collective experience, expertise and wisdom of various data processing entities would lead to a better product than is possible from individual efforts. Since no two people have the same experiences, training, or abilities, a team approach that brings multiple viewpoints to the team mix could lead to a consensus assessment that is far better than the usually more restrictive assessment that results from individual assessments.

3. AN EXAMPLE: INTEGRATING DAILY DATA

An example of the approach described above is the cooperative effort between the NCDC and the RCCs to assess the quality of daily surface data for the United States. Both the NCDC and RCC support customer servicing with historical and real-time daily data. Recognizing that these agencies have common interests, representatives held many discussions both between and among themselves relating to the feasibility of a cooperative approach to data quality assessment. A consensus was reached that all servicing agencies should use the same basic data set and quality assurance methods. Note that this consensus was

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

reached after several years of incubation and experimentation with prototype systems.

We agreed that the best way to accomplish the goal of integration was first to combine all daily data into one file. The NCDC is currently working on this aspect. Historical digital data for the United States that were collected from several networks (National Weather Service first order and cooperative network data streams, SNOTEL, Navy and Air Force) and data sets that were compiled retrospectively (Midwest RCC, Climate Database Modernization Program "pre-1948" data) have been merged into one data set that retains information about the data source and carries into the merged set the original elements, data values, and flags associated with the values.

The data assurance efforts are being developed separately to exploit the expertise of the individual partners. Individual quality assurance modules will be tested and, if accepted by the partnership, made part of a suite of algorithms through which the data will be passed.

Examples of separate modules are format and consistency checks being developed by the NCDC, a regression-based spatial assessment being developed by the High Plains RCC, and an anomaly based spatial assessment being developed by the Midwest RCC.

The format checks being developed at the NCDC include checks to insure that all data and metadata that is in a given data set conform to the documentation that exists for the set. This documentation (e.g., NCDC reference manuals, Federal Meteorological Handbooks, and Circular N) describes data element ranges, data formats, data value flag definitions, units of measurement, and other similar information. The software checks the integrated data to insure that the information in the data set conforms to the source-specific rules; no meteorological data assessments are made. Check failures are indicated by appropriate flags being set in the integrated data set and by output files generated by the checking software.

Consistency checks are mathematical and meteorological in nature. They consist of insuring that maxima are greater than or equal to minima, that data values fall within physical limits, and that relationships among elements are valid. At the time of this writing, the software developed by the NCDC performs about 40 different consistency checks on the integrated data. Data values are also checked to see if they fall between extremes. Thresholds for the extremes are set by physical limits, by the .005 probability of a value based on a Wakeby probability model calculated from L-moments, by empirical curve fitting of the observed data if the probability model is not appropriate, and by, as a last resort, state-wide extremes. Not only are the observed values checked, but replacement (estimated) values are also checked against other observed, other replacement values and extreme thresholds.

The assessment system being developed by the NCDC is linear. In the first level, data are first passed through the format checks. If any information fails a check, a flag is set. It is intended that the flagged information will be examined and corrected, and then once again passed through the format checks, before the data set is passed through the consistency and extreme checks (second level). Any information that fails a format check is ignored in the consistency and extremes checks. Similarly, the flags set by the consistency and extremes checks should be evaluated and corrections made, and then the data should be passed through the checks once again. Third level, which has not yet been developed, will be spatial checks. The last level, which also has not yet been developed, will be to assess multiple values of an element for a given station and time that originate from different sources and to (hopefully) produce one value for this element, station, and time.

The system is also modular. All algorithms are being written as subroutines so that individual components can be called as needed. The modular approach allows other routines to be included in the future. These routines could include those developed by the NCDC, RCCs, or any other agency.

The separate modules being developed by the RCCs will be described in other presentations at this Conference.

Linking the data into one system will be accomplished through mirror sites on a distributed network. The quality assessment software and documentation will also be linked through this distributed system so that all partners will have access to the same software suite.

4. THE FUTURE

The integration process is evolving and can continue indefinitely. To date, the climatological service community has recognized the need for integration, has agreed on a conceptual plan to achieve integration, and has begun work on building a flexible system. A lot of work still needs to be done to a) develop, test and implement new assessment algorithms, b) complete the distributed network, c) integrate real-time and historical data so that *any* data can be provided to users within servicing time constraints, and d) integrating data on all time scales into one system.