Creating surface temperature datasets to meet 21st Century challenges

Met Office Hadley Centre, Exeter, UK

7th-9th September 2010

White papers background

Each white paper has been prepared in a matter of a few weeks by a small set of experts who were pre-defined by the International Organising Committee to represent a broad range of expert backgrounds and perspectives. We are very grateful to these authors for giving their time so willingly to this task at such short notice. They are not intended to constitute publication quality pieces – a process that would naturally take somewhat longer to achieve.

The white papers have been written to raise the big ticket items that require further consideration for the successful implementation of a holistic project that encompasses all aspects from data recovery through analysis and delivery to end users. They provide a framework for undertaking the breakout and plenary discussions at the workshop. The IOC felt strongly that starting from a blank sheet of paper would not be conducive to agreement in a relatively short meeting.

It is important to stress that the white papers are very definitely not meant to be interpreted as providing a definitive plan. There are two stages of review that will inform the finally agreed meeting outcome:

1. The white papers have been made publicly available for a comment period through a moderated blog.
2. At the meeting the approx. 75 experts in attendance will discuss and finesse plans both in breakout groups and in plenary. Stringent efforts will be made to ensure that public comments are taken into account to the extent possible.

Data provenance, version control, configuration management

John R. Christy (UAHuntsville)
Nick Barnes (Clear Climate Code)
Amy Luers (Google)
Shawn Smith (FSU/COAPS)
Steve Worley (ICOADS/NCAR)
Karl Taylor (CMIP)
Jay Lawrimore (NCDC)

White paper authors are requested to consider: The metadata requirements to ensure traceability back to the original data record; the version controlling requirements (what represents Business As Usual and what constitutes a fundamental update to the databank); how to best retain previous versions of the databank; how to configure the archives to maximize usefulness (data format, file indexing, appending the homogenized data estimates etc.); how to version control and archive the datasets produced from the initial databank (see Section 8).

Introduction

Through the years it has become apparent that the development of an internationally-organized, transparent, and comprehensive data management system of surface air temperature has been sorely needed. Nations have stepped forward with extensive efforts to perform these functions, e.g. the U.S. National Climatic Data Center Global Historical Climatology Network (NCDC), the UK Met Office HadCRUT3 surface temperature data set, and the International Comprehensive Ocean-Atmosphere Data Set (ICOADS). These efforts, we believe, can be built upon to achieve an even greater level of effectiveness with international support and oversight for global land surface air temperature measurements managed in a centralized system.

The emphasis on this white paper is to address the requirements for establishing a data system whereby (a) original observations, (b) information about the observing system which made those observations (metadata), and (c) products generated from those observations, have the potential of meeting the reliability requirements desired not only of the scientific community but of many other communities (e.g. policy, legal, etc.) who now rely on climate data. Such a system will require robust methods of data provenance, version control and configuration management as described below and defined in Appendix A, which in this project relate primarily to Level 0 and 1 data categories (see wp3.) If such a system is established, the scientific requirements regarding climate data activities, including accessibility, traceability, reproducibility and reliability, will be met, and, we will have provided to other communities the information about climate data they need.

To aid in the production of derived products (e.g. Level 5 such as HadCRUTv3, GISS, ERSST) which may exclude and/or modify the archived, primary-source data through quality control, homogenization, and interpolation algorithms, we shall provide various

standardized testing datasets which may be used for assessing the skill of the algorithm. However, it is not the purview of this project to assess derived products as to whether they meet higher standards required by specific communities (i.e. scientific, legal, etc.) It is one aim of this project to archive and disseminate derived products as long as the homogenization algorithm is documented by the peer review process.

Date Provenance (Traceability to primary source)

Primary sources, referred to in this project as Level 0 data (see wp3), of instrumental temperature data fall into two broad categories (a) paper documents and (b) digital computer files from electronic sensors. Establishing an archive that provides a pathway from the heavily-used, uniformly formatted data records (Level 2 and 3 data) in the project's archive back to their primary source (or relevant information if no primary source exists) is a critical function of the project.

Paper Documents

The earliest primary-source climate records, and even many today, consist of hand-written or machine-printed values, or pin-traces on paper documents. Scientific (and other) research is performed using digital files created from these data after they have been electronically keyed, becoming Level 1 data.

Associated with these primary-source paper documents are many metadata documents that describe the instrument location (including maps), type of instrument, condition of site and instrument, directions for taking observations, calibration of instruments, etc. These metadata will also be archived according to "Levels" as described in wp3. This metadata information will require archiving as we anticipate that one of the major uses of our primary-source archive and its associated data records is the construction of homogenized long-time series for which metadata are vital.

One goal of this project is to make all images of the primary-source (Level 0) paper documents available to investigators to address traceability and authenticity. Unfortunately, in many cases the primary source documents have been lost, destroyed or for some reason have become unreadable. In their place quite often are secondary-source documents (i.e. official monthly summaries, newspaper reports, etc.) or digital files that may have been derived from a primary-source document before its demise. These data are known as Level 1 data but which may not have traceability to an archived Level 0 document or file. Traceability and authenticity in these cases are more difficult. Thus, there are two types of Level 1 data – that which is traceable to an archived Level 0 primary source and that which is not. As indicated below, there will likely be multiple versions of Level 1 data due to the loss of Level 0 for a particular station.

Electronically measured and transmitted data

There has been a relatively rapid conversion from printed data collection methods to electronic measuring, reporting, and quality control so that the human eye never

witnesses the observation nor its transmission and archival. In some of these cases, the electronically measured observation is produced originally as a geophysical parameter and reported on an electronic network, usually in an obscure, digitally-packed file structure. In these cases it is important to archive the original transmission as well as the unpacking algorithm so that traceability to and from the primary-source evidence may be achieved.

The output of some electronic sensors is recorded in raw data files that require specialized unpacking, conversion and calibration algorithms to generate temperatures. In the purest sense, the Level 0 data are machine-readable files of, for example, the voltages, digital counts, refractivities, etc. In such cases, both the fundamental data stream and the associated conversion algorithm would be considered together as Level 0, primary-source information.

As with the instrumental data records recorded by hand on paper, these electronic measurements and algorithms will require metadata documentation that defines the process. While these activities appear especially onerous to climate researchers, the reliability, reproducibility, and traceability requirements insist that such burdens be accommodated wherever possible. We recognize that in many electronic systems, these primary-source data and algorithms may be impossible to recover. In such circumstances where fundamental source data are unavailable, the project should provide in the associated Level 0 metadata archive of information, i.e. technical manuals, to describe the instrumentation and conversion techniques which generated the archived geophysical value at Level 1.

Because of the differing methods needed to discover and archive primary and secondary sources based on their original form (i.e. documents vs. digital files) and the time frame covered by these time series, it is anticipated that multiple, parallel (or perhaps sequential) efforts will be required by teams of experts. In other words, if funding is limited, the project may begin with data records deemed most vulnerable to loss, e.g. pre-1950 paper documents in developing countries.

- *As an outcome of the workshop, there should be a clear definition of primary (Level 0) and secondary (Level 1) source database across the spectrum of observing systems which may contribute data to the land surface temperature database.*

- *We should establish a coordinated international search and rescue of Level 0, primary-source climate data and metadata both documentary and electronic (see wp3.) This effort would recognize and support similar on-going national projects. Once located, the project should (a) provide, if necessary, a secure storage facility for these documents or hard-copies of same, (b) create, where appropriate, digital images of the documents for the archive for traceability and authenticity requirements, (c) key documentary information into digital files (native format in Level 1 and uniform format in Level 2), (d) archive, test and quality-assure raw data files, technical manuals and conversion algorithms which are necessary to understand how the geophysical variable*

may be unpacked and generated from electronic instrumentation, and (e) securely archive the files for public access and use.

- *A certification panel will be selected to rate the authenticity of source material as to its relation to the "primary-source", i.e. to certify a level of confidence that the Level 1 data, as archived, represents the original values from the Level 0 primary source. The process will often be dynamic, since we anticipate that new information will always become available to confirm or cast doubt on the current authenticity rating.*

Version Control

Real-world experience indicates that archives necessarily become dynamic libraries that must adapt to the unforeseen appearances of new, competing or corrected data sets. Time series which proceed through the Levels to the integrated databank (Level 3) must be carefully indexed for traceability. Indeed, given the wide variety of surface temperature observational methods and institutions that have been involved at some point in their collection and use, this project understands that version control is a vital and complex requirement.

Archives of historical observations can and should be updated as new information becomes available. This, at the outset, presents a very difficult problem for the project at hand, namely, at what scale (spatial and temporal) should version control be applied when new or revised data are discovered and accepted as authentic? Should the time series of each individual station be assigned a version number which may be up-versioned when new information is gathered? Should temporal blocks be considered for versioning (i.e. decade by decade or pre-1950 vs. post-1950, etc.)? Should the block of documents pertaining to metadata for a station be up-versioned when a new document is discovered? Should time series of stations be bundled into regional, national, continental or global extents for version designations and updated only occasionally when a significant number of changes are warranted? In cases with unformatted, raw digital data files, what should be done when a change is warranted in applying the unpacking and conversion algorithm of the original file?

A practical guideline to follow is that new versions should be adopted when there is a significant addition or change to be instituted. This change would be considered significant if products generated from the archive may be altered in some noticeable way with the new information. In the meantime, new data may be placed in a pre-authentication database, with appropriate caveats and indexing, for usage until an authenticity investigation is completed and the data deemed suitable for the permanent archive.

Flexibility will of course be required as versions of the most recent decade or two will be subjected to a greater frequency of added observations, and thus a versioning system will need to accommodate newer versus older portions of the archive. In addition, the constant inclusion of current observations as time goes on will be accommodated.

The project will also supply to the community datasets for testing algorithms for accuracy in the various aspects of spatial and temporal homogenization. These specialized datasets will also require versioning so that for publications (or litigation) an investigator will have a clear pathway defined to replicate the findings. It is possible that version control of electronic data and software can be handled through commercial off-the-shelf systems or open source systems such as Subversion (http://subversion.apache.org/).

- *Given the extent of this project and the unpredictable nature of the evolution of the archive, the reliance on an active panel to address version-control issues as they arise will be necessary. The panel will investigate the possibility of utilizing commercial off-the-shelf or open-source version control software for electronic files and software code (e.g. Subversion (http://subversion.apache.org/).*

- *Since one requirement of this project is to preserve older versions of the archive, and that a considerable amount of tedious research will be performed on any one version, it is generally assumed that up-versioning will be performed of the basic, Level 2 digital archive as sparingly as possible.*

- *The algorithms that produce the datasets used for testing and the datasets themselves must be documented and version-controlled.*

Configuration Management

The data system being proposed will be a dynamic system that satisfies a number of requirements in order to establish and maintain a consistent set of products. An example of a system designed to store and provide access to document images of the original paper records is EDADS operated by NOAA/NCDC. [http://www.ncdc.noaa.gov/oa/climate/cdmp/edads.html]. Configuration management includes the designation of format selection of accessible files for convenient public use. These are the Level 2 products that provide a uniform format structure for ready analysis and when integrated with other stations become the Level 3 databank. In most cases a fundamental station record format that includes the core information (station identifier, location, date, time (i.e. 0900) or category (i.e. TMax), temperature value, version, etc.) will be the most commonly accessed files and provide the greatest utility for the users. Certain relatively vital pieces of information may also be contained in this fundamental data record such as type of instrument, type of shelter, and length of record with consistent parameters (i.e. location, instrument, etc.) Associated with the Level 2 data will also be pathways back to the primary source data (Levels 0 and 1) and to the available metadata documents which explain the observation and establish traceability.

Though most surface stations are assigned a World Meteorological Organization identifying number, there are many for which this has not been done. For these stations the CM team will work with the WMO to assign new WMO-qualified identifiers. However, if it is deemed necessary, the team may construct a new system of identification that is highly expandable and contains within it an apparency feature which

allows an individual to easily recognize the region wherein the station resides. This could include the use of a FIPS country code and network identifier.

Many builders of large datasets have encountered the existence of duplicate datasets of a particular station whose values differ or data records whose values seem at odds with the climate of the identified station. If the Level 0 primary source evidence is available, the time series can be reconstituted and authenticated. If the primary source document or datafile has been lost, a pathway to a decision is required. In a parallel to "textual criticism" in literature, it is expected that investigators, in many cases external to this project, will need to sift through the information when multiple copies of a station record are discovered, determine which files or documents are most closely related to the missing primary source, and express (and document) an expert judgment as to the decision which is ultimately made. It is likely that in many cases, it will be necessary to archive multiple versions of a station record at Level 1. In some cases there may be greater confidence in some records than others. Confidence ratings and discussions may help in documenting the unique character of each record. This process will likely also occur in situations where a primary source paper form (Level 0) of daily observations may disagree with an official report of a monthly summary supposedly derived from the primary source. The CM functions will obviously be closely linked with the version control functions. Defining other requirements associated with hardware, security, support, and financial resources will be necessary as part of this project.

- *A configuration management board will be selected to initially define the necessary infrastructure, formats and other aspects of archive practices. A permanent board will then be selected to oversee the operation. This board and the version-control panel may be coincident or at least overlapping in membership.*

Summary

To provide primary-source surface temperature data for the research community through a suitable and convenient archive, and to meet new requirements now being demanded of climate observations (i.e. "admissible evidence" in the legal sense) a significant and on-going investment will be required. To the extent resources are limited, the utility of this project will also be limited in meeting the various requirements placed on data by the differing communities who now utilize climate observations. Now is the time to initiate this project (and support similar on-going projects) as institutional memory, documents and critical material are being lost with each passing day. To whatever extent this project is successful with data provenance, version control and configuration management, a century from now, our descendants will either thank us or criticize us.

Appendix A

Definitions

Data Provenance (DP) refers to the confirmation or gathering of evidence as to the time, place, and -- when appropriate -- the person responsible for the creation, production, or discovery of the data.

Version Control (**VC**, also known as revision control, source control, or software configuration management) is the management of changes to documents, programs, and other information stored as computer files.

Configuration Management (CM) is a field of management that focuses on establishing and maintaining consistency of a system's or product's performance and its functional and physical attributes with its requirements, design, and operational information throughout its life.