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.

Publication & Collation of Data Products and Presentation of Audit Trails

Song Lianchun[1] Zhou Zijiang[2] Peter Thorne[3] Kate Willett[4] Guo Jianxia[5] Karl Taylor[6]

1．National Climate Center， China
2．National Meteorological Information Center，China
3．NOAA's National Climatic Data Center，USA
4．Met Office Hadley Centre, Exeter, UK
5．Meteorological Observation Center of CMA，China
6．PCMDI, LLNL, CA, USA

Discussion Areas:

i． What an acceptable audit trail for dataset acceptance would be (intermediate steps, code, others) and what would be desirable beyond this minimum requirement;

ii．Whether datasets should be served from a common platform in a common format;

iii．How to enable intercomparison of the results between the datasets that will result;

iv．What will constitute acceptable publication criteria (peer-reviewed literature; other?);

v．How to ensure that datasets are correctly credited to their creators and related Intellectual Property Rights issues on these value added products;

vi．How to handle approaches that may be substantially different e.g. reanalyses.

Introduction：

The creation of a databank that holds the original data and associated metadata has been covered in earlier white papers. Here, discussion is restricted to the requirements of the derived data products that are assumed to have had some degree of quality control and / or homogeneity assessment performed. An example of a “data product” would be a gridded version of the original data (e.g., interpolated to a regular grid with infilling in regions of sparse data). These data products will be constructed by a diverse range of groups using independent approaches and considering different periods, regions and temporal resolutions (monthly, daily, observation-by-observation). It is necessary to consider what minimum requirements there are to ensure the scientific value of these analyses and to enable users to undertake meaningful intercomparisons. These requirements form the focus of this white paper. We would propose that only those datasets meeting such a minimum requirement set be considered as constituting an output of the overaching envisaged project

outlined across all white papers.

Effective publication, comprehensive analysis and objective collation of quality controlled and homogenized data and products derived thereof are the prerequisites for easy access to and effective use of data by the full range of envisaged users of climate services. Therefore, it is essential to define a controlled standard procedure for data production in order to assess the appropriateness of candidate data processing methodologies, and to ensure minimum standards with regards to data publications. To improve the credibility of the data products, when data sets and related products are released, both sources of original data, and information about data processing should be clarified, and the results of inter-comparisons with other data and products that fall into the same categories should also be provided to the extent possible. When data and related products are published, the data processing should be adequately described such that independent replication of the result is possible. For the sake of effective and efficient data applications by end users, there is also a need for simplicity and consistency and avoidance of the over-excessive and complex links to third party data sources or sources in multiple formats. Since the data processing techniques have already been addressed in details in other white papers, they will not be further dwelled on in this section.

Description of Data Source

The producer of a data product should provide all relevant metadata about the observational measurements used. This would normally include the version of the core databank used along with a full list of the stations used and the criteria by which they were chosen (e.g., based on the region of interest and specified quality control criteria). It is assumed that the primary databank contains all available information regarding the station location, instrumentation, exposure and changes over the period of record. Given that the original databank will be under version control, as long as this version is also recorded in the data product metadata and the selection criteria are thoroughly documented, the data selection steps will be replicable. An independent investigator should therefore be able to retrieve the same raw data.

Publication of methodology

The production of data products includes generation of quality controlled, homogeneous station data and unified gridded products. To meet the needs for the envisaged applications (e.g.regional climate change detection studies, monitoring, aiding planners, policy makers, transport and construction [finite list]), data and derived products in most cases need to have undergone homogeneity checks and adjustments. For a homogenized dataset to be acceptable, the expectation would be that at a minimum a description of the processing algorithm would already have been described in a paper accepted by a peer-reviewed publication.

Quality Assessment of Data Products

Previous white papers have discussed the construction of a common set of benchmarks. To be acceptable we would propose that the algorithms must have been run against at least an agreed minimum subset of these benchmarks subsampled in space and time exactly as the real world observations were by the data analyst and the results assessed. This metadata regarding performance, likely provided by an assessment group using a common set of assessment criteria

(see white paper 9 and 10) needs to be directly related to the dataset archive version. Additional metadata would be encouraged that further outlines performance against the benchmarks in addition to any high level performance summary undertaken by the assessment team. In addition metadata relating to assessment against alternative datasets should be published to enable users to easily ascertain whether the current dataset is right for their needs or better alternatives exist.

Presentation of Data Products

Data products from a given analysis effort may include both data and graphics. Data files should be generated in commonly understood formats such as ASCII characters or binary data (e.g. NETCDF), while a graphic file may be represented in GIF, JPEG and other formats. Where standards exist governing the recording of metadata (e.g., the Climate and Forecasting Conventions), files should follow those conventions. The data files should be named according to a unified standard, which should include such information as region, temporal & spatial resolution, producers, etc, should clearly reflect the content of data products, and should use common characters for adding a suffix to a given data file name to reflect its data format and version number.

Conceptually it would be of most use to end users if the datasets were delivered or at the very least mirrored in a common format through a common portal akin to the CMIP-5 portal. This would enable ease of intercomparison by end users. This portal may have additional capabilities such as an interactive interface enabling users to derive timeseries and maps for their given region or location from all datasets that produce an estimate.

Audit Trail Minimum Requirements

All derived data products hosted by the databank must be accompanied by a sufficiently detailed audit trail that they are reproducible. There is a strong argument for this to include code/software written to create the data product although this may not be possible in some cases (e.g. where the employer retains copyright on the employee's code) – perhaps a step by step list of what the code actually does as opposed to actual code will suffice here in such cases. The audit trail would include at a likely minimum:

- List of stations used and periods considered
- Quality control flag metadata
- A full listing of breakpoint locations and adjustment factors
- Ancillary datasets utilized
- Intermediate steps, particularly for homogenization procedures that are iterative in nature.

However, as a general principle as much metadata regarding the processing undertaken as is practical should be encouraged.

Platforms for dataset intercomparison

A master website should host a table that summarizes all available derived data products. The table will serve as an index to the products and provide brief descriptions of what they contain

(spatial coverage, temporal coverage, gridded or station etc.), along with a list of creators and a quality assessment stamp (quality controlled?, homogenised? Or some indicator of quality that implies these). There could also be a search engine to search by resolution, region etc. This should allow manual comparison by the user. There may be substantial value in providing basic comparison tools such as difference maps, trend/timeseries comparions and climatology comparisons visually to the user.

Acknowledgements requirements

When using the data and derived products, the users should clearly indicate the data sources and the references to data products in a prominent location of their applications. Appropriate acknowledgement statements can be provided on the website for users to copy into any publication. At a minimum any scientific paper should reference at least one paper describing the construction of each dataset used. Furthermore, if a common portal is pursued then the portal should be acknowledged in a similar way to that of the CMIP portal.

Recommendations

1. When releasing data products, we would recommend that the following information must be provided about process of product generation, apart from data itself for the product to be considered an output of the project:

(1) A listing of the source data (databank version, stations, period) along with methodological rationale

(2) A file describing the quality control method and quality-control metadata flags;

(3) Homogeneous and / or gridded version of the data;

(4) Quality assessment report produced by running against at least a minimum set of the common test cases described in previous white papers;

(5) A published paper based on the data construction method and related products in the peer reviewed press in a journal recognized by ISI;

(6) Publication of an audit trail describing all intermediate processing steps and with a strong preference to inclusion of the source code used..

2. Datasets should be served or at the very least mirrored in a common format through a common portal akin to the CMIP portal to improve their utility.

3. Utility tools should be considered that manipulate these data in ways that end users wish.