A World Data Center Perspective on Data Management

Jay Lawrimore, Matt Menne, Claude Williams
NOAA National Climatic Data Center
The focus of this presentation is on 10 data management issues associated with global data sets developed at NCDC

- Past Successes and Future Opportunities

Three timescales
- MONTHLY: Global Historical Climatology Network–Monthly
- DAILY: Global Historical Climatology Network–Daily
- SUB–DAILY: Integrated Surface Data (ISD)
Ten Issues

1. Separate and Independent Dataset Development
2. Building and Maintaining Historical Data Holdings
3. Maintaining Datasets Through Operational Ingest
4. Other Near Real-Time Data Acquisition Issues
5. Data Rescue and Digitization
6. Building a Dataset from Multiple Sources
   - Formatting, Station Matching, Duplicates
7. Quality Control
8. Metadata
9. Bias Adjustments
10. Data and Metadata Access
NCDC monitors mean changes in global land surface temperature using the Global Historical Climatology Network (GHCN) dataset
- Composed of 7,280 stations that record monthly mean temperatures

In use since the early 1990s. Version 3 Beta release in September 2010
Global daily *in situ* dataset derived from multiple sources
- ~25,000 temperature stations
- ~60,000 precipitation stations
- ~25,000 snowfall or snowdepth stations

Approaching 2 billion daily observations
- Earliest value from January 2, 1833
- Latest value from yesterday
Integrated Surface Data (Sub-Daily)

- ISD—formerly called ISH (Integrated Surface Hourly)—an integration of dozens of datasets
  - include the Automated Surface Observing System (ASOS), Synoptic, Airways, METAR, Coastal Marine (CMAN), Buoy, and various others, from both military and civilian stations and including both automated and manual observations
- Released to the public in 2003
- Relational Database table structure (data and metadata)

- Archived dataset, with over 20,000 worldwide stations for 1901 – present (< 100 sites in early years, > 10,000 stations now active, pre–1900 data being added)
The Monthly, Daily, and Sub-daily data sets were developed independently of each other. When GHCN-M was developed there were only small amounts of digitized daily data from which to construct monthly means. There is little connection between the monthly, daily, and sub-daily data sets. Today there is greater access to digitized daily (and sub-daily) data but coverage before 1950 remains relatively sparse.
Issue 1: Separate and Independent Data Set Development

1901–1930

GHCN–Monthly

1931–1960

GHCN–Monthly

GHCN–Daily
The greater availability of digitized daily observations and potential for even more provides better opportunities to more closely link monthly and daily datasets.

But improvements in transmission and sharing of daily climate observations still needed.

Determining which of the GHCN-D stations are already in GHCN-M and which can be summarized and added or augmented to the Monthly is a time-consuming effort.

- Station names often don’t match.
- Location (latitude/longitude) even more often don’t match.
- Summarized data are similar but not exact.

Hourly and Synoptic Observations are not best sources for development of daily and monthly data sets.
Datasets have been developed using a variety of sources acquired through various means:

1. Related Projects (e.g., WMO normals, World Weather Records, Colonial Era Archives)
   - World Weather Records (Decadal data collections)
     - Excellent source of data during this decade AND data from the previous decades
     - Greatest success in collecting data from countries was in association with a WMO Request
     - Time consuming processing of converting multiple and diverse formats. Some data required keying from e-mails
2. Direct Agreements with other Meteorological and Data Centers that collect, archive, manage and distribute climate data
   - e.g., Australian Bureau of Meteorology recently provided daily data for 17,000 stations
   - Data Sharing Restrictions within some countries limit data access

3. Personal Contacts

4. International Workshops
   - No one single source of data is sufficient for developing and maintaining a global data set.
   - Requires continual efforts to maintain a data set that is healthy and can meet the needs of a large and diverse user community
Data sharing agreements can provide major improvements in data coverage.

Australian Bureau of Meteorology recently provided daily temperature records for more than 1500 additional stations:

- A more than 10-fold increase in previous GHCN-Daily data holdings from that country.
Issue 3: Maintaining Datasets Through Operational Ingest

- Ongoing updates to Datasets are accomplished largely through operational acquisition and ingest processes
  1. GTS transmissions
     - Monthly Means (via CLIMAT bulletins)
     - Synoptic (observations every 3 or 6 hours, can include daily max and min temperature)
     - METAR (hourly observations, typically from airports)
  2. Other (NOAAPort)
  3. Web Services (Country-to-Country)
     - Currently receiving updates to more than 1000 stations from Canada daily via web services
Issue 3: Maintaining Datasets Through Operational Ingest

4. FTP
   - Currently receiving updates to data from Australia on the 4th of each month

5. E-mail
   - GSN data from some countries received via e-mail
     - Requires a great amount of human interaction to review and verify data received, convert from various formats and place into common digital format
Issue 3: Maintaining Datasets Through Operational Ingest

- Operational Ingest and Acquisition alone provides only a portion of all data
- Without continued attention to acquire and incorporate data from other sources, the number of stations in the datasets suffer decline
  - Many stations do not report in near real-time
  - Addition of other stations requires years of effort to locate, collect, process, and incorporate
- Other reasons for decline in stations
  - Stations stop operating (e.g., 20% decline in Russia since 1990)

All GHCN data are available at www.ncdc.noaa.gov
Issue 4: Other near–real Time Data Acquisition Issues

- A data transmission pathway for daily climate observations does not exist
  - No analog to CLIMAT bulletins on a daily timescale
  - Forced to rely on transmission of daily summaries in Synoptic Bulletins
    - Not Required of Member Nations
    - No Regular transmission
    - Not midnight–to–midnight observations and often not clear what 24–hour period the observations cover
Issue 4: Other near–real Time Data Acquisition Issues

- Central Repository for daily and/or sub–daily data needed to serve community
  - Adopt international mechanisms for standardizing exchange of the highest resolution data for climate purposes
  - Provide formal mechanism for disseminating daily climate observations

- Bilateral arrangements have proven effective at improving data sharing (e.g., US–Canada and U.S.–Australia)
  - Consider other opportunities for expansion of bilateral arrangements for sharing daily and sub–daily data
NCDC leads the Climate Database Modernization Program

- Major effort supported by contractors in 3 states to photograph and digitize paper and microfiche

- Positives
  - 56 million images made
  - US Forts data (1800s) digitized
  - Data from Uruguay and Mexico
  - Upper air data from several countries in Africa.

- Negatives
  - Much less data digitized than imaged
  - Resource intensive
  - Very costly
  - Resources required to incorporate data into data sets
Issue 5: Data Rescue and Digitization: Other Solutions?

- **Crowdsourcing**
  - A distributed labor network
  - Like Wikipedia
  - Exploits spare power of millions of others. Distributes burden
  - Being used in numerous ways
    - Open source software
    - Media inquires: sifting through thousands of documents
  - Requires clear and standardized guidelines
    - Strict quality control procedures by experts
      - Multiple cultures and languages

- High cost and resource intensive programs like NCDC’s CDMP program restricts wider uses.
Issue 6: Building a Dataset from Multiple Sources: Formatting and Duplicates

- Formatting
  - Labor intensive efforts to convert from one of many formats
  - Manual effort involved to decipher, align, remove extraneous lines, columns, and data
  - Ensure inventory (metadata) contains stations in data file

- Duplicates
  - There are reportedly over 100 different ways to calculate monthly mean temperature
  - Multiple sources are often similar but not exactly the same
  - Often not possible to determine one “correct” time series
Issue 6: Building a Dataset from Multiple Sources: Station Matching

- Metadata inaccuracies complicate merging of station records
  - Even if large percentages of data match, large distances between stations make station matching difficult or impossible

- Strong reliance on station ID matching
  - Complicated by ID changes that occur unrelated to station moves and some stations are members of multiple networks

- Comparisons of overlapping observational data between sources often required to match stations
  - Rarely exact data matches
  - For example in ISD dataset; Temperature, Dew Point Temperature, and Wind Direction all used to verify station match if 70% of data are within pre-selected ranges

Decisions and compromises must be made in matching stations from different sources
Issue 7: Quality Control

- NCDC relies heavily on automated, objective and repeatable processes for quality control of monthly, daily, and sub-daily data.
  - As new algorithms are developed entire period of record can be re-quality controlled.

- Manual corrections are NOT made operationally but there is a mechanism for incorporating manual edits retrospectively.
  - Web-based reporting mechanism (DATZILLA).

Daily maximum temperatures during April and May 1967 at Lärdeau, Canada (GHCN-Daily station ID = CA001144580), showing an example of data duplication identified by the duplicate check comparing data from different months within a year.
Many more errors could be resolved at the source of the problem or during the transmission pathway
- Improved training in NHMC’s on data collection and submission

GCOS Monitoring Centers proven very effective at identifying problems on the monthly timescale
- Expand role of GCOS Monitoring Centers to include greater focus on daily observations
- Expand emphasis on collection and transmission of daily climate observations
Issue 8: Metadata

- Metadata often consists of no more than station location and elevation
- The GHCN–Monthly dataset development process included development of additional metadata using other sources such as
  - Digital Elevation Models
  - Night–Light data from Satellites to determine urban versus rural areas
  - Operational Navigation Charts to assess distance to airports, oceans, lakes, and other topographical features.
Issue 8: Metadata Needs

- Metadata collection at NCDC for stations outside U.S. networks has received little attention in comparison to U.S. networks
- Detailed station histories have yet to be fully exchanged internationally
  - Is a standard station history format needed? or more attention needed?
- Most available global metadata at NCDC is outdated
- Little to no information on observing instruments, practices, and surrounding areas
- In the last two decades greater capabilities emerged such as higher density gridded topographical data and mapping capabilities such as Google Maps
Widespread land surface measurements since 1880
- ~650 stations in 1880
- ~6000 stations in 1970
- ~2500 stations provide observations on a monthly basis today

Changes through time artificially influence the climate record and must be removed

Changes due to:
- Instrumentation
- Observing practices
- Location
- Environment
Adjustments to remove the influence of artificial influences on the climate record can result in a **Smaller** overall station trend.

- **Petropavlovsk, Russia**
  - Unadjusted
  - Adjusted

- **Cape Town, South Africa**
  - Unadjusted
  - Adjusted

- Or a **Larger** overall station trend.
The need for bias adjustments adds to uncertainty in global temperature measurements

More attention needs to be given to benchmarking bias adjustment methodologies

More could be done to assess structural uncertainty associated with selection of parameters used in bias adjustment algorithms
Issue 10: Data and Metadata Access

- All GHCN–Monthly and GHCN–Daily data are provided for free
  - Restrictions have been placed on release of data from some countries
- Little attention in the past was given to providing the underlying software (quality control, bias adjustment, analysis)
- Source data in general have not been provided in the past
- Information on methods generally available only in the peer–review literature
More attention is being dedicated to openly providing software (quality control, bias adjustment, analysis)

As well as providing the underlying source data used to develop the baseline datasets

Providing more diagnostic analysis tools so that users can more easily assess the datasets including impact of bias adjustments is a high priority
NCDC’s monthly, daily, and sub-daily datasets serve a wide array of climate service needs.

Attention is given to a range of issues associated with dataset development and maintenance including data acquisition, data rescue, quality control, bias adjustment, and data access.

International collaboration has been key to the success of these efforts.

Closer partnerships are required to address remaining gaps and better meet the expanding needs of the climate community.
Take-Aways

Issue 1: The greater availability of digitized daily observations and potential for even more provides better opportunities to more closely link monthly and daily datasets

- But improvements in transmission and sharing of daily climate observations still needed

Issue 2: No one single source of data is sufficient for developing and maintaining a global data set.

- Requires continual efforts to maintain a data set that is healthy and can meet the needs of a large and diverse user community
Take-Aways

- **Issue 3:** Without continued attention to acquire and incorporate data from other sources, the number of stations in the datasets suffer decline
  - Many stations do not report in near real-time
  - Addition of other stations requires years of effort to locate, collect, process, and incorporate

- **Issue 4:** Need a data transmission pathway for daily climate observations (e.g., similar to monthly CLIMAT)
  - Adopt international mechanisms for standardizing exchange of the highest resolution data for climate purposes
  - Consider other opportunities for expansion of bilateral arrangements for sharing daily and sub-daily data
Issue 5: Formal data modernization efforts like NCDC’s CDMP Program are effective but costly and resource intensive
   ◦ Consider other possibilities such as Crowdsourcing

Issue 6: Addressing Formatting, Duplicates, Metadata issues is resource intensive but essential parts of dataset development

Issue 7: Automated Quality Control processes are effective but improved monitoring and education can help correct problems nearer their source
Take-Aways

- Issue 8: Metadata collection at NCDC for stations outside U.S. networks has received little attention in comparison to U.S. networks
  - Detailed station histories have yet to be fully exchanged internationally (Is a standard station history format needed? or more attention needed?)

- Issue 9: More evaluation of effectiveness of bias adjustment methodologies needed

- Issue 10: More open access to source data and algorithms/software desired