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.

Creating surface temperature datasets to meet 21st Century challenges

- 14: Solicitation of input from the community at large including non-climate fields and discussion of web presence

Authors:

Nigel Fox	NPL
Ian Joliffe	Univ of Exeter
Nick Barnes	Ravenbrook
Amy Luers	Google inc
Rob Allan	UK met office
Philip Brohan	UK met office
Michael De Podesta	NPL
Richard Chandler	UCL

Scope of document/task:

- Mechanisms to facilitate communication with all stakeholders including non-climate science community.
  - Out going - Awareness
  - Incoming – input/views particularly from those with relevant expertise
  - Leading to "buy-in"
- Use of internet tools and web presence
  - Test-bed for new methodologies: blogs, social networking sites, data visualisation …
  - Provision of data, tools and products that can be easily utilised by all stakeholders
- Maximising productive input and debate, not fuelling controversy and confusion.
  - Providing adequate background / education
  - Controlled but flexibly moderated debate
  - Responsive feedback
  - Access to data
- Need to identify desired stakeholder input

Introduction

Today there is an unprecedented demand for climate information by all sectors of society. Local, national and international organizations grappling with development and security issues are increasingly seeking information about how the climate is changing and what the impacts of these changes are likely to be, to inform how to respond to these risks. Adequately meeting society's growing need for trusted climate information will require a more open, transparent and participatory approach to climate change science than we have seen in the past— one that engages scientists, governments and the public.

In this age of widespread internet access, information is available and shared almost instantaneously on so many topics—through websites, facebook and twitter among other tools – yet, the scientific community has not yet adapted. If science is to effectively support society's

needs, the research and publication process must find its position in this new transparent information culture.

This global surface temperatures project is a critical high profile opportunity to define a new model of open, transparent and collaborative science and establish a precedent for this new approach.

The challenge – is how to do this in a manner that maintains the rigor of the scientific process? In this paper we explore how to integrate ICT tools to enable openness, transparency and collaboration in the three key components of the global temperature project: 1) Data Sharing; 2) Data integration and derivative products; 3) Data access and useability.

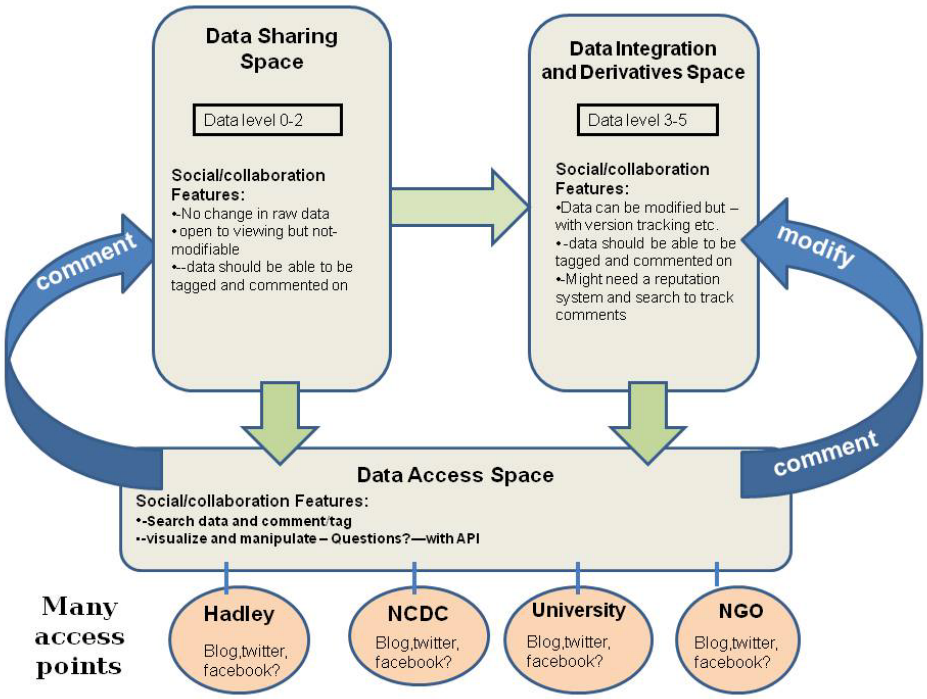

Figure1 highlights how we might integrate social and collaboration tools in each of these three areas. See White Paper 3 Figure 1 for elucidation of the data levels.

In considering the above scope it is essential that we take note of recent media and public interest/sensitivity to the topic of climate change and temperature data in particular. We need to be sure that we make every effort to provide full transparency of the activities in the project and the methodologies being used or associated with it. However, to minimise risk of further confusion from this process we will need to be sure that there is adequate public awareness of basic scientific principles, statistics and uncertainty statements, the latter both in terms of data and prediction.

The nature of this project also provides the opportunity, by making this a highly visible activity, to rebuild public confidence in climate science (and science as a whole) and facilitate "buy-in" to the need for a lower carbon future. Of course such visibility brings with it obvious risks.

The global nature of this initiative together with its high profile can take the advantage to experiment with new media/communication approaches which may also attract resources in their own right (e.g. open source dataset creation algorithms, novel applications and visualisation tools). The use of such tools particularly to visualise data or in communication, may also help to engage with the younger generation.
One of the most important issues that needs to be addressed early on is "what type of active input is being sought"?. Is it – ideas/agreement on methodologies for data collection, analysis etc, or something else. For the purpose of this document it is assumed that from a technical perspective the answer is largely as indicated i.e. data collection, analysis and visualisation. However, there is also the presumption that we seek input or at least feedback that the public as a whole understand the issues and importance of the activity and its consequences.

The other key aspect is the need for openness and transparency. Without the entire process being as open and transparent as feasible with exceptions to such an ethos clearly flagged and justified (e.g. employer asserts IPR to some aspect of a given effort) it will be hard to justify this project meeting these aims.

Stakeholder engagement for what and how?

There are three distinct groups that need to be considered when seeking an answer to this question:

i/ those involved in the community already or with technical expertise from another discipline or community e.g. statisticians, data processing etc but also others engaged in activities involving multiple-sampling, thermometry, metrology and probably many others. Input would be solicited on all aspects of the process - methodologies being proposed for data analysis, including any data screening criteria and visualisation as well as data collection and its quality assurance.

If all data were accessible then this would also allow any user to apply independent tools and analysis and report back to the project for subsequent discussion.

ii/ those likely to provide little technical input but who will be an active user of derivative products from these core data sets to make decisions of environmental, societal and economic importance.

iii/ those likely to provide little technical input e.g. General public, policy makers, but whom we need to be sure understand and have full buy-in to what is being done

Technically proficient/aware

In the first case, "technically proficient/aware", it will probably be adequate to make use of normal established communication methods, e.g. scientific journals, open meetings/workshops and a website. The latter would of course require some level of interactability and feedback probably with the ability for wiki/blog discussion to seeded questions and or unsolicited input. We should explore new collaborative dataspace models to support tagging and commenting whilst recognising that not all scientsists will be comfortable with such technologies (e.g. the blog for discussion has seen minimal use despite being flagged to numerous expert communities through multiple listservs).
However, as a starting point a few questions phrased to allow simple YES/NO responses

wherever possible, but with options for free flow comment may be simpler and less intimidating. For this to work effectively, in terms of data useage, it will be important to give clear examples and clarity on any analysis approach or methodology that is being proposed for discussion with real or simulated data to scope that expected in practise. Because it is important that this project not be seen to be owned by a single entity but rather be a truly international multi-partner effort we should actively consider a cloud based model taht is not centered on one website but can be accessed from several sites. This may yield an additional organisational overhead to achieve in total but would ensure that the overhead did not fall on a single institution which may be more sensible. For all practical purposes the website would become the principle focal point for interaction.

It would need to have sufficient resources to ensure it was maintained and also responsive to input including feedback to correspondents. Here, again, a distributed model may work better whereby the comments relevant to a given component of the project are managed by those with relevant expertise. This will of course also require some level of appropriate and consistent moderation. Any answers to questions would need to be automatically compiled into a summary table and any free-flow answers/comments copied into a blog-like format to allow further community discussion.

It would be appropriate to have some form of registration before any inputs are submittable to the site, however full visibility should be available without registration. Registration should not normally preclude anyone, but some reasonable filtering system may need to be employed to reduce poor quality, non-attributable, input e.g. correspondents should be happy to provide real contact details and use regular email addresses, and of course in many countries this would probably be subject to the data protection act. More important is that any input will have to meet certain norms: be on topic; non-inflammatory; avoid accusations of mal-practice etc. if we are to ensure the project retains focus.

Whilst dedicated workshops and meetings are likely to bring the most rapid and comprehensive interaction they are very costly in terms of organisation and therefore probably best arranged as a session as part of other already organised events/conferences which a critical mass of project participants are already planning to attend rather than as stand alone events. However, after some pre-defined timescale or if some significant controversy were to emerge then there may be justification to have an open event to ratify any decision. There may also be a need for some meetings of governance structures (which should be covered under the governance white paper and its discussions).

The scientific media, particularly those of a more general nature and wide circulation such as “new scientist” but also Nature should be used as “pointers” to the web-site. General awareness news articles rather than costly adverts being the preferred approach using appropriate press-release mechanisms to support this. If written correctly these would likely trigger broad media coverage and so the web presence would need to be ready for this and also appropriate speakers available to answer questions.

In terms of data access the most transparent but also challenging approach that could be followed is to provide full access to all data sets and support information for anyone to analyse in their own way. Whilst this is attractive it does bring with it some significant risks and issues not least the ability to obtain a uniform data access policy from all data suppliers and a uniform data format to permit easy intercomparisons. There may also be

issues regarding misunderstandings of aspects of data quality which could lead to confusing analysis being carried out and/or the need for time consuming discussions with individual data providers to clarify issues. However, this would be the ideal scenario. The application of any restrictions would be difficult. It is relatively easy to justify a limit on access to data providers or funders to the project but as soon as this is extended to try to include any one else, and not everyone it is hard to have what would be seen as fair criteria.

Non-technical specialist

Whilst it is unlikely that there will be detailed comment and input on specific methodologies being used or data quality from the non-technical specialists we should not ignore their importance. This project is likely to be of wide interest and provides an opportunity to increase awareness in climate change issues and further the public understanding of science in general. It is thus important to give them the same access to the web presence and opportunity to comment. Again, it would be beneficial to have a cloud-based presence with multiple "front ends" to the portal and redundant serving capacity There may be benefit to having a presence aimed at different levels for the technical expert, the non-technical end-user and the lay person (and possibly others). One such entry point should be set-up with an educational objective and have sufficient resources attached to take the technical data and distill this in an unbiased way into information taht is usable to a lay-audience. This copuld, for example, provide information and background on the key scientific practises that are likely to cause confusion, e.g. uncertainty and the need/methods for data screening/weighting etc. The use of good graphics and ideally interactive visualisation tools to support this would be expected by the public and organisations like Google are probably well placed to develop an appropriate learning and visualisation environment. Many related organisations; NOAA, NASA, ESA etc have similar educational sections of their websites and encourage the same for individual projects as well.

It may possible to propose small experimental projects to encourage engagement with schools and help them gain an understanding of some of the sampling issues, e.g. establishing their own small local networks of temperature monitoring stations and observing variations due to location, sampling strategies etc. They could also be provided with access to typical or real data for their own analysis, (sub-setted and stored independent of the primary data sets) and perhaps comparing results and approaches with other schools, again as a learning and awareness activity. Such educational activities are likely to be relatively low cost but could have major benefit in terms of public awareness and buy-in. These could also include efforts to digitise local data to augment the databank (see White Paper 3).

Data and metadata accessibility

In considering data accessibility, it is important to consider two things,

- Ease of access to the user
- Responsibility for data quality and access

Databank (see also white papers 3-6)

In considering data accessibility from a user perspective, particularly if the users are to be global and of varying levels of expertise, it is important that all data from whatever source is provided in a similar (ideally identical) format and with consistent and transparent

metadata and quality descriptors. This will ensure users can make reasonable decisions on data selection and useage. Since such datasets are expected to be used for climate studies it is important that they are stored in an invariant form for all future generations. Any modification of the data (even if for good reason) should be done and described as part of any processing activity to avoid potential confusion through use of different source data and version controlled. From a simplistic perspective a single (and of course mirrored) archive of data, accessible from a single portal would appear to be the easiest solution as the cost of data storage is relatively low. If the archive was truly an archive, where data was only added or accessed and not subject to any modification, then this may be the best solution and would ensure that any and all users would always have access to the same source data. Once submitted it would be fixed for all time. If a distributed data system were established, this would place severe requirements on data suppliers to freeze and maintain links, and guarantee not to update any data sets. There are opportunities to engage the community in the creation of this databank through crowd-sourcing of digitisation with suitable quality control (double keying from unique IPs with quality checks by full-time experts).

In both cases there is an assumption that data suppliers will allow free and open access to their data and take responsibility for ensuring it is correct at the time of submission. If any subsequent error is found or an update required the data-base will need to be able to tag the data set with some indicator that a change is required and that this should be stated in any future use of the data. It will of course be critical to have a common set of guidelines for data format, and metadata with an appropriate QA process to be demonstrated by data suppliers going forwards. For existing data, an appropriate screening and subsequent quality indicators should be created for each data set to ensure consistent analyses.

It seems reasonable that any user of the data archive should be required to register and agree to reasonable acknowledgements in any published useage. Data should be searchable and downloadable for off-line processing, ideally making use of graphical interfaces. However, there should also be on-line tools to visualise and manipulate data available. The nature of the data is not highly complex and so such tools should be readily available from many commercial suppliers who might be encouraged to volunteer to provide them as part of a promotional activity. This can be solicited in the early stages of the projects development and facilitated in the seeded questions section of the web front end.

Datasets (see also white paper 13)

Datasets should be made available through the cloud-based web presence along with their attendant metadata upon acceptance for publication. The metadata should include all intermediate processing steps, a description of the algorithm and its code. As has been proven by e.g. www.clearclimatecode.org there is substantial value to release of the code enabling a greater understanding of the processing by involvement of interested members of the public on a pro bono basis as well as scientific peers. However, a distinction needs to be made between the scientifically useful reproduction activity undertaken by e.g. clearclimatecode.org and the scientifically meaningless replication process of running the same code on my PC as you ran on yours which only proves I can replicate your mistakes on my PC. The scientific value in providing the process metadata accrues through reproduction and assessment of sensitivity to methodologically uncertain choices.

Datasets and metadata, including their quality assessment against the benchmarks should be made available in a common format to aid users. In the first instance it would make sense for the dataset, its metadata and associated web presence to be hosted by the dataset creators or a group nominated by them. They will be best placed to answer questions related to their particular product.

Data products

Once multiple groups have produced datasets for a given station / region / timescale then tools will be required that enable this data to be collated along with attendant metadata pertaining to dataset quality and additional flags. Then additional capabilities will be required to aid in interpretation. This may require the user to input what is of most interest to them: characterisation of extremes; seasonal variability; trend characterisation etc. as each dataset will have different capabilities and strengths and weaknesses. A suite of graphical capabilities will be required as well as a capability to tabulate the data in the way the user desires.

New web based communications

If we are to maximise impact and exposure of the project we should take advantage of all media/communication opportunities. This should include, at least for promotional purposes, tools such as Twitter, Facebook etc and also virtual worlds such as Second Life for those aspects deemed appropriate whilst recognising that not all project aspects will necessarily be suited to these tools. All of the above are being used by many organisations to promote awareness but also in some cases to solicit opinions and views. Whilst it is arguable that the use of these tools will not provide any new information or key stakeholders they do offer different opportunities to promote the project and its outputs through a secondary marketing approach, i.e. there will be communication opportunities from those interested primarily in the tool being used and where our project may be of secondary interest.

Organisations like Google might be able to help promote the project and provide significant support in visualisation. For example through adding additional layers or markers on for example “Google Earth”, identifying the location of all sample sites, together with the associated high resolution satellite imagery this could help identify sampling errors due to environmental effects as well as historical images for the location. These pointers could also be linked to the source data records and with some relatively simply additional tools could allow time series visualisation for any site and subsequently global or grouped visualisation of any data set in a simple intuitive manner.