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. Christy [2], M. Flannery [3], B. Gleason [4], A. Klein [5], A. Mhanda [6], K. Ishihara [7], J. H. Lawrimore [8], D. Lister [9], M. Menne [10], V. Razuvaev [11], M. Renom [12], J. Rennie [13], M. Rusticucci [14], J. Tandy [15], P. W. Thorne [16], S. Worley [17], A.S.B. Pereira.[18]

[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

rumental record of temperature has its roots in the development rsal temperature scales in the 18th century.

nthly mean temperature series for De Bilt, Netherlands extends 06 to the present.

eral other long European series exist going back over 200 years.

out the 1800s measurements expanded across other continents. onal Meteorological and Hydrological Services (NMHS) around d have operated networks to support weather and climate obser- since the late 19th Century.

DATABANK HERITAGE

ot until the 1980s and 1990s that major efforts were made to col- ervations and create consolidated global datasets.

e Global Historical Climatology Network-Monthly dataset: 7280 with monthly mean, maximum, and minimum temperature.

CRUTEM: a global dataset of more than 4000 stations is still ned today.

SA GISS: Global surface temperatures based on GHCN-M data.

atasets are the foundation for understanding trends in surface ture.

DATA LIMITATIONS

datasets met needs for our basic understanding of climate

existing deficiencies in data collection and exchange practices eatened the credibility of climate assessments;

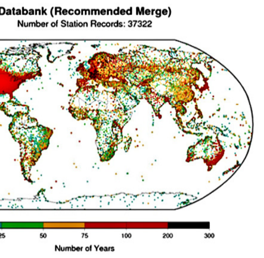
fficient coverage, particularly before 1950;

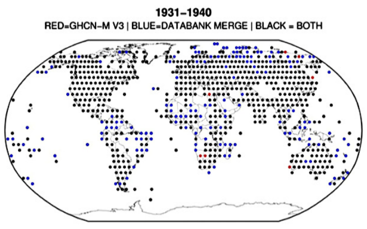
of adequate metadata;

r data provenance;

ted data accessibility;

n additional sources of data exist, efforts to collect and integrate ngle database have lagged.


Databank (Recommended Merge)
Number of Station Records: 37322
25
50
75
100
200
300
Number of Years



METADATA DEFICIENCIES

e metadata records are incomplete and inadequate for fully erizing uncertainty.

n consists of no more than station location and elevation

S. metadata collection for stations outside U.S. networks has re- ttle attention;

l station histories have yet to be fully exchanged internationally.

ailable global metadata at NCDC is outdated.

no information on observing instruments, station moves, obs s, and station environment.

h metadata are especially important in the assessment and cor- of inhomogeneities in the climate record.

DATABANK SOLUTION

nse to these needs, efforts to develop a global land surface Data- re initiated as part of the International Surface Temperature Ini-

activity is overseen by a Databank Working Group (DWG) which o the ISTI Steering Committee

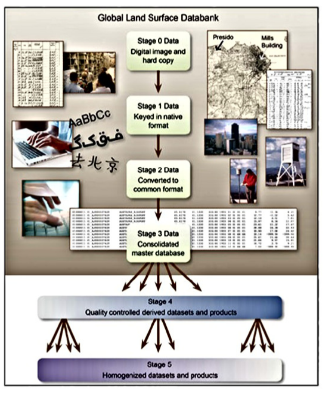
verages design principles and lessons learned from the Interna- omprehensive Ocean-Atmosphere Data Set (ICOADS) effort.

successful program that has produced and maintained an inte- nd up-to-date dataset of global ocean measurements since the 0s.

DATABANK DESIGN

tabank is being constructed le available in six Stages from inal observation to the final controlled and bias corrected

ial focus is on temperature the daily and monthly times- hough other elements and tim- will be added later.


Global Land Surface Databank
Stage 0 Data
Digital image and hard copy
Stage 1 Data
Keyed in native format
Stage 2 Data
Converted to common format
Stage 3 Data
Consolidated master database
Stage 4
Quality controlled derived datasets and products
Stage 5
Homogenized datasets and products
Presido
Mills Building
AaBbCc

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:

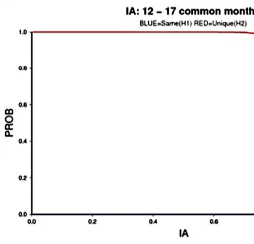
- The metadata threshold is set relatively low to account for the possibility that there are errors in metadata such as inaccurate location or elevation.

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 whether the candidate should be merged with a station in the source or added as a unique station:

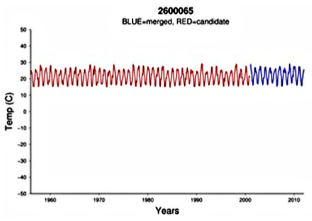
- ProbSameStn = Pmeta * H1;
- If ProbSameStn > 0.5, merge stn with highest Prob;
- ProbUniqStn = (1-Pmeta) + H2;
- If ProbUniqStn > 0.75, add candidate as unique stn;

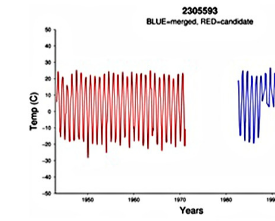
TESTS FOR NON-OVERLAPPING DATA

For master and candidate records that don't have at least 5 year lapping data.
This potentially includes tests for differences in mean and varia

- 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



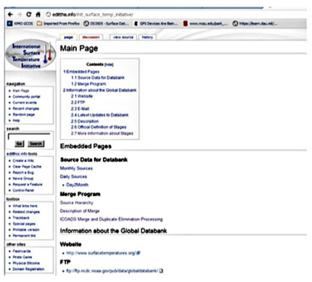


DATABANK ACCESS

The Databank is provided via the Global Observing System Info 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/



Main Page

Contents
1 Embedded Pages
1.1 Source Data for Databank
1.2 Merge Program
2 Information about the Global Databank
2.1 Website
2.2 FTP
2.3 E-Mail
2.4 Latest Updates to Databank
2.5 Description
2.6 Official Definition of Stages
2.7 More information about Stages

Embedded Pages

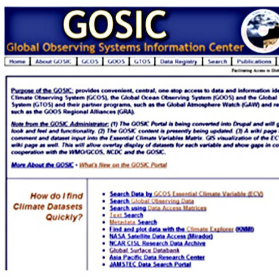
Source Data for Databank
Monthly Sources
Daily Sources
- Day2Month

Merge Program
Source Hierarchy
Description of Merge
ICOADS Merge and Duplicate Elimination Processing

Information about the Global Databank

Website
- http://www.surfacetemperatures.org/

FTP
- ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/



Purpose of the GOSIC: provides convenient, central, one-stop access to data and information ide Climate Observing System (GCOS), the Global Ocean Observing System (GOOS) and the Global System (GTOS) and their partner programs, such as the Global Atmosphere Watch (GAW) and re such as the GOOS Regional Alliances (GRA).

Note from the GOSIC Administrator: (1) The GOSIC Portal is being converted into Drupal and will g look and feel and functionality. (2) The GOSIC content is presently being updated. (3) A wiki page comment and dataset input into the Essential Climate Variables Matrix. GIS visualization of the EC wiki page as well. This will allow overlay display of datasets for each variable and show gaps in co cooperation with the WMO/GCOS, NCDC and the GOSIC.

More About the GOSIC - What's New on the GOSIC Portal

How do I find Climate Datasets Quickly?

- Search Data by GCOS Essential Climate Variable (ECV)
- Search Global Observing Data
- Search using Data Access Matrices
- Text Search
- Metadata Search
- Find and plot data with the Climate Explorer (KNMI)
- NASA Satellite Data Access (Mirador)
- NCAR CISL Research Data Archive
- Global Surface Databank
- Asia Pacific Data Research Center
- JAMSTEC Data Search Portal

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 de

- available at http://www.surfacetemperatures.org/databank;

Please contact Jay.Lawrimore@noaa.gov and Jared.Rennie@n with any questions or to submit data.

FOLLOWING STAGE 3

Stage 3 of the Databank provides the foundation from which ne ods of analysis, consistent benchmarking of performance and d ing to end-users will be established.
Development of quality controlled (Stage 4) and Homogeneity data (Stage 5) is being led by the Benchmarking Working Grou

- Accompanying presentation.

FINAL THOUGHTS

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

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

This is an open process that will only be successful by providing u access to the data, all methods, software, and provenance infor
It also depends on the contributions from many working group m and continuous feedback from the user community.