7.8 THE QUALITY CONTROL OF THE INTEGRATED SURFACE HOURLY DATABASE

J. Neal Lott *
National Climatic Data Center, Asheville, North Carolina

ABSTRACT

The National Climatic Data Center (NCDC), in conjunction with the Federal Climate Complex (FCC), developed the global Integrated Surface Hourly (ISH) database to address a pressing need for an integrated global database of hourly land surface climatological data. The database of approximately 20,000 stations has data from as early as 1900 (many stations beginning in 1948-1973 timeframe), is operationally updated with the latest data, and is now being used by numerous customers in many varied applications. ISH is being quality-controlled in several phases, with two phases now completed. This paper addresses: a) the challenges and lessons learned in ISH development, b) the quality control (QC) applied during the initial development, c) the more extended QC applied after the initial development, d) the current shortcomings and needs for the database, and e) the future plans for QC and for partnerships.

1. INTRODUCTION

The FCC is comprised of the Department of Commerce's NCDC, and two components of the Department of Defense -- the Air Force Combat Climatology Center (AFCCC) and the US Navy's Fleet Numerical Meteorological and Oceanographic Command Detachment (FNMOC Det). The FCC provides much of the Nation's climatological support. 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.

As a result of Environmental Services Data and Information Management funding, Office of Global Programs funding, and extensive contributions from member agencies in the FCC, the NCDC has completed two phases of the ISH database project:

a) The "database build" phase, producing ISH Version 1 – The new database collects all of the NCDC and Navy surface hourly data (DSI 3280), NCDC hourly precipitation data (DSI 3240), and Air Force Datsav3 surface hourly data (DSI 9956), 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. See Figure 1 for a high-level flow diagram of the process.

b) The first two phases of QC, resulting in ISH Version 2 – Phase one involved the quality assurance of the Version 1 database build, to detect and correct any errors identified during this phase (e.g., due to input data file problems). Phase two involved the research, development, and programming of algorithms to correct random and systematic errors in the data, to improve the overall quality of the database; and the data processing of the full period of record archive through these QC algorithms.

The database has been archived on NCDC's Hierarchical Data Storage System (HDSS, tape-robotic system). All surface hourly climatic elements are now stored in one consistent format for the full period of record. The database is operationally updated with the latest data on a routine basis.

Surface hourly is one of the most-used types of climatic data for NOAA customer-servicing and research, involving requests for the hourly data and for applications/products produced from the data. ISH is greatly simplifying servicing and use of the data, in that users do not have to acquire portions of three datasets with differing formats, and do not have to deal with and program for the inconsistencies and overlaps

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

between the three input datasets. Also, this has resulted in an end-to-end process for routine database updates, the database is being placed on-line for Web access, and the more recent data have been collected into a CD-ROM product with a map interface.

2. PROJECT CHALLENGES AND ISSUES

Though surface-based meteorological data are the most-used, most-requested type of climatological data, a single integrated database of global surface hourly meteorological observations over land 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 often would receive data from three datasets which differed in format, units of storage, and levels of QC. 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. Therefore, this project was undertaken to produce a single, integrated, quality-controlled, complete global climatic database of hourly data, for use in numerous applications, by private and public researchers, corporations, universities, and government agencies. However, the integration of disparate data sources presented many challenges. The three input datasets were the most logical starting point for ISH, as they were the most-used hourly datasets available, having also been subjected to considerable QC, and having adequate station history information available.

All necessary metadata within the FCC were collected, coordinated, and loaded into a set of relational database tables. The metadata include important information about the data: station histories; dataset documentation; inventories; and other 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 had to 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.

The station number changes over time were an added challenge, as some stations had three to as many as six different station numbers to identify the same location. It was critical to merge these data into a single "entity" over time, so that users would have a consistent set of data. Also, in looking ahead (at that time) to having the data on-line in a web-based interface, it was important to consider what would be presented to the user through the interface (i.e., a poplist of stations by country, state, etc).

A time conversion control file was used to convert the NCDC and Navy hourly data (DSI 3280 and 3240) to Greenwich Mean Time (GMT), so that all input data were then in GMT time convention. (The Air Force surface hourly data (DSI 9956) were already in GMT.) The ISH data are therefore in the same time convention as global upper air 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. This required fully accounting for time zone changes to properly merge the data.

Finally, toward the end of the development phase, the original workstation for development and testing was replaced with a newer 64-bit workstation. Although the change should have been transparent to the ISH system, it was not. Many problems began to appear with code that had been working before the change. After much research, no cause could be identified, although evidence seemed to point to a system memory utilization problem. Several work-arounds were put into place in order for all of the components to function again.

3. QUALITY CONTROL APPLIED INITIALLY

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.

As part of the "Version 1" building of the database, we included QC checks to ensure that the input 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 QC 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. Temperature, dew point, and wind direction were compared for each data value (e.g., temperature at a given station-date-time in DSI 9956 vs. temperature at same station-date-time in DSI 3280) to obtain a percent score for the day for coincident data. Criteria of 1-degree Celsius for temperature and dew point, and 10 degrees for wind direction were used as the pass/fail limits for each element, with an overall 70% score for the day required to perform the intra-observational merge for that day. In other words, 70% of the data values compared would have to meet the limit checks to "pass" for the day. Failure of these checks sometimes pointed to time conversion problems, updates to the control file, and re-processing of those stations. Subjective analysis of the data before and after processing proved this QC check and the limits applied to be quite effective.

A complete inventory system was included to fully verify that no data were "lost" during the processing. This involved an inventory (i.e., number of observations by station-month) for each of the input datasets, with the inventories stored in Oracle relational tables. Then, the final "output" ISH database produced a similar inventory stored in a database table. The inventory tables were then compared to check for any loss of data. This proved to be a critical component of the process, and revealed a number of problems that would have otherwise gone undetected. The database processing was not considered complete until the inventory verification process was complete. Also, the final inventory thereby provided a very useful inventory of the ISH archive for use in placing the data online and servicing customers. See Table 1 for an inventory by WMO block of the number of stations in the ISH database.

Another critical component of the Version 1 database build was the development, processing, and verification of "test data" to attempt to check as many possible paths through the process that the data might follow. The painstaking process of creating, processing, and checking the test data, though very time-consuming, was critical to the success of the project. This, in conjunction with the checking of actual data from the archive, proved to be very valuable, and more than worth the time invested in this component of QC. An added benefit of test data is the re-validation of the system periodically, such as when a source code change or operating system upgrade is required. Then, an automated comparison (e.g., Unix diff) of baseline output test data vs. the new output test data quickly reveal if any problems are present.

Needless to say, it was very important to randomly check the results of the final process—i.e., selected output files, for any unforeseen problems. This component of the process revealed very few problems, due to the intensive nature of the QC described above.

4. PHASE 2 OF QUALITY CONTROL

To develop the Version 2 database, we researched, developed, programmed, and processed the data through 57 QC algorithms. This phase of QC subjects each observation to a series of validity checks, extreme value checks, internal (within observation) consistency checks, and temporal (versus another observation for the same station) continuity checks. Therefore, it may be referred to as an inter/intra-observational QC, and is entirely automated during the processing stage. However, it does not include any spatial QC ("buddy" checks with nearby stations), which is planned for later development.

An example of one of the algorithms performed is the continuity check for temperature, which does a "two-sided" continuity validation on each temperature value for periods ranging from 1 hour to 24 hours. An increase in the temperature of 8 degrees Celsius in one hour (e.g., from 10° C to 18° C) prompts a check on the next available (i.e., next reported) temperature–if that value then decreases by at least 8 degrees in an hour (e.g., 18° C to 9° C), then that indicates a very improbable "spike" in the data, and the erroneous value (e.g., 18° C) is changed to indicate "missing" for that observation. However, the original value is saved in a separate section of the data record for future use if needed. The same would apply

for a downward "spike" in the data, and similar checks are performed for periods out to 24 hours, to allow for missing data and for part-time stations which do not report hourly or three-hourly data. The validation always checks the closest values available temporally (i.e., before and after the data point being checked), and the limit is automatically adjusted based on the elapsed time between values. Temporal continuity checks are performed for continuous elements such as temperature, dew point, wind speed, and pressure (station, sea-level, and altimeter setting).

Another example of the algorithms is the consistency check for present weather vs temperature, to ensure that, for example, frozen precipitation is not reported at unrealistic temperatures. There are a number of these QC checks for various types of present weather reports. Similar checks are performed for various other elements such as cloud data, precipitation amount, snow depth, etc.

Though all climatic elements are checked to some extent, the elements validated to the greatest extent are: wind data, temperature and dew point data, pressure data, cloud data, visibility and present weather data, precipitation amounts, snowfall and snow depth. In addition, a selected number of systematic deficiencies are addressed with specific algorithms to correct those problems. As mentioned above, the input datasets had already been subjected to a great deal of QC, so this phase of QC was designed to address problems which were less likely to have already been corrected.

The creation and verification of test data for each algorithm was just as critical in this phase as in the creation of ISH Version 1. As mentioned above, an automated comparison of baseline output test data vs. the new output test data quickly reveal if any problems are present in the system.

We do not consider this to be an "end-all" QC process, but merely the next step in producing a better quality database for NOAA customers. Detailed documentation on each of these QC algorithms is available (Lott, 2003).

5. FUTURE PLANS

One of the goals for ISH is to have the entire dataset available for query via the NNDC Climate Data Online (CDO) system (Lott and Anders, 2000). With the difficult and tedious task of blending the data from the three sources completed, end-users may then extract what is needed with relative ease and can focus on their research or studies, rather than getting bogged down in the merging process. A second goal is to continue to add to and improve this global baseline database for research and applications requiring data of this type, by adding additional datasets (i.e., merge into ISH), and developing/applying more extensive QC checks.

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

a) Store the entire database in relational tables, with access provided via the NNDC CDO system. This process is well underway.

b) Continue the routine updating of the database using the established procedures and software, and revise the NCDC processing software for U.S. surface data to perform a near real-time (daily) ingest and QC of all surface data into ISH format, thereby providing users with near real-time access to quality-controlled data. This process is also well underway, and is being referred to as the Integrated Surface Data Processing System.

c) As funding permits, add additional datasets to ISH, via the merging process. This is now planned for selected U.S. mesonet data, and several datasets of non-U.S. data.

d) As funding permits, add additional station history/metadata to the "system," to include as much instrumentation information as possible; thereby making the data more useful for climate change research.

e) As funding permits, research, develop, and apply more sophisticated time series and spatial QC checks to ISH; thereby making the data more robust and useful for all applications.

f) Develop partnerships with other government agencies and groups, such as the Regional Climate Centers. This includes partnerships for additional data sources, enhanced QC techniques, and online applications.

6. SUMMARY AND CONCLUSIONS

The development and QC of ISH has been a rather long and arduous process, but well worth

the effort. The QC has included the following phases/components (as described in more detail above):

Phase 1:

a) Validation of the merging process through element value comparisons, such as temperature.

b) A complete inventory of all input and output data, to ensure no data loss during the processing.

c) Thorough checking of test data and archive data, to fully test the software before full database processing began.

d) Random checks of the final output database (ISH).

Phase 2:

a) Extremes / validation checks, to ensure no obviously erroneous values are present in the data.

b) Temporal continuity checks, to look for "spikes" in continuous elements such as temperature, dew point, wind speed, and pressure (station, sea-level, and altimeter setting).

c) Consistency checks of one element vs another within a given data record/observation, such as temperature vs present weather (e.g., no snow at 10° C)

The lessons learned include:

a) Thorough test data are critical to any process such as this. Though proven to be true in many of the author's previous projects, it certainly proved to be critical in this one.

b) Peer review is very important to ensure that the overall process and the individual QC checks are not merely the ideas of one individual, but are consistent with good science and good data processing standards.

c) The concept of "phases" in a project of this magnitude is critical to success. There is a tendency to "bite off more than we can chew" with any project, and the phased-in approach was one of the keys to success with ISH.

d) Finally – expect to reprocess. No matter how many checks and balances are in place, further improvements and some reprocessing will be necessary. The key is in limiting its frequency, while at the same time having a willingness to do it when needed.

7. 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, TMC Corporation; and Fred Smith, TMC Corporation. As mentioned previously, the Air Force (AFCCC) and Navy (FNMOC Det) contributed to the effort. Bob Boreman (NCDC), who devoted a great deal of time and effort to ISH, especially in the research and development of the time conversion control file, passed away in July 2001. He is greatly missed, and his contributions to this effort are gratefully acknowledged.

8. 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.]: National Climatic Data Center, 2000.

Lott, Neal. Quality Control Documentation for Federal Climate Complex Integrated Surface Hourly Data. [Asheville, N.C.]: National Climatic Data Center, 2003.

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.

Lott, Neal, Rich Baldwin, and Pete Jones. NCDC Technical Report 2001-01, The FCC Integrated Surface Hourly Database, A New Resource of

Global Climate Data. [Asheville, N.C.]: National Climatic Data Center, 2001.

Plantico, Marc and J. Neal Lott. Foreign Weather Data Servicing at NCDC. ASHRAE Transactions 1995, V. 101, Pt 1.

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

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

Table 1 - Database Inventory

The following table provides an inventory by WMO block of the number of stations in the ISH database having at least one year of data.

WMO Bk	Stations	WMO Bk	Stations	WMO Bk	Stations	WMO Bk	Stations
01	297	30	230	58	189	88	36
02	413	31	227	59	149	89	124
03	593	32	102	60	193	91	234
04	105	33	190	61	128	92	12
06	275	34	170	62	125	93	82
07	297	35	143	63	143	94	622
08	165	36	130	64	155	95	163
09	74	37	208	65	123	96	92
10	491	38	202	66	30	97	67
11	226	40	356	67	186	98	85
12	219	41	202	68	270	99	334
13	153	42	225	69	548		
14	71	43	151	70	223		
15	322	44	100	71	1085		
16	280	45	13	72	2113		
17	122	46	60	74	700		
20	59	47	521	76	261		
21	58	48	350	78	244		
22	180	50	52	80	130		
23	177	51	68	81	27		
24	116	52	107	82	144		
25	99	53	118	83	290		
26	220	54	178	84	101		
27	147	55	22	85	163		
28	237	56	163	86	57		
29	178	57	243	87	183		

Figure 1 - ISH Database Build Process Flow