The International Surface Temperature Initiative's Global Land Surface Databank

J. H. Lawrimore[1], J. Rennie[2], W. Gambi de Almeida[3], J. Christy[4], M. Flannery[5], B. Gleason[1], A. Klein-Tank[6], A. Mhanda[7], K. Ishihara[8], D. Lister[9], M. Menne[1], V. Razuvaev[10], M. Renom[11], M. Rusticucci[12], J. Tandy[13], P. W. Thorne[2], S. Worley[14]

[1]*NOAA's National Climatic Data Center, Asheville, NC, USA*
[2]*Cooperative Institute for Climate and Satellites, NCSU & NOAA's National Climatic Data Center, Asheville, NC*
[3]*Instituto Nacional de Pesquisas Espaciais, Centro de Previsão de Tempo e Estudos Climáticos, Brazil*
[4]*University of Alabama-Huntsville, Huntsville, AL, USA*
[5]*Bureau of Meteorology, Melbourne, Australia*
[6]*Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands*
[7]*African Centre of Meteorological Applications for Development, Niamey, Niger*
[8]*Japan Meteorological Agency, Tokyo, Japan*
[9]*Climatic Research Unit, UEA, Norwich, UK*
[10]*Russian Research Institute of Hydrometeorological Information, Obninsk, Russia*
[11]*Universidad de la Republica, Montevideo, Uruguay*
[12]*University of Buenos Aires, Argentina*
[13]*Met Office Hadley Centre, Exeter, United Kingdom*
[14]*National Center for Atmospheric Research, Boulder, CO, USA*

Abstract. The International Surface Temperature Initiative (ISTI) consists of an end-to-end process for land surface air temperature analyses. The foundation is the establishment of a global land surface Databank. This builds upon the groundbreaking efforts of scientists in the 1980s and 1990s. While using many of their principles, a primary aim is to improve aspects including data provenance, version control, openness and transparency, temporal and spatial coverage, and improved methods for merging disparate sources. The initial focus is on daily and monthly timescales. A Databank Working Group is focused on establishing Stage-0 (original observation forms) through Stage-3 data (merged dataset without quality control). More than 35 sources of data have already been added and efforts have now turned to development of the initial version of the merged dataset. Methods have been established for ensuring to the extent possible the provenance of all data from the point of observation through all intermediate steps to final archive and access. Databank submission procedures were designed to make the process of contributing data as easy as possible. All data are provided openly and without charge. We encourage the use of these data and feedback from interested users.

Keywords: Climate change, climate dataset construction, data provenance.

INTRODUCTION

Universal temperature scales were not developed until the early 18th century when the scale most closely resembling today's Fahrenheit scale was developed. This was followed by the work of Anders Celsius that was eventually extended to become today's standard scientific temperature scale (1).

These efforts made possible the record of temperature that today provides insight into the Earth's climate. The Central England Temperature record began in 1659. More than 100 years of this record were based on instrumental measurements, some estimated from measurements in indoor unheated rooms, combined with non-instrumental weather diary entries. A daily series considered to be truly representative did not begin until 1772 (2). An even longer continuous record of temperature for a single location is the monthly mean temperature series for De Bilt, Netherlands, which extends from 1706 to the present (3). Several other long European series exist going back over 200 years.

Some early records were made in North America in the late 1700s. Throughout the 1800s measurements expanded across other continents. These early records were carefully made by professionals who had the skills and training to operate and care for the delicate meteorological instruments. As instruments became cheaper and more durable, it became possible for an even greater expansion (4). National Meteorological and Hydrological Services (NMHS) around the world have operated networks to support weather and climate observations since the late 19th Century.

It was 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 contained more than 6000 stations when it was released in 1992 (5). A second version contained 7280 stations with monthly mean, maximum, and minimum temperature (6). An independent effort was made to create CRUTEM at about the same time and this global dataset of more than 4000 stations is still maintained today (7).

In the early part of the 21st century attention turned to daily data. The Global Historical Climatology Network-Daily dataset (8) provides daily maximum, minimum, and mean temperature for more than 25,000 stations. These records are generally shorter duration than monthly means with most not beginning until the middle of the 20th century and large gaps still present, particularly in the Southern Hemisphere.

These and other monthly and daily global datasets provide the foundation for studying variation and change in the Earth's climate over the past 100 to 200 years. While these have led to tremendous advances in understanding, there remain impediments due to residual deficiencies in global collections. Regional and local scale assessments are constrained by limited spatial coverage in many regions, especially in the 1800s (Figure 1). Although additional sources of data exist, often in their original manuscript form or more recently as images of the original forms, digitization efforts that make the integration into datasets possible have lagged.

Available metadata records are incomplete and inadequate for fully characterizing uncertainty associated with changing observing practices, instrumentation, and environmental conditions surrounding the station. Such metadata are especially important in the assessment and correction of inhomogeneities in the climate record (9). Although this information is often maintained in NMHS archives, in most cases metadata have not been included in data exchange activities.

There also has been limited attention given to the need for version control and provenance tracking in the construction of datasets. For decades, climate scientists were focused on building datasets with the best temporal and spatial coverage possible, separating valid from invalid reports, and developing methods to remove inhomogeneities from the record. External pressures that would lead to doubts about the integrity of the underlying data and the steps involved in the calculation of global temperatures were never envisioned. Only in recent years has there developed a need for scientists to better document the provenance and implement version control from the point of measurement through dissemination, quality control, bias correction, archive and access. The requirement to be fully open and transparent as to the details associated with each processing step includes the need to provide access to the software associated with each data processing step, quality control, and bias correction. By putting in place new practices the wider community will have the opportunity to more fully engage in the process. This should engender greater public confidence and understanding.

In response to these needs, efforts to develop a global land surface Databank were initiated as part of the International Surface Temperature Initiative (ISTI). This activity is overseen by a Databank Working Group (DWG) which reports to the ISTI Steering Committee (10). It leverages design principles and lessons learned from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) effort; a highly successful program that has produced and maintained an integrated and up-to-date dataset of global ocean measurements since the mid-1980s (11).

DATABANK DESIGN

The Databank is being constructed and made available in six Stages from the original observation to the final quality controlled and bias corrected product (refer to Figure 2 in accompanying ISTI conference proceedings paper). The initial focus is on temperature data on the daily and monthly timescale, although other elements and timescales will be added later.

Stage-zero consists of observations in their original form. The historical record consists primarily of observations recorded on paper and housed in NMHS archives and other locations such as national museums. Over the past two decades there has been a transition towards fully automated networks that operate without the need for an observer. However, there remain thousands of stations that continue to rely on paper records. Many paper records have been converted to photographic or scanned images over the past decade through programs such as NOAA's Climate Database Modernization Program (12), and the International Environmental Data Rescue Organization (IEDRO) (13). Such images are essential to preserve the original observations. In other cases only the original paper form or possibly a microfiche or microfilm copy exists. These sources may not physically reside on the Databank server, but when the location is known its archived location is documented.

Stage-1 consists of digital data in native format. This is beneficial in that it does not require extra effort on the part of the data provider to perform reprocessing and reformatting while reducing the possibility that errors could occur during translation.

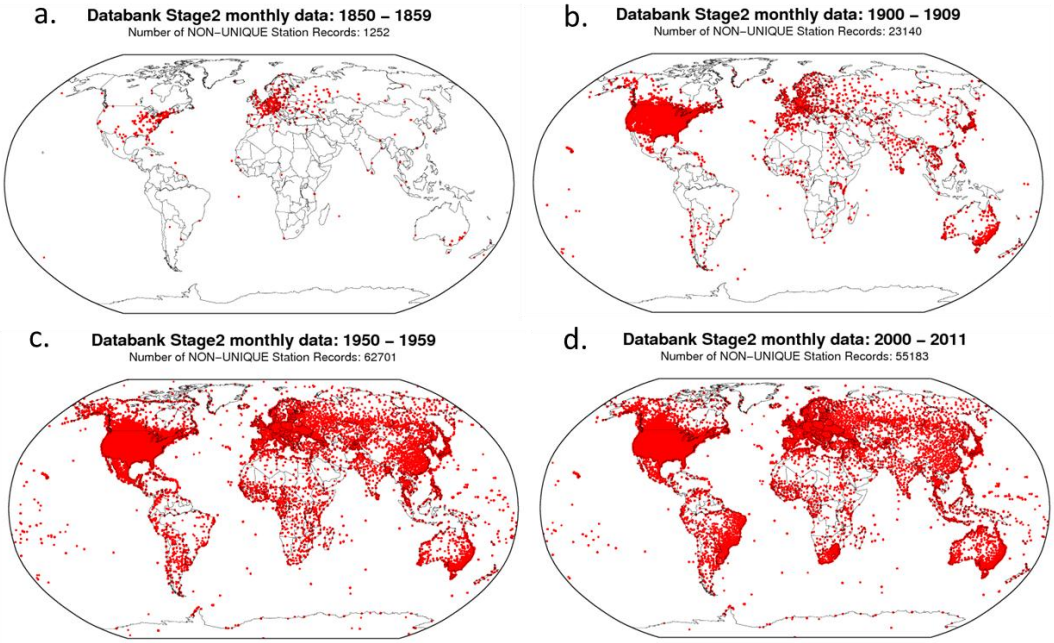

FIGURE 1. Number of non-unique stations available to the Databank in digital form as of January 14, 2012 for each of four periods from 1850 through 2010. Stations have at least one observation of monthly mean temperature during each period.

Databank policy encourages data be provided in its rawest form; that closest to the measurements that were first reported by the observer (10). Ideally no quality control or homogenization should be applied prior to submission so that the provenance can be better assured leading up to and through the point where quality control of the Databank is accomplished in a consolidated and automated way.

However if the original raw observations do not exist, the quality controlled and/or homogeneity adjusted data will be accepted. The details of such processing applied prior to submission are collected and retained. This information is used in merging and remains with the source data to support future decisions regarding its use.

Following Stage-1, all data are converted to a common format in Stage-2. 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. Accompanying this is a map which shows the locations and the number of years of data in their record.

Data Provenance

To provide a traceable record Data Provenance Tracking (DPT) flags are required. Stage-2 data provides the first opportunity to assign such flags. 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. The information contained within each DPT flag completely defines an observation.

DPT 1 defines the Stage-0 Source from which the observation originated. Sources include NMHS hosts such as the Japan Meteorological Agency and the Australian Bureau of Meteorology, universities such as University Rovira I Virgili, University of Alabama-Huntsville, and internationally sponsored programs such as the World Meteorological Organization's World Weather Records.

DPT 2 describes the source of the Stage-1 data. This source may differ from the Stage-0 data provider or provide additional information such as the name of the host's dataset from which the data originated.

DPT 3 indicates if the data provided by the host had been previously quality controlled or homogeneity adjusted.

DPT 4 describes the mode of digitization and the institution responsible.

DPT 5 provides the mode of transmission and collection. This describes the process used to transfer the data to the Databank.

Data Merging

Next data are merged into a single Stage-3 dataset. This is fraught with many complexities associated with the nature of weather and climate data which were collected by hundreds of thousands of observers in hundreds of countries often using differing languages, observing methods, and documenting and archive procedures. Often metadata provide only the most basic information such as station name and location, and often even this information is inaccurately recorded.

Because all stages of data are provided within the Databank, it is possible for any interested individual or organization to implement their own unique merging technique for creating a merged dataset and this is encouraged. Nevertheless, the ISTI is currently developing a merging methodology which will be applied to development of a Stage-3 dataset. This will be fully documented and made available along with all source code used in performing the merge. This is an evolving process with refinements expected to be made on a continuing basis in coming months and years.

Because many sources may contain records for the same station it is necessary to create a process for identifying and removing duplicate stations, merging some sources to produce a longer station record, and in other cases for determining when a station should be brought in as a new record.

First a source hierarchy is created. Prioritization is based on a number of criteria. Because ISTI places special emphasis on data provenance, the Stage-3 databank holdings are envisaged to constitute as close to the raw data as possible, ideally with provenance tracking back to the raw, hard copy record. Monthly mean maximum and minimum temperature are preferred because they can be directly used to calculate monthly mean temperature. In cases where only monthly mean temperature data are provided it is often unclear what method was used for its computation (3). In addition, biases affecting maximum temperature can differ from those affecting minimum temperature, necessitating different corrections (14).

With this framework in mind, prioritization of the 36 Stage-2 sources for the merge process is accomplished. One possible hierarchy is shown in Table 1. Different decisions may lead to a different hierarchy, and further development is needed before a final hierarchy is established. The merge process occurs iteratively, starting from the highest priority data source (target) and progressing through all the source decks (candidates). Potential approaches include the use of a Bayesian approach based upon metadata matching and data equivalence criteria.

Metadata Comparisons

There are at least three geolocation characteristics which can be used to identify potentially matching or definitively unique stations. The distance between stations based upon latitude and longitude fitted to an exponential decay function decaying from 1 at no distance to zero at 100km, and the probability that the two stations are the same returned as this value. This can be followed by a similar approach using the height difference between the two stations. A third involves a test of the similarity of the station name, using a measure such as the Jaccard Index (JI), which is defined as the intersection divided by the union of two sample sets. The Jaccard Index looks for cases in which certain letters exist in both station names, as well as the number of times letters occur in one name, but not in the other.

These three geolocation metrics have a probability from 0 to 1. Using a simple Bayesian approach, they can be multiplied and a combined probability returned that the two stations are the same. If this surpasses a threshold further evaluation based on data comparisons can begin. This threshold should be set low enough to account for the possibility that there are errors in metadata.

Data Comparisons

There are two distinct types of scenarios for data comparisons. Those where station data overlap, and those where they do not. For cases with overlap, a direct comparison of observations during the same months and years can made. For cases in which data do not overlap, testing for data equivalence is required. Potential approaches include the generation of a Bayesian probability, combining both the geolocation and data probabilities. This probability can be evaluated to determine whether the target and candidate observations are from the same or different stations.

If it is concluded that the candidate station is the same as the target station a merge can be performed. Only data not already in the target station record will be added. Preference is always given to the target, since it contains data that were higher in priority. If a candidate station goes through the entire target dataset and no match is found, then the station is deemed unique. Further details associated with the merge process are under development.

Table 1. The 36 Stage-2 sources in the Databank and their priority as of January 24, 2012. These sources establish the foundation from which the Stage-3 merged dataset will be created.

Priority	Source	Priority	Source
1	GHCN-Daily raw (NCDC)	19	Spain (Univ. Rovira i Virgili)
2	Mexico (CDMP)	20	Russia (Roshydromet)
3	Vietnam (CDMP)	21	Uruguay (Inst. Nacional de Invest. Agropecuaria)
4	US Forts (CDMP)	22	Switzerland (Digihom/MeteoSwiss/IAC-ETH)
5	Channel Islands (States of Jersey Met)	23	Tunisia/Morocco (ISPD)
6	Ecuador (Inst. Nacional De Met E Hidrologia)	24	Europe/N. Africa (ECA Daily/KNMI)
7	Pitcairn Island (Met Service of New Zealand)	25	Southeast Asia (SACA/KNMI)
8	Beirut (Univ. of Giessen)	26	Japan (Japan Met Agency)
9	Brazil (INPE, Nat. Institute for Space Research)	27	UK Met Office Historical (UKMO)
10	Miscellaneous (NCDC)	28	Europe/N. Africa (ECA Monthly/KNMI)
11	World Weather Records (WMO)	29	GHCN-M v2 Source (NCDC)
12	Colonial Era Archives (Griffith)	30	GHCN-M v2 (NCDC)
13	East Africa (Univ. of Alabama-Huntsville)	31	Central Asia (NSIDC)
14	Antarctica South Pole (Univ. of Wisc.-Madison)	32	Canada (Env. Canada)
15	Switzerland (ISPD)	33	Australia (BOM)
16	Polar (ISPD)	34	Arctic (IARC/Univ. of Alaska Fairbanks)
17	Sydney (ISPD)	35	Greater Alpine Region (Histalp/ZAMG)
18	Antarctica (SCAR Reader Project)	36	HadCRUT3 (UKMO)

DATA ACCESS AND VERSION CONTROL

Data are provided from a primary ftp site hosted by the Global Observing Systems Information Center (GOSIC; http://gosic.org) and World Data Center A at NOAA/NCDC. In addition World Data Center B at Oblinsk, Russia established an ftp site that is routinely updated to mirror the data on the primary site.

ftp://ftp.ncdc.noaa.gov/pub/data/globaldatabank/
ftp://ftp.meteo.ru/pub/data/globaldatabank/

All data are provided in ASCII to facilitate access and ease of use. Future efforts may include conversion to NetCDF Climate and Forecast (CF) convention.

In some cases the data provider has agreed to contribute regular data updates. Upon updates the previous version is moved to an archive directory and permanently stored. Within the archive directory each version is maintained and designated by the year, month, and day the data were first received. It is preferable that the entire source dataset be transferred as updates are made rather than collection of only the most recent observations. Acquiring the full source better ensures the most up-to-date data.

A version number is assigned to new sources or updates to sources as they are added. All files from a single source are combined into a single tar file compressed using gzip. The version number is contained within a naming structure:

source.timescale.stage#.X.Y.yyyymmdd.tar.gz

where

1. *source* identifies the data provider.
2. *timescale* is monthly, daily, or hourly.
3. *stage#* is currently either Stage-1 or Stage-2.
4. X is incremented when there is a major change to the source dataset such as replacement or addition of a large percentage of data.
5. Y is incremented when there are small updates to the source dataset such as real-time updates to existing stations.
6. yyyymmdd is the year, month, and day the data source was provided or updated.

OPPORTUNITIES TO CONTRIBUTE

Databank submission procedures are designed to make the process easy while ensuring the submitted data are of high quality and traceable. Policies require submission of information about the contributed data

including file formats and metadata such as station location and name. Data should be provided in the original native format e.g ASCII text, Microsoft Excel, XML, NetCDF. A complete guide to data submission procedures is available online (http://www.surfacetemperatures.org/databank).

CONCLUDING REMARKS

Construction of a land surface databank is a major undertaking requiring time and international coordination. It has been preceded by many groundbreaking efforts and comes at a time when the need for high quality, traceable, and complete data is clearer than ever.

As an integral part of the ISTI, the global land surface Databank provides the foundation from which new methods of analysis, consistent benchmarking of performance and data serving to end-users will be established. Information regarding how the Databank effort fits within the broad effort of the ISTI is provided in an accompanying paper. Further information also can be found through www.surfacetemperatures.org. Constructive comments are encouraged and can be provided at http://surfacetemperatures.blogspot.com/.

ACKNOWLEDGMENTS

We thank the many contributors of data that have made establishment of the Databank possible.

REFERENCES

1. Knowles Middleton, W.E., *A History of the Thermometer and its use in Meteorology*, Baltimore, Maryland: The John Hopkins Press, 2002, pp. 5-104.
2. Parker, D.E., T.P. Legg, C.K. Folland, 1992: A New Daily Central England Temperature Series, 1772-1991. International Journal of Climatology. Vol. 12, 317-342.
3. Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie (2011), An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3, J. Geophys. Res., 116, D19121, doi:10.1029/2011JD016187.
4. National Weather Service, 2011: *What is the COOP Program?* [web site] http://www.nws.noaa.gov/om/coop/what-is-coop.html
5. Vose, R. S., R. L. Schmoyer, P. M. Steurer, T. C. Peterson, R. Heim, T. R. Karl, and J. Eischeid, 1992: The Global Historical Climatology Network: Long-term monthly temperature, precipitation, sea level pressure, and station pressure data. ORNL/CDIAC-53, NDP-041, 325 pp. [Available from Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831.]
6. Peterson, T. C., and R. S. Vose (1997), An overview of the Global Historical Climatology Network temperature database, Bull. Amer. Meteorol. Soc., 78, 2837–2849.
7. Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, Hemispheric and large-scale land surface air temperature variations: An extensive revision and an update to 2010. *J. Geophys. Res.*, doi:10.1029/2011JD017139, in press.
8. Menne, M.J., I. Durre, R.S. Vose, B.E. Gleason, and T.G. Houston, 2011: An overview of the Global Historical Climatology Network Daily Database. *Journal of Atmospheric and Oceanic Technology*, submitted.
9. Thorne, Peter W., and Coauthors, 2011: Guiding the Creation of A Comprehensive Surface Temperature Resource for Twenty-First-Century Climate Science. *Bull. Amer. Meteor. Soc.*, **92**, ES40–ES47. doi: http://dx.doi.org/10.1175/2011BAMS3124.1
10. International Surface Temperature Initiative, 2011: "Databank effort" [web site] http://www.surfacetemperatures.org/databank.
11. Woodruff, S.D., Worley, S.J., Lubker, S.J., Ji, Z., Freeman, J.E., Berry, D.I., Brohan, P., Kent, E.C., Reynolds, R.W., Smith, S.R. & Wilkinson, C. (2009). ICOADS Release 2.5: Extensions and Enhancements to the Surface Marine Meteorological Archive. Int. J. Climatol., 31, 951-967, doi:10.1002/joc.2103.
12. Dupigny-Giroux, Lesley-Ann, Thomas F. Ross, Joe D. Elms, Raymond Truesdell, Stephen R. Doty, 2007: RESOURCES - NOAA's Climate Database Modernization Program: Rescuing, Archiving, and Digitizing History. *Bull. Amer. Meteor. Soc.*, **88**, 1015–1017.
13. International Environmental Data Rescue Organization (IEDRO), 2011: What is IEDRO? [web site] http://iedro.org/en/about/ourmission.html
14. Williams, C., M. J. Menne, P. W. Thorne, 2012: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geophys. Res.*, Accepted.