6.2 THE FCC INTEGRATED SURFACE HOURLY DATABASE, A NEW RESOURCE OF GLOBAL CLIMATE DATA

J. Neal Lott and Rich Baldwin
National Climatic Data Center
Asheville, North Carolina

1. Introduction

As a result of Environmental Services Data and Information Management (ESDIM) funding, Office of Global Programs (OGP) funding, and extensive contributions from member agencies in the Federal Climate Complex (FCC), the National Climatic Data Center (NCDC) has completed two phases of the Integrated Surface Hourly (ISH) database project:

1) The "database build" phase, producing ISH version 1 – The new database collects all of the NCDC and Navy surface hourly data (TD3280), NCDC hourly precipitation data (TD3240), and Air Force Datsav3 surface hourly data (TD9956), into one global database. The database totals approximately 350 gigabytes, for nearly 20,000 stations, with data from as early as 1900 to present. The building of the database involved extensive research, data format conversions, time-of-observation conversions, and development of extensive metadata to drive the processing and merging. This included the complex handling of input data stored in three different station-numbering/ID systems.

2) The first phase of quality control (QC), producing ISH version 2 -- This includes a) correction of errors identified after the "database build" phase (e.g., due to input data file problems); b) research, development, and programming of algorithms to correct random and systematic errors in the data, to improve the overall quality of the database; and c) data processing of the full period of record archive through the QC software. The database has been archived on NCDC's Hierarchical Data Storage System (HDSS, tape-robotic system), and is now on-line for WWW access.

2. Background

2.1 Participating Agencies

The Federal Climate Complex (FCC) in Asheville, NC is comprised of the Department of Commerce's National Climatic Data Center (NCDC), and two components of the Department of Defense (DOD)--the Air Force Combat Climatology Center (AFCCC) and the US Navy's Fleet Numerical Meteorological and Oceanographic Command Detachment (FNMOC Det). The FCC provides the nation's climatological support.

* Corresponding author address: J. Neal Lott, National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801; email neal.lott@noaa.gov.

The purpose of the FCC is to provide a single location for the long term stewardship of the nation's climatological data, and to provide the opportunity for customers to request any climatological data product from a single location.

2.2 The Problem

Surface-based meteorological data are by far the most-used, most-requested type of climatological data. However, a single integrated database of global hourly meteorological observations did not previously exist. Researchers requiring surface climatic data often acquired the data from several sources in a variety of formats, greatly complicating the design of their applications and increasing the cost of using the data. For example, when someone needed all available surface hourly data for a selected region (U.S. or worldwide) and time, they would receive data from three datasets which differed in format, units of storage, and levels of quality control. Alternately, the user would simply choose which one of the datasets might be able to meet their requirements, which often resulted in incomplete or inaccurate results. Many users complained about the problems this created in data usage, and in getting complete, accurate results.

Additionally, the currently available datasets of monthly resolution (e.g., monthly precipitation totals) being used for climate change research do not provide the hourly and daily resolution required to describe regional climatologies and investigate extreme events, such as a 1-day rainfall event causing flash flooding. These monthly datasets also do not usually provide other needed parameters such as cloud data, visibility, wind speed, snow depth, etc. Therefore, this project was required to produce a single, integrated, quality-controlled, complete global climatic database of hourly data.

3. Development of ISH

3.1 Input Data Sources

The datasets used as input for the ISH database were:

1) NCDC and Navy surface hourly data, referred to as TD3280 (Steurer, Bodosky, 2000). (TD is an old reference to "tape deck.")
- Data Origin and QC: Data originate mainly from the Automated Surface Observing System (ASOS), and previously from diskettes from the stations and key-entered data. Data undergo extensive automated and manual quality control.
- Content/Elements: Comprises about 1400 National Weather Service (NWS) and U.S. Navy stations historically, generally for 1948 to present. About 380

stations are currently active. It includes most surface elements observed in the U.S. (wind speed and direction, temperature, dew point, cloud data, sea level pressure, altimeter setting, station pressure, present weather, and visibility). Wind gust, precipitation amount, and snow depth are not included. "Specials" are not included and only synoptic hours (every 3rd hour) are included for 1965-1981 (for most stations).

2) AFCCC Datsav3 surface hourly data, referred to as TD9956 (AFCCC, 1998).
- Data Origin and QC: Data originate mainly from the Global Telecommunication System, various other sources, and keyed data prior to 1973. The keyed data include over 100 "tapedecks" which were laboriously converted to Datsav3 format by AFCCC, thus providing a great deal of data prior to 1973 for a number of international stations. Data undergo extensive automated quality control, and manual quality control for USAF stations.
- Content/Elements: Comprises about 20,000 global stations historically, with many stations for 1973 to present, and some as far back as 1900. About 10,000 stations are currently active. It includes all elements mentioned above for TD3280, along with wind gust, precipitation amount, snow depth, and other elements as reported by each station. It also includes "Specials."

3) NCDC hourly precipitation data, referred to as TD3240 (Steurer, Hammer, 2000).
- Data Origin and QC: Data originate from various sources including ASOS and weighing rain gage data from stations. The data undergo automated and manual quality control.
- Content/Elements: Contains hour-by-hour precipitation amounts. It includes NWS stations and various cooperative observing sites. About 2800 stations are currently active, with data generally for 1948 to present. ISH includes TD3240 stations which also have data available in either TD9956 or TD3280.

These three datasets were the most logical starting point for ISH, as they were the most-used hourly datasets available, and have also been subjected to considerable quality control and have adequate station history information available.

3.2 The Process

All necessary metadata within the FCC were collected, coordinated, and loaded into a set of relational database tables. NCDC uses Oracle for this requirement. The metadata includes important information about the data such as station histories, dataset documentation, and inventories; along with critical information to control the process of merging the data. Since the three input data sources are archived in dissimilar station numbering/identification systems, the metadata must provide a cross-reference to identify data for the same location (i.e., same station with data in each of the three input datasets). This station history then controls the overall process flow and data merging, and also must account for station number changes over time.

A time conversion control file was used to convert the TD3280 and TD3240 data to Greenwich Mean Time (GMT), so that all input data were then in GMT time convention. The ISH data are therefore in the same time convention as upper air data and many other global databases, model output, satellite data, etc. This is quite important for potential GIS applications. The creation of this time conversion control file was very cumbersome, involving research of several sources of information concerning time zones, time zone changes historically, etc. We had to account for time zone changes to properly merge the data.

We included a quality control check for TD9956 vs. TD3280 to ensure that the data were actually for the same location at the same time before performing the intra-observational merge, which creates a composite observation for that date-time. The quality control check was conducted on a daily basis (i.e., on each day's data) to determine if the data for that day should actually be merged into composite observations.

Procedures, algorithms, and then computer programs were written to merge the surface hourly datasets into one common database. More than one billion surface weather observations (covering 1900 to present) for approximately 20,000 global locations were accessed and merged during this process. Examples of input data types were: Automated Surface Observing System (ASOS), Automated Weather Observing System (AWOS), Synoptic, Airways, Metar, Coastal Marine (C-MAN), Buoy, and various others. Both military (e.g., USAF, Navy, etc.) and civilian stations, automated and manual observations, are included in the database.

The final version 1 database (see Lott, 2000) includes all observed elements at hourly intervals (every three hours for some stations), such as temperature, dew point, visibility, wind speed and direction, wind gust, cloud data, precipitation, snow depth, and many others.

To develop the version 2 database, we researched, developed, programmed, and processed the data through 54 quality control (QC) algorithms. This phase of quality control subjects each observation to a series of validity checks, extreme value checks, internal (within observation) consistency checks, and external (versus another observation for the same station) continuity checks. Therefore, it may be referred to as an inter/intra-observational quality control, and is entirely automated during the processing stage. However, it does not include any spatial quality control ("buddy" checks with nearby stations), which is planned for later development.

Since the input datasets had already been subjected to a good deal of quality control, this additional quality control was designed to address problems which were less likely to have already been corrected. Detailed documentation on each of these QC algorithms is available (Lott, 2001). The final database and subsequent products are available to all three organizations to support the FCC mission and its numerous customers.

3.3 Goals and Milestones

As is the case with most software systems, ISH is designed to evolve, within the limitations of current funding and technology, to reach its goals. Here are the milestones for the overall effort:

- Develop and test the initial population system, including all required metadata and process control files. (Completed)
- Use this system to create the ISH database from the beginning of record to the present. (Completed)
- Store a portion of the database (as much of the most recent data as possible) in relational tables and develop software for online access via the NOAA National Data Center Climate Data Online (NNDC CDO) system (Lott, Anders, 2000). (Completed)
- Modify this system to operate in a production environment on a routine (weekly to monthly) basis. (Completed)
- Develop, program, and process the database through a set of quality control algorithms, to improve the overall quality of ISH. (Completed)
- As resources become available, make the entire database accessible on-line via NNDC CDO.
- Add additional datasets to ISH, via the merging process.
- Add additional station history/metadata to the database, to include as much instrumentation information as possible; thereby making the data more useful for climate change research.
- Research, develop, and apply more sophisticated time series and spatial quality control checks to ISH; thereby making the data more robust and useful for all applications.

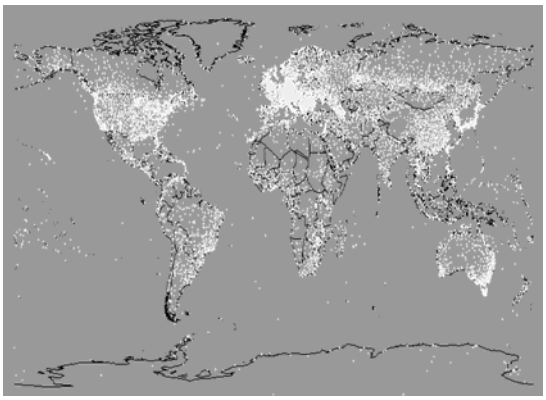
4. Results and Conclusion

The new, integrated database is a global baseline database and the first of its kind, and can be used by both U.S. and global customers. All surface hourly climatic elements are now stored in one format for the full period of record. The database will be operationally updated on a weekly basis for 2001 and beyond.

Researchers will be serviced more quickly and in a much more efficient manner at a lower cost to the FCC agencies. This in turn will reduce the overall operating costs to the government and the access and development costs for the users of the data.

ISH provides a common data format for development of FCC applications software, for NNDC applications, and for all researchers and users of the data to develop application programs easily and efficiently. Also, the database contains daily parameters (e.g., maximum/minimum temperature, daily precipitation) which can be used to generate daily summaries.

Finally, this global map depicts the general density of station locations in the ISH database, in this case, for 1995:



5. Acknowledgments

There were many people who contributed to this project's success. The key members of the team were: Rich Baldwin, NCDC; Vickie Wright, NCDC; Dee Dee Anders, NCDC; Danny Brinegar, NCDC; Neal Lott, NCDC; Pete Jones, Marada Corporation; and Fred Smith, Marada Corporation. Of particular note is Bob Boreman (NCDC) who devoted a great deal of time and effort to ISH, especially in development of the time conversion control file. Bob passed away in July 2001 and is greatly missed.

6. References

AFCCC. Documentation for Datsav3 Surface Hourly Data. [Asheville, N.C.]: Air Force Combat Climatology Center, 1998.

Lott, Neal. Quality Control of USAF Datsav3 Surface Hourly Data–Versions 7 and 8. [Asheville, N.C.]: USAF Environmental Technical Applications Center, 1991.

Lott, Neal. Data Documentation for Federal Climate Complex Integrated Surface Hourly Data. [Asheville, N.C.]: NCDC, 2000.

Lott, Neal. Quality Control Documentation for Federal Climate Complex Integrated Surface Hourly Data. [Asheville, N.C.]: NCDC, 2001.

Lott, Neal, Rich Baldwin, and Pete Jones. NCDC Technical Report 2001-01, The FCC Integrated Surface Hourly Database [Asheville, N.C.]: NCDC, 2001.

Lott, Neal and Dee Dee Anders. NNDC Climate Data Online for Use in Research, Applications, and Education. Twelfth Conference on Applied Climatology. Pages 36-39. American Meteorological Society, May 8-11, 2000, Asheville, NC.

Steurer, Pete and Matt Bodosky. Data Documentation for Surface Hourly Airways Data, TD3280. [Asheville, N.C.]: NCDC, 2000.

Steurer, Pete and Greg Hammer. Data Documentation for Hourly Precipitation Data, TD3240. [Asheville, N.C.]: NCDC, 2000.