

WCRP CONFERENCE FOR LATIN AMERICA AND THE CARIBBEAN: DEVELOPING, LINKING AND APPLYING CLIMATE KNOWLEDGE

THE INTERNACIONAL SURFACE TEMPERATURE INITIATIVE GLOBAL LAND SURFACE DATABANK: UPDATE AND RECENT DEVELOPMENT

W. G. Almeida^{1*}, J. Christy², M. Flannery³, B. Gleason⁴, A. Klein⁵, A. Mhanda⁶, K. Ishihara⁷, J. H. Lawrimore⁸, D. Lister⁹, M. Menne¹⁰, V. Razuvaev¹¹, M. Renom¹², J. Rennie¹³, M. Rusticucci¹⁴, J. Tandy¹⁵, P. W. Thorne¹⁶, S. Worley¹⁷, A.S.B. Pereira¹⁸

^{1*,18} CPTEC/INPE, ² UAH, ³ BOM, ^{4,8,10} NOAA, ⁵ KNMI, ⁶ ACMAD, ⁷ JMA, ⁹ CRU, ¹¹ RIHMI-WDC, ¹² UdelaR, ^{13,16} CICS-NC, ¹⁴ UBA, ¹⁵ MetOffice, ¹⁶ NCAR.

E-mail waldenio.almeida@cptec.inpe.br

SURFACE TEMPERATURE RECORDS THE BEGINNINGS

Instrumental record of temperature has its roots in the development of standard temperature scales in the 18th century. Monthly mean temperature series for De Bilt, Netherlands extends from 1666 to the present. Several other long European series exist going back over 200 years.

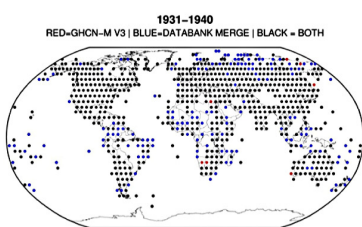
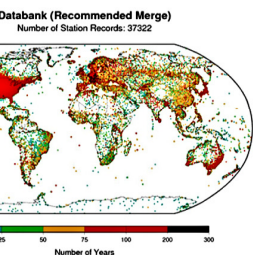
Throughout the 1800s measurements expanded across other continents. National Meteorological and Hydrological Services (NMHS) around the world have operated networks to support weather and climate observations since the late 19th Century.

DATABANK HERITAGE

Not until the 1980s and 1990s that major efforts were made to collect observations and create consolidated global datasets. The Global Historical Climatology Network-Monthly dataset: 7280 stations with monthly mean, maximum, and minimum temperature. CRUTEM: a global dataset of more than 4000 stations is still used today. GISS: Global surface temperatures based on GHCN-M data. These datasets are the foundation for understanding trends in surface temperature.

DATA LIMITATIONS

Global datasets met needs for our basic understanding of climate change. Existing deficiencies in data collection and exchange practices have weakened the credibility of climate assessments; insufficient coverage, particularly before 1950; lack of adequate metadata; lack of data provenance; limited data accessibility; and where additional sources of data exist, efforts to collect and integrate into a single database have lagged.



METADATA DEFICIENCIES

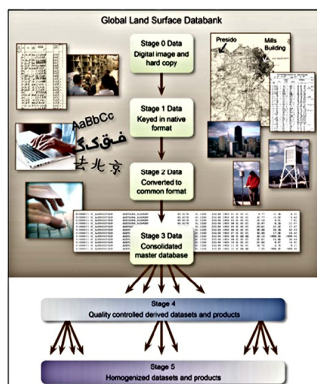
Metadata records are incomplete and inadequate for fully understanding uncertainty. Metadata consists of no more than station location and elevation. U.S. metadata collection for stations outside U.S. networks has received little attention; station histories have yet to be fully exchanged internationally. Available global metadata at NCDC is outdated. Lack of information on observing instruments, station moves, observations, and station environment. Metadata are especially important in the assessment and correction of inhomogeneities in the climate record.

DATABANK SOLUTION

In response to these needs, efforts to develop a global land surface Databank were initiated as part of the International Surface Temperature Initiative. The activity is overseen by a Databank Working Group (DWG) which reports to the ISTI Steering Committee. It leverages design principles and lessons learned from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) effort. It is a successful program that has produced and maintained an international up-to-date dataset of global ocean measurements since the 1950s.

DATABANK DESIGN

The Databank is being constructed and will be available in six Stages from initial observation to the final quality controlled and bias corrected dataset. The initial focus is on temperature, with the daily and monthly times-series, though other elements and time-series will be added later.



SOURCE DATA

Stage 0: Original observation

Stage 1: Native keyed format.

- Databank policy encourages data be provided in its rawest form; that closest to the measurements that were first reported by the observer.
- Ideally no quality control or homogenization should be applied prior to submission

Stage 2: Common Formatted Data

- This step appends data provenance to help users understand the history of each observation.

- Stage 2 format is ASCII and each data source is in a separate subdirectory.

- An inventory file is produced containing any available metadata. At a minimum this typically consists of a station id, name, latitude, longitude, elevation, and beginning and ending year.

DATA PROVENANCE

To provide a traceable record, Data Provenance Tracking (DPT) flags are added to the data in Stage 2:

- A DPT flag is a 3- to 4-digit numeral or alpha character representing unique information regarding each observation.
- There are currently five DPT flags: (1) Stage-0 Source, (2) Stage-1 Source, (3) Data Type, (4) Mode of Digitization, and (5) Mode of Transmission/Collection.
- Additional flags can be added in the future, for example to specify instrument type as sufficient metadata becomes available.

STAGE 3: MERGED DATASET

Each source is evaluated for merging into a single Stage 3 dataset. A source hierarchy is established based on factors such as:

- Whether the monthly data was calculated from dailies held in the databank;
- Whether the data arises from World Weather Records / national holdings;
- Average length of station record in the deck;
- Number of stations in the data deck;

The process of merging sources is complex due to the nature of weather and climate data:

- Collected by hundreds of thousands of observers in hundreds of countries often using differing languages, observing methods, and documenting and archive procedures.

MERGE PROCESS

The process of merging sources is based on comparisons of Metadata and Data between master and candidate stations. Temperature records for a station may be provided in many different source datasets:

- In some cases the records may overlap in time and be non-overlapping in others

Monthly mean temperatures can be calculated in many different ways:

- The temperatures for the same station from different sources are often similar but not exactly the same

In densely populated areas it also can be especially difficult to distinguish two or more unique stations because the temperatures are very similar.

METADATA TEST

The first step of the merge process is based on metadata comparisons:

- Station name (Jaccard Index);
- Location;
- Elevation;

For every master and candidate station a probability that they are the same station (and should be merged) is calculated:

- Likelihood of a match decreases with increasing differences in name, distance and elevation;

Those for which there is at least a 50% probability of a match are held over for data comparisons to further evaluate the likelihood of a match:

- The metadata threshold is set relatively low to account for the possibility that there are errors in metadata such as inaccurate location or elevation.

TESTS FOR OVERLAPPING DATA

Stations for which there is a possible Master to Candidate match based on metadata are held over for data testing.

Overlap Tests: For stations having at least 60 months of overlapping data.

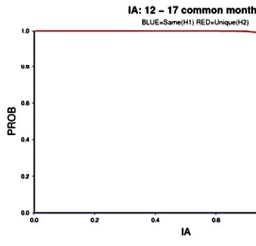
Metrics based on Index of Agreement and Normalized RMSD currently being evaluated.

Probability of same station (H1) and probability that candidate station is unique (H2) are calculated:

- Based on value of Metric and Number of Months of overlap.

Jared could produce a merge and a unique example. E.g., showing a Master station series and the overlapping data for candidate stations pointing out which one was selected for merge.

And/or a graphic with a Master and a Candidate where the IA test and metadata showed it to be unique.



TESTS FOR OVERLAPPING DATA

Probabilities from Metadata and Overlap tests are combined to determine whether the candidate should be merged with a station in the source or added as a unique station:

- ProbSameStn = Pmeta * H1;
- If ProbSameStn > 0.5, merge stn with highest Prob;
- ProbUniqStn = (1-Pmeta) * H2;
- If ProbUniqStn > 0.75, add candidate as unique stn;

TESTS FOR NON-OVERLAPPING DATA

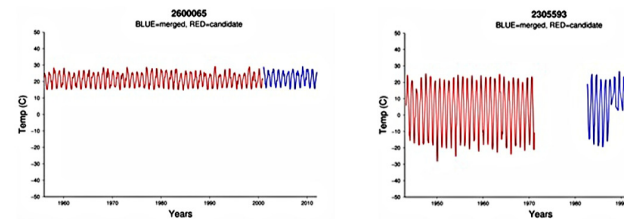
For master and candidate records that don't have at least 5 years of overlapping data.

This potentially includes tests for differences in mean and variance:

- the t-test for mean, and the F-test for variance;

A minimum number of non-overlapping years in the target and source will be required to perform such tests.

Using the resulting p-value, as well as the degrees of freedom, t is fit into their respective distributions, and a probability value is calculated.

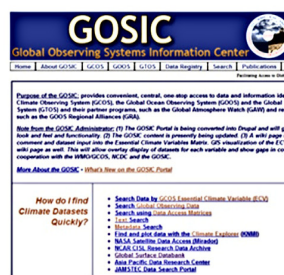


DATABANK ACCESS

The Databank is provided via the Global Observing System Information Center (GOSIC) website at: http://gosic.org/GLOBAL_SURFACE_DATABANK/GBD.html

A primary and mirror ftp site are available: <ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/> <ftp://ftp.meteo.ru/pub/data/globaldatabank/>

An open access wiki for the Databank WG is available at: http://editthis.info/intl_surface_temp_initiative/



DATA SUBMISSION

Submitting Data to the Databank is Easy:

- Data submissions are accepted in any format;
- Data can be provided via FTP, E-mail, or CD-ROM;

Our Databank submission guidance letter provides additional details:

- available at <http://www.surface-temperatures.org/databank/>;
- Please contact Jay.Lawrimore@noaa.gov and Jared.Rennie@noaa.gov with any questions or to submit data.

FOLLOWING STAGE 3

Stage 3 of the Databank provides the foundation from which methods of analysis, consistent benchmarking of performance and data access to end-users will be established.

Development of quality controlled (Stage 4) and Homogeneity data (Stage 5) is being led by the Benchmarking Working Group:

- Accompanying presentation.

FINAL THOUGHTS

Version 1 of the Databank will be released by the start of the summer. This is its first development cycle:

- Methods developed for version 1 will be improved upon and incorporated in future releases;

This is an open process that will only be successful by providing open access to the data, all methods, software, and provenance information.

It also depends on the contributions from many working group members and continuous feedback from the user community.

