Building trust in climate science: data products for the 21st century[†]

Richard E. Chandler[a*], Peter Thorne[b], Jay Lawrimore[c] and Kate Willett[d]

Climate science has a key role to play in informing strategies for adapting to and mitigating the effects of climate change. However, given the magnitude of the issues involved and their implications, it is imperative that the scientific process is–and is seen to be—rigorous, defensible, and transparent so as to ensure trust in the results. A key element in building such trust is to provide access to underlying data, so that interested parties can check published results and compare with their own analyses. A further priority is to provide data at the fine space and time scales that are relevant for user needs. Until recently, the ability to meet these requirements has been constrained by data-sharing agreements and limitations on digital storage and processing. This is now changing however, thanks to improved global collaboration, communication and computing capability. This article describes current efforts to exploit these opportunities via the International Surface Temperature Initiative, an international and multidisciplinary effort that aims: firstly, to create a single comprehensive global databank of surface meteorological observations at monthly, daily, and sub-daily resolutions; and secondly, to encourage the contribution of multiple independent data products, subject to common performance assessment and benchmarking criteria, thus providing the opportunity for a detailed assessment of uncertainties. The rationale for the initiative is discussed, along with logistical and technical challenges, as well as opportunities for involvement from the statistical and wider scientific and user communities. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: benchmarking; climate; governance; public trust in science; transparency

1. INTRODUCTION

Many lines of evidence suggest that the earth's climate has changed rapidly since the mid-20th century (Trenberth *et al.*, 2007). This is of particular concern to society and policymakers and, as such, climate change adaptation and mitigation strategies form a core part of the political and social landscape at both national and international levels. Climate science has a key role to play in informing such strategies. However, given the magnitude of the issues involved and their implications, it is imperative that the scientific process is—and is perceived to be —rigorous, defensible, and transparent so as to ensure trust, promote understanding, and encourage critical appraisal of the results by scientists and citizen scientists alike.

   The need for trust and transparency was well illustrated following the unauthorized release, in November 2009, of approximately 1000 e-mails from the Climatic Research Unit (CRU) of the University of East Anglia in the UK. According to the subsequent report of the Independent Climate Change E-mails Review (Russell *et al.* 2010), e-mails fueled challenges to the work of CRU, to the reliability of climate science generally, and to the conclusions of the Intergovernmental Panel on Climate Change. The incident received extensive media coverage, and subsequent polls in the USA and UK suggested that it contributed to a decline in public trust in climate science and in scientists more generally (Jasanoff, 2010; Tollefson, 2010; Maibach *et al.*, 2011). Although the Independent Climate Change E-mails Review and several other reviews exonerated the CRU scientists from any malpractice in data processing and analysis, recommendations emphasized the need for increased openness, transparency, and access to underlying data (Russell *et al.*, 2010, p.15). In fact, several long-standing efforts in this direction were already in existence, such as the Global Historical Climatology Network (Peterson and Vose, 1997) and International Comprehensive Ocean-Atmosphere Data Set (icoads.noaa.gov). Nonetheless, many commentators saw the incident as fundamentally changing the way in which climate scientists, and the scientific community more generally, must operate in order to maintain credibility (see, for example, Horton 2010; Jasanoff 2010).

* Correspondence to: Richard E. ChandlerDepartment of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K. E-mail: richard@stats.ucl.ac.uk

a   Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, U.K.

b   Cooperative Institute for Climate and Satellites, North Carolina State University and NOAA's National Climatic Data Center, 151 Patton Avenue, Asheville, NC, U.S.A.

c   NOAA's National Climatic Data Center, 151 Patton Avenue, Asheville, NC, U.S.A.

d   Met Office Hadley Centre, FitzRoy Road, Exeter, Devon EX1 3PB, U.K.

Maintaining credibility is only one potential benefit of providing open access to climate data. Societal infrastructure is vulnerable to a range of climate and weather-related events such as floods, droughts, windstorms, and heatwaves; and plans for the infrastructure of the future must be informed by the best possible evidence. In many situations, climate data are required at fine time and space scales to provide decision-relevant assessments of risk and infrastructure performance; however, fine-resolution data have often been difficult to access.

Furthermore, it is essential to acknowledge the inherent uncertainties in any quest for robust climate information. One way to explore uncertainties and to establish confidence in the results is to develop multiple methodologically independent estimates of quantities of interest. The provision of a framework that includes comprehensive and traceable open access data, intercomparison tools and product platforms, will facilitate the development of such products.

Against this background, in February 2010 the UK Meteorological Office produced a submission to the World Meteorological Organization (WMO) Commission for Climatology, calling for the creation of a new suite of data products to address the issues outlined earlier. Following WMO approval of this submission, a meeting was held in Exeter, UK in September 2010, with a view to establishing an open access global data bank of land surface temperatures and associated climate products that have been subject to rigorous assessment procedures. The meeting was attended by 80 international experts from a wide variety of disciplines including climate science, metrology, statistics and software engineering, as well as journalists and citizen scientists: white papers were posted online in advance, and public comments solicited. The meeting outcomes are summarized in Thorne *et al.* (2011b). In particular, the meeting marked the inception of the International Surface Temperature Initiative, henceforth "the Initiative" (see http://www.surfacetemperatures.org).

The aim of the Initiative is to create, for the first time, a single comprehensive global databank of land surface meteorological observations at monthly, daily, and sub-daily resolutions; to solicit creation of multiple products from this; to assess performance of these products; and to serve those products to end-users. The scope is thus much broader than the creation of a global data set, for example, along the lines currently being developed by the Berkeley Earth project (berkeleyearth.org). Uniquely, the Initiative will provide access to extended archives of raw observations along with metadata and provenance tracking facilities; and, rather than creating data products itself, it will provide resources for the development, assessment, and hosting of data products contributed by the international community. It is run by an international and multidisciplinary team and is officially endorsed by the WMO and The International Environmetric Society, as professional bodies representing the climate and statistics communities, respectively. Endorsement from the Bureau International de Poids et Mésures, representing the metrology community, is also in process.

This paper provides a brief introduction to the Initiative summarizing progress to date and setting out the road ahead as currently envisaged. As such, it is, in some sense, an update to Thorne *et al.* (2011b). As a contribution to this *Environmetrics* volume however, the main aim is to highlight the statistical issues involved as well as the opportunities for the statistical community to contribute. The next section summarizes some of the difficulties and challenges associated with climate data provision, Section 3 sets out the main tasks for the initiative, and Section 4 concludes.

2. CLIMATE DATA PROVISION: DIFFICULTIES AND CHALLENGES

As noted earlier, increased data access and transparency are a key priority for the climate community. Whereas its importance has long been recognized, only now in the age of global Internet and massive digital storage is it even starting to be possible. In this section, we briefly review some common barriers to open data access, and also set out some of the issues underlying the call for transparency.

2.1. Barriers to open data access

In many parts of the world, weather observations have been made routinely since the 19th century and before; and many of these records are freely available via initiatives such as the Global Historical Climatology Network (see earlier). However, although most national meteorological services have signed up to a WMO commitment (http://www.wmo.int/pages/about/Resolution40_en.html) on the exchange of data, attitudes to this commitment vary from one country to another. In reality, most countries interpret it as pertaining only to real-time data exchange. Some meteorological services are prepared to release essentially all of their data without restriction, others to restrict their usage (for example, to non-commercial purposes), and others still to release some subset of the data (e.g., from a small number of stations, or at a monthly but not a daily time scale). There are several reasons for this. In some countries, data are seen as having commercial value: to make them freely available would therefore close down a key income stream for the meteorological service. Elsewhere, there may be geopolitical reasons underlying a country's reluctance to share information with its neighbors. Finally, access to recent data from some countries is currently effectively impossible because the necessary resources and facilities simply are not available. A key challenge for the Initiative will be to engage with national meteorological services, via their representatives in the WMO, so as to understand better their individual requirements and concerns over the release of data. A further challenge will be to identify other organizations that may possess valuable data holdings: organizations responsible for water supply and sanitation, for example, often run their own independent networks of rain gages and meteorological stations.

Even where data are nominally available, access in a form suitable for quantitative analysis (as opposed to, say, hard copy records, images or graphs) can be difficult for non-specialists. Some user-friendly interfaces exist: examples include the KNMI Climate Explorer (climexp.knmi.nl), the IRI/LDEO climate data library (iridl.ldeo.columbia.edu/) and the U.S. National Climatic Data Center (http://www.ncdc.noaa.gov/). Elsewhere, however, to extract the information required can be a major logistical exercise owing to both the format of the supplied data files and to opaque and inconsistent documentation. As an example: one of the authors is currently involved in research requiring the use of data on multiple weather variables for two moderately sized river catchments, obtained from a national archive that provides data free of charge for (most) research purposes. Although the archive has a relatively intuitive "data extractor" tool, this has limited capabilities and could not be used here; the data therefore had to be extracted manually. The data were supplied in separate

files for each year: each file contained data for all stations operational during that year in the country concerned. To obtain the long records required, it was therefore necessary to download many yearly files: this was slow because the files contained information for all stations in the country whereas only a small fraction of these were needed. Software then had to be written to extract and combine the data for the stations of interest, taking into account data quality flags and the possibility that data might not be present for stations that were nominally operational. Finally, decisions had to be taken regarding the handling of the duplicate (and conflicting) records that were present for many stations. In total, the production of this data set required several person-weeks of researcher effort, interspersed with days of computer time, over a 6-month period. This is not unusual in careful scientific work, although it is clearly an inefficient use of resources given that many of the difficulties could be avoided by a more user-friendly form of data delivery. A further challenge for the Initiative will therefore be to bridge the divide between providers and users of climate data, for example, by providing a user interface in which issues such as duplicate merging and provenance tracking are easily handled, and in which flexible search criteria are available for users to locate and obtain data.

2.2. The need for transparency

A novice user could be forgiven for thinking that once the data access problem has been solved, all that remains is to provide a web-based portal that allows users to locate and download whatever data they require via an intuitive interface. Unfortunately, however, the realities of climate data dictate otherwise. In this short article, we cannot hope to describe all of the reasons for this: instead, we provide some examples that indicate some of the difficulties involved.

One obvious difficulty, mentioned in the anecdote earlier, is the presence of duplicate and conflicting records. One mechanism by which this can occur is that the original weather station readings are used by multiple "official" organizations, each of which carries out its own averaging and quality control procedures and might make adjustments (documented or otherwise) to the data as supplied to correct for non-climate influences (e.g., instrument changes): the duplication then occurs when these organizations are subsequently asked to release their data holdings back to the national or international archive. Even more fundamentally, different organizations may assign their own station identifiers (and names!) and may store location information to different precisions, so that the existence of duplicate records is not always obvious and requires careful checking.

Another difficulty is that the potential for errors and inconsistencies in climate observations is high: careful quality control and appropriate treatment of inconsistencies are therefore essential to ensure the credibility of results from any observation-based analysis. Sources of errors and inconsistencies include instrumental error, instrument failure or degradation over time, observer (i.e., human) error, change in measurement technique or instrumentation, relocation of an observing station, human or animal interference with the measurement device, and gradual changes in the environment surrounding the observing station (for example, the growth of trees, construction of new buildings, or urban expansion). Many of these issues are often undocumented in data records; careful detective work is therefore required to identify them. For a review of the kinds of problems that can occur and an indication of some of the procedures that can be used to detect and correct for them, see Baker *et al.* (1994); Peterson and Easterling (1994); Durre *et al.* (2008, 2010); Menne and Williams (2009). Clearly, however, there are uncertainties associated with the application of any such procedure: spurious inhomogeneities may be detected, genuine problems may be missed, and estimated corrections are subject to sampling error (see Thorne *et al.* 2005, 2011a; Kennedy *et al.* 2011a, 2011b for example). In some instances—for example, where times of measurement regime changes are known—and for some applications, it may be possible to adjust for the known inconsistencies directly in an analysis of the raw data (see, for example, Yang *et al.* 2006 for an example involving a change in the measurement resolution of precipitation data, and Ambrosino *et al.* 2011 for an example involving the introduction of satellite observations into reanalysis data). An advantage of such an approach is that the uncertainties are automatically propagated through to the conclusions of the final analysis. However, this is not always possible, and there will always be a demand for correction procedures that attempt to retrieve the true climate signal from the raw data. Of course, different procedures have different characteristics, so that the choice of procedure is itself a source of uncertainty (Thorne *et al.*, 2005). To understand this uncertainty better, it would be helpful to assemble an archive of multiple independent efforts based on a variety of different algorithms and modeling approaches.

Menne *et al.* (2009) provide an example illustrating some of the issues discussed earlier: this relates to the time series of annual mean minimum temperatures (i.e., the mean of the daily minima) recorded at Reno, Nevada, USA from 1901 to 2004. Figure 1 shows these data, obtained from the U.S. Historical Climatology Network (USHCN) database (cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html). A naïve inspection suggests, among other things, that there was an abrupt cooling in this location at the end of the 1930s, and that temperatures started to increase gradually from about 1975 onwards. However, the first of these features is, in fact, associated with a relocation of the observing station from a city center rooftop location to the (at the time rural) airport–a not uncommon occurrence with the advent of mass aviation. The second is probably associated with urban expansion: the airport site was initially well outside the city of Reno, but as the city expanded so the station is likely to have been increasingly affected by the associated "urban heat island".

Of course, if interest lies solely in the climate at station locations, effects such as those of urban heat islands are unproblematic: the post-1975 warming indicated in Figure 1 is a genuine reflection of the changing environment around the station. However, one often wishes to extrapolate from a station network to a wider spatial region, and, in this case, differential effects such as those of urban heat islands (which are particularly common because measurements are made preferentially where people live so that urban and suburban locations tend to be oversampled) can introduce biases unless appropriate adjustments are made (see Trenberth *et al.* 2007, Section 3.2.2.2 for a discussion). A further source of potential bias is that observing station networks evolve over time: for example, high-latitude station densities have decreased in recent years so that changes in these latitudes are less well sampled now than, say, during the 1970s. If rates of systematic change are latitude-dependent (and evidence suggests that they are—see Simmons *et al.* 2010)—then this is potentially problematic, although an appropriate assessment of uncertainty should account for the difficulty.

The potential for bias when characterizing trends in aspects of regional or global climate is one of the most oft-cited sources of controversy associated with the use of climate observations (see, for example, McKitrick and Michaels, 2007; Pielke *et al.*, 2007; Schmidt, 2009; Parker,
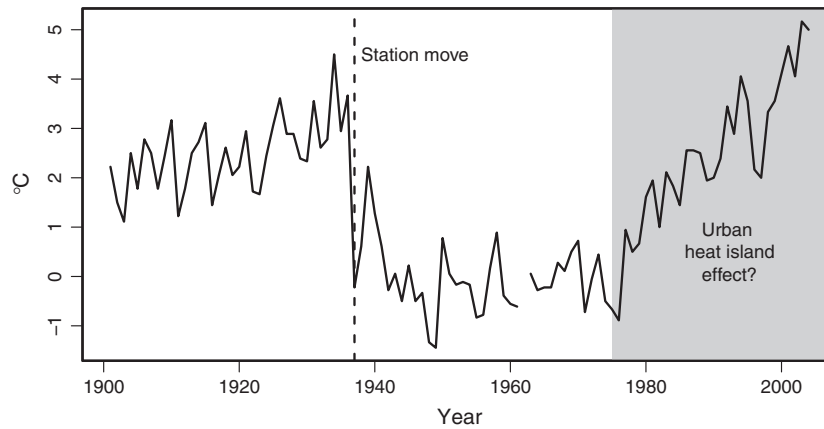
Figure 1. Annual mean minimum temperatures recorded at Reno, Nevada, in 1901–2004 (Station ID 266779 in USHCN version 2). Vertical dashed line indicates time of a known station relocation; shading indicates period of possible warming caused by urban expansion. The value for 1962 is not plotted because too many of the contributing daily observations failed quality checks during August of this year

2010, Fall *et al.*, 2011). From a statistical perspective, the issue is one of potential confounding variables in an observational study where time plays the role of the "exposure" variable: to eliminate the associated bias, it is necessary to adjust appropriately for the confounders (see Hansen *et al.* 2010; Jones and Moberg 2003; Menne and Williams 2009; Smith *et al.* 2008 for examples of this). For many potential users of climate data, however, the need to carry out such adjustments is yet another barrier: the use of derived data products, in which the available observations have been interpolated onto a regular grid with (hopefully) appropriate adjustment for confounders, has therefore become routine practice in climate science. Figure 2 provides an example: the stations shown in Figure 2(a) form the basis for the widely used monthly CRUTEM3 temperature product (Brohan *et al.*, 2006), and Figure 2(b) shows a specimen field of temperature anomalies (i.e., deviations from the 1961–1990 mean field) from the data product. The gridded structure of the product is clearly visible here: the grid spacing is 5 degrees in both the latitude and longitude directions. Also clear is that the link from observations at the stations shown in Figure 2(a) (not all of which were operational throughout the period covered by the data product) to the gridded data is indirect and spatially varying: some regions are very well sampled whereas others (notably the high-latitude regions, along with the equatorial regions of Africa and South America) have very few data points. Indeed, in some regions, the CRUTEM3 product contains missing values because there are no observations in the corresponding grid squares.

For the purposes of estimating trends in global or hemispheric temperatures—which are often cited as "headline" indicators of climatic change—the undersampling of equatorial and high-latitude locations is potentially contentious: certainly, the scarcity of stations near either of the poles is striking in Figure 2(a). The impression is exaggerated, however, by the use of a rectangular mapping projection: Figures 2(c) and (d) show the same information as 2(a) and (b) but using an equal-area projection, which gives a more accurate impression of the extent of undersampling at different latitudes. Moreover, Figure 2(d) appears overall to be lighter in color than Figure 2(b). This is illusory however: both images use identical grey scales, but the distorted projection in Figure 2(b) gives undue prominence to regions with darker (i.e., warmer) anomalies for this particular month and hence gives the appearance that temperatures were warmer overall than was really the case. This illustrates that relatively simple considerations—here, the choice of an appropriate map projection—can be helpful in communicating the issues to be considered when interpreting any given set of climate data, and can also affect the visual perception of structures in the data.

From the discussion earlier, it should be clear that "raw" climate data must typically go through extensive processing prior to any serious analysis. Furthermore, much of this processing requires decisions to be taken and judgments to be made: for example, to decide whether an apparent shift in the level of a temperature time series represents an undocumented change in some aspect of the measurement process, or whether it represents a genuine abrupt change in the local climate; or to decide on exactly what statistical procedure will be used to create a gridded data set from a spatially irregular network of observations (for example, whether to use neighbor-based algorithms or procedures based on spatial or spatiotemporal statistical models of varying degrees of complexity). The importance of these decisions cannot be overemphasized. Often, the choice of methodology for gridding or for calculating temperature anomalies can have as much impact on the final result as the application of bias corrections (Lawrimore *et al.*, 2011; Vose *et al.*, 2005). Naturally therefore, these decisions and judgments will affect the eventual outcome: from a statistical perspective, this is a fundamental consequence of the uncertainties associated with imperfect sampling and measurement. Thus, once again, it will be useful to assemble an archive of multiple independent—and methodologically distinct—products to better sample the solution space.

The need for this is highlighted by difficulties that are clearly experienced by non-scientists in understanding the necessity for, and validity of, adjustments. For example, a 2010 report from the Office of the Inspector General of the U.S. Department of Commerce (Zinser, 2010) is entirely devoted to addressing concerns expressed by U.S. Congress members, regarding the quality of USHCN data and the validity of adjustments made in deriving these data. Such concerns are understandable from the perspective of individuals who are unfamiliar with the difficulties associated with the interpretation of climatic data. However, as noted by Russell *et. al* (2010, p.15) and others, the "blogosphere"—whereas bringing undoubted benefits by providing the opportunity for citizen scientists to engage with and contribute to the scientific process—now provides a vehicle by which views that are not always constructive or fully informed can be promoted without moderation. This represents a considerable challenge to a scientific community whose culture has until now been firmly based on the
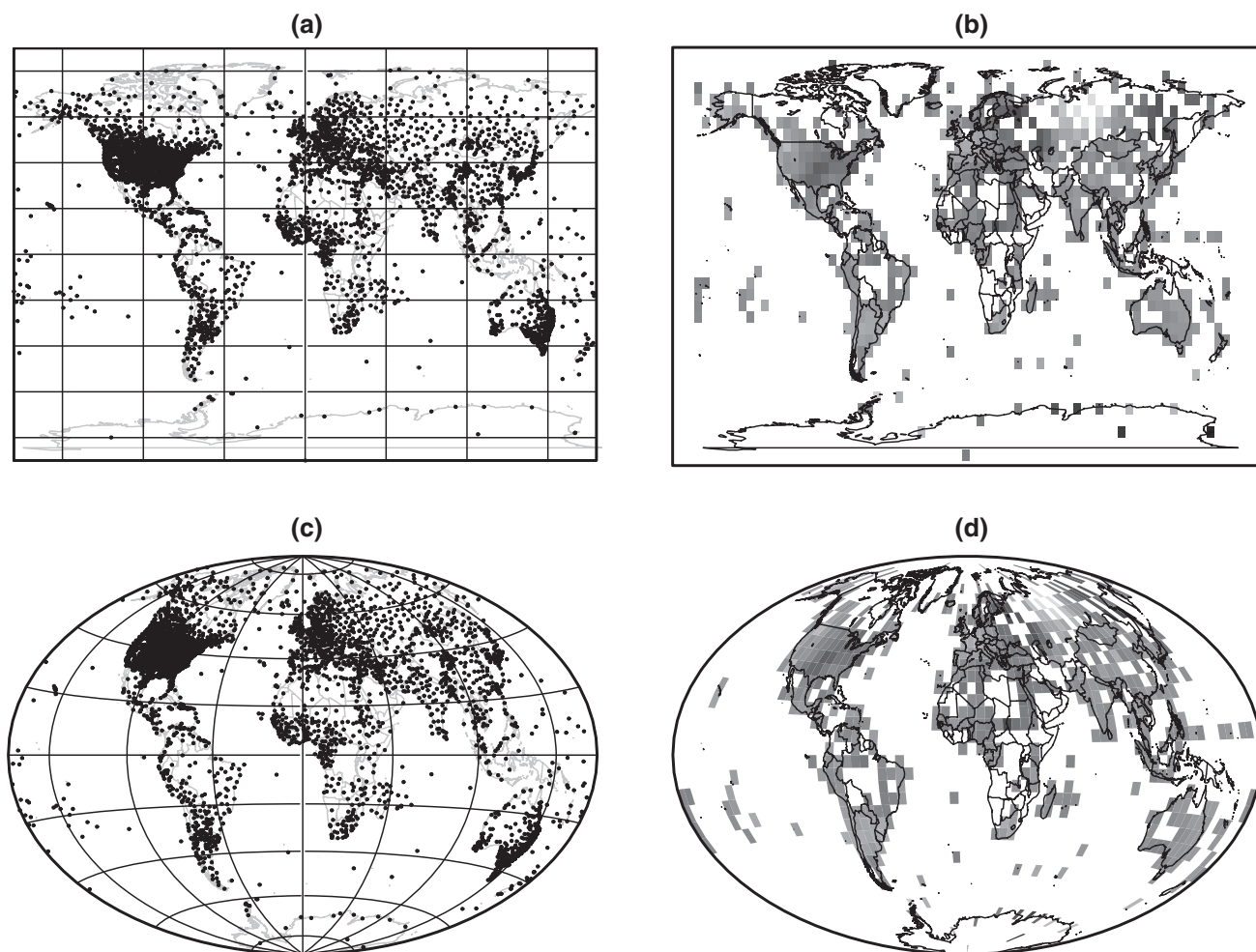
(a)

(b)

(c)

(d)

Figure 2. Temperature station networks and derived products: (a) stations used for construction of CRUTEM3 gridded data set (Brohan *et al.*, 2006), (b) gridded field of mean temperature anomalies for July 2011 (darker grid cells indicate warmer anomalies; unshaded grid cells indicate missing data). Plots (c) and (d) show the same information as (a) and (b) but using an equal-area projection

notion of peer review as the benchmark of acceptability. To respond effectively to this challenge, climate scientists must provide as much information as possible on the provenance of the data upon which they rely, as well as on the performance of any preprocessing algorithms that have been applied.

3. MAIN TASKS FOR THE INITIATIVE

To address the issues outlined earlier, the work of the Initiative has been divided into three main tasks. The first of these is data rescue and databank creation; the second is the creation of a program of common benchmarking and assessment of contributed data products; and the third relates to the provision of a user-friendly interface for accessing and visualizing the raw data and derived products. Working groups are responsible for each of these tasks, with overall coordination by an international steering committee. The governance of the Initiative involves oversight from the WMO, The International Environmetric Society, and (hopefully) Bureau International de Poids et Mésures as described previously, and includes defined reporting requirements and checkpoints. To date, most of the effort has been focused on the creation of an initial databank and on the development of benchmarking and testing procedures for contributed data products: we briefly review these processes next. Work on the development of a user interface cannot start in earnest until some data and products are in place: version 1 of the databank is scheduled for release during 2012.

3.1. Data rescue and databank creation

Within "the Initiative," the transition from initial data record to final databank entry is considered to take place in several stages. Stage 0 represents a hard copy or digital image of a written record; Stage 1 represents a digitized version; Stage 2 represents a digitized version following a standard formatting and data provenance convention; and Stage 3 represents a "merged" version of the Stage 2 data containing a unique record for each station.

A key task will be to identify as many Stage 0 data sources as possible, and to tap into the huge potential that these represent. Many sources have already been identified. In the USA alone, for example, there are more than 50 million digital images in the US National Oceanic and Atmospheric Administration Foreign Data Library (docs.lib.noaa.gov/rescue/data_rescue_home.html), and more than two thousand boxes of written data records in the National Climatic Data Center archives. Similar holdings exist in other libraries and repositories, particularly those of former colonial powers, and paper-based archives are literally rotting away or seen as a nuisance in many countries.

The digitization of Stage 0 data sources clearly will be a huge effort, and innovative solutions are needed. One possibility being considered is that of "crowdsourcing", whereby images are made available to interested members of the public for transcription and uploading to the Initiative databank in a similar manner to the "Old Weather" project (http://www.oldweather.org). More generally, the Initiative will work in partnership with existing efforts that have similar aims, such as the "Atmospheric Circulation Reconstructions over the Earth" project (www.met-acre.org - see Allan *et al.* 2011), the International Environmental Data Rescue Organisation (http://www.iedro.org), and other national and international programs. There is also an opportunity here for the use of statistical network design methods, to identify those spatial locations and time periods for which additional data would be most beneficial and hence to provide a means of prioritizing Stage 0 sources for digitization.

In addition to Stage 0 data, there are of course numerous sources of existing Stage 1 data that have already been digitized. Many of these data are held in regional and national archives. Although access is sometimes restricted for commercial or geopolitical reasons as discussed in Section 2.1 earlier, there is an expectation that, by working with national meteorological services, barriers to free and open exchange can be removed (Thorne *et al.*, 2011b). Moreover, digital formats vary from one source to another: it will therefore be necessary to reconcile these into a common format (Stage 2), prior to the final stage of merging sources to obtain a unique record for each station (Stage 3). The common data format will contain information on data provenance, in order that every data value can be traced back through the earlier stages to the original record from which it was derived. Moreover, metadata (i.e., information about the data—such as details of the measuring equipment used along with photographs or descriptions of the observing station, which can be valuable aids to the identification and confirmation of non-climatic influences on the observations) will be provided where available. The Stage 3 data are likely to form the basis for the development of most data products.

Because of the large volume of data involved, the transition from Stage 2 to Stage 3 will need to be (at least semi-)automated: at the time of writing (September 2011), however, the protocol for merging duplicate records is still to be finalized. A potential concern is that the results obtained from Stage 3 data may be sensitive to the specification of this protocol; however, Stage 2 data will also be available via the data portal so that concerned users can carry out their own analyses to assess the extent of such sensitivities. Moreover, no other data processing will be carried out up to this stage: the aim is merely to collate and document the available data as completely as possible.

3.2. Data products: benchmarking and assessment

As described earlier, climate observations often contain errors, inconsistencies, and inhomogeneities so that careful checking and preprocessing is required. An "expert" user may be content to work directly with the data from Stages 2 or 3 of the databank, carrying out diagnostics and checks prior to subsequent analysis. However, users who are less experienced, or with less resources available, will often prefer to work with data products in which the Stage 2 or 3 data have been "cleaned" and, perhaps, interpolated onto a regular grid as with the CRUTEM3 example from Figure 2.

As with the merging of Stage 2 data to arrive at Stage 3, data product creation will require automated procedures owing to the volumes of data involved. Moreover, multiple decisions will be required: which stations to select, what period(s) of record to use from each station, the temporal and spatial resolution of the final product, choices relating to quality control, and possible removal of inhomogeneities from the original records, the choice of interpolation methodology when moving from an irregular set of spatial locations to a grid, and so forth. Many of these issues are fundamentally statistical: the statistical community therefore is encouraged to play a key role here. A particular challenge is to provide appropriate, credible, and user-relevant assessments of the uncertainty in any product: this is an area where there is huge potential for the application of modern statistical methods. For example, whereas the CRUTEM3 product contains missing values in grid squares containing no data, an appropriate application of geostatistical techniques would provide the opportunity both to estimate the temperatures in these grid squares and to quantify the associated uncertainty. The computational burden of such techniques at a global scale can be prohibitive, however, and methodological developments, such as those of Lindgren *et al.* (2011), are needed to overcome this.

Given the many different decisions that could be made in deriving a data product, it is clear that one could envisage many different products being created from the same underlying data. As discussed in Section 2.2, this is an inevitable consequence of imperfect sampling and measurement. The Initiative encourages all data product creators to upload their (appropriately documented—see Thorne *et al.* 2011b) products to the data portal, with strong encouragement to participate in the benchmarking and assessment program discussed next. The availability of multiple products will provide further information on key uncertainties, and hence is potentially of fundamental scientific value. Of course, given the inadequacies of data documentation and the impossibility of sampling fully every possible source of uncertainty, it will never be possible to quantify uncertainty completely. Nonetheless, confidence in conclusions would be increased if, for example, they could be shown to be insensitive to a wide range of plausible but distinct approaches to data product creation. An additional benefit is that all approaches have strengths and weaknesses so that no single product can serve all user needs: by making multiple products available, and providing documentation along with information on situations for which each one may be more or less suitable, users will be able to make informed choices as to which products may be more or less appropriate for them.

If a user is to have confidence in the appropriateness of a data product for use in a specific application, it is necessary to provide some independent verification of its fitness for purpose. Of course, it is impossible to assess the accuracy of any product directly: to do so requires knowledge of the true evolution of the climate system that the product purports to represent (and if this were known, there would be no need for the product in the first place). However, some independent verification can be provided by focusing not upon the accuracy of the product

itself, but rather upon the performance of the algorithm that was used to create it. The Initiative is therefore developing synthetic test cases that will be used for blind testing of algorithm performance: test results will be made available alongside products on the data portal, so that users can see for themselves how the underlying algorithms perform. This will build on the work of the European Union COST HOME Action on developing and testing homogenization algorithms (http://www.homogenisation.org—see Venema *et al.* 2012) and pioneering efforts of Williams *et al.* (2012).

The basis for this benchmarking exercise will be the creation of around 10 different "analog worlds," each generated from a climate model to produce a complete and known realization of an "earth-like" climate system. The temperature records from these analog worlds will then be sampled using an irregular network of observing stations that mimics the network present in the Stage 3 data bank, and random and systematic errors will be added to the sampled records to approximate the kinds of error structures that may exist in reality. Data product providers will then be invited to run the resulting corrupted data sets through their algorithms to try and recover the original analog data: crucially, this exercise will be "blind" so that providers are not able to tune their algorithms to try and reproduce a known answer. Performance will then be assessed on the basis of measures that reflect an algorithm's ability to recover the original data: statistical input will be required here, too, to ensure that the performance measures are appropriate to the questions of interest. The first round of analogs is planned for late 2012, based on the version 1 databank release.

4. SUMMARY

The work described here represents a genuinely new approach to the provision of global land surface climate data; and, although the focus is solely on temperature for the time being, it is anticipated that other variables such as precipitation, humidity, and wind speed will eventually be incorporated as well. The main innovations include the truly international and multidisciplinary nature of the effort, with contributions from disciplines such as statistics, metrology, and software engineering in addition to climate science; and the emphasis on transparency, engagement, and participation from a wide range of interested parties, including end users and citizen, as well as "professional" scientists. For the predominantly statistical *Environmetrics* readership, the main opportunities for participation include: helping to locate and prioritize raw data sources in Stages 0 and 1 (those working in national statistical services may be able to contribute here), development of methodologies for data cleaning and data product creation, development of methods to create realistic benchmarks and assess algorithm performance, participation in the data product benchmarking exercise, and the provision of constructive feedback on the Initiative as it develops. Interested parties can make contact either via the corresponding author of this paper, or via the Initiative web site at http://www.surfacetemperatures.org.

It should be clear that the task facing the Initiative is challenging. However, international data repositories exist for other types of data, for example, the International Comprehensive Ocean-Atmosphere Data Set (see Section 1) for surface ocean measurements. By liaising closely with such activities, the Initiative will be able capitalize on lessons learned and experience gained. Nonetheless, the ability of the Initiative to provide a useful characterization of uncertainties in the historical record will depend critically upon contributions from a sufficient number of analysis teams to explore the full range of possible analysis decisions: this is an area where the statistical community has particular expertise and where, hopefully, many statisticians can be encouraged to contribute.

As indicated in Section 3 earlier, the development of a user interface to the Initiative data portal has not yet started because this task must follow from the databank and data product development. At the time of writing, several unresolved issues and technical challenges remain. These include the precise formats for the data, the degree of user interaction that will be possible, for example, the amount of graphical and tabular output that can be created "on the fly", and—which is likely to be of particular interest to statistical community—how effectively to communicate information about uncertainties in data products. Such uncertainties are an inevitable consequence of measurement error and sampling variation, and modern statistical methods coupled with advances in computing power are increasingly able to provide realistic representations of uncertainty when synthesizing evidence from multiple sources (the statistical literature on "data fusion" or "data assimilation" is particularly relevant in this context—see, for example, Fuentes and Raftery 2005; Sahu *et al.* 2009). From the perspective of stimulating public trust, it is important that scientists acknowledge clearly the uncertainties in their data and in the conclusions drawn from them (Russell *et al.* 2010, Section 5.6; Horton 2010). To some extent, this can be achieved by providing "error bars" on data products, derived from a realistic and credible statistical representation of the error structures and sampling mechanisms in the observations. For data users, however, handling uncertainty is less straightforward. Although there are efforts to communicate uncertainty in the analysis of global temperature (e.g., the monthly State of the Climate reports from http://www.ncdc.noaa.gov/sotc/global/2011/8#temp), it is probably fair to say that, at present, most data users prefer to ignore uncertainties attached to data products, for the simple reason that they do not know what to do with them (Rougier *et al.*, 2010). However, ignoring uncertainty leads to overconfidence in any subsequent conclusions: this is neither good scientific practice nor helpful to decision makers who need to be fully informed about what is and is not known. It is a major challenge for data providers to provide uncertainty assessments in a form that is appropriate for end users, and also a challenge for users to become comfortable with the idea of integrating input data uncertainty assessments into their work. One possibility that emerged from discussions at the Exeter meeting in 2010 (see Section 1), with strong support from the statisticians present, was that data providers should not attempt to provide a single "best" reconstruction, but rather a set of realizations from the joint distribution of the (unknown) quantities of interest given the available observations (see, for example, Kennedy *et al.* 2011a). This would force end users to confront the issue of uncertainty head-on; but it would require a substantial shift in the way that data products are perceived. Alternative ideas and suggestions are welcome!

REFERENCES

Allan R, Brohan P, Compo GP, Stone R, Luterbacher J, Brönnimann S. 2011. The international Atmospheric Circulation Reconstructions over the Earth (ACRE) initiative. *Bulletin of the American Meteorological Society* **92**: 1421–1425.

Ambrosino C, Chandler RE, Todd MC. 2011. Southern African monthly rainfall variability: an analysis based on generalized linear models. *Journal of Climate* **24**: 4600–4617, DOI: 10.1175/2010JCLI3924.1.

Baker CB, Eischeid JK, Karl TR, Diaz HF. 1994. *The Quality Control of Long-Term Climatological Data Using Objective Data Analysis: Preprints of AMS Ninth Conference on Applied Climatology*: Dallas, TX. 15–20 January 1995.

Brohan P, Kennedy JJ, Harris I, Tett S F B, Jones PD. 2006. Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *Journal of Geophysical Research* **111**: D12106, DOI: 10.1029/2005JD006548.

Durre I, Menne MJ, Vose RS. 2008. Strategies for evaluating quality assurance procedures. *Journal of Applied Meteorology and Climatology* **47**: 1785–1791.

Durre I, Menne MJ, Gleason BE, Houston TG, Vose RS. 2010. Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology* **49**: 1615–1633, DOI: 10.1175/2010JAMC2375.1.

Fall S, Watts A, Nielsen-Gammon J, Jones E, Niyogi D, Christy JR, Pielke RA. 2011. Analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends. *Journal of Geophysical Research (Atmospheres)* **116**: D14120, DOI: 10.1029/2010JD015146.

Fuentes M, Raftery A. 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61**(1): 36–45.

Hansen J, Ruedy R, Sato M, Lo K. 2010. Global surface temperature change. *Reviews of Geophysics* **48**: 29pp., DOI: 10.1029/2010RG000345. RG4004.

Horton R. 2010. Science will never be the same again. *Lancet* **376**: 143–144.

Jasanoff S. 2010. Testing time for climate science. *Science* **238**: 695–696, DOI: 10.1126/science.1189420.

Jones PD, Moberg A. 2003. Hemispheric and large-scale surface air temperature variations: an extensive revision and an update to 2001. *Journal of Climate* **16**: 206–223, DOI: 10.1175/1520-0442(2003)016<0206:HALSSA>2.0.CO;2.

Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M. 2011a. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. *Journal of Geophysical Research* **116**: D14103, DOI: 10.1029/2010JD015218. Measurement and sampling uncertainties.

Kennedy JJ, Rayner NA, Smith RO, Parker DE, Saunby M. 2011b. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. *Journal of Geophysical Research* **116**: D14104, DOI: 10.1029/2010JD015220. Biases and homogenization.

Lawrimore JH, Menne MJ, Gleason BE, Williams CN, Wuertz DB, Vose RS, Rennie J. 2011. An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *Journal of Geophysical Research* **116**: D19121, DOI: 10.1029/2011JD016187.

Lindgren F, Rue H, Lindström J. 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B* **73**: 423–498.

McKitrick RR, Michaels PJ. 2007. Quantifying the influence of anthropogenic surface processes and inhomogeneities on gridded global climate data. *Journal of Geophysical Research (Atmospheres)* **112**(D24): D24S09, DOI: 10.1029/2007JD008465.

Maibach E, Witte J, Wilson K. 2011. Climategate undermined belief in global warming among many American TV meteorologists. *Bulletin of the American Meteorological Society* **92**(1): 31–37, DOI: 10.1175/2010BAMS3094.1.

Menne MJ, Williams CN. 2009. Homogenization of temperature series via pairwise comparisons. *Journal of Climate* **22**(7): 1700–1717.

Menne MJ, Williams CN, Vose RS. 2009. The U.S. historical climatology network monthly temperature data, ersion 2. *Bulletin of the American Meteorological Society* **90**(7): 993–1007.

Parker DE. 2010. Urban heat island effects on estimates of observed climate change. *Wiley Interdisciplinary Reviews – Climate Change* **1**(1): 123–133.

Peterson TC, Easterling DR. 1994. Creation of homogeneous composite climatological reference series. *International Journal of Climatology* **14**(6): 671–679.

Peterson TC, Vose RS. 1997. An overview of the Global Historical Climatology Network temperature data base. *Bulletin of the American Meteorological Society* **78**(12): 2837–2849, DOI: 10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2.

Pielke RA, Davey CA, Niyogi D, Fall S, Steinweg-Woods J, Hubbard K, Lin X, Cai M, Lim YK, Li H, Nielsen-Gammon J, Gallo K, Hale R, Mahmood R, Foster S, McNider RT, Blanken P. 2007. Unresolved issues with the assessment of multidecadal global land surface temperature trends. *Journal of Geophysical Research (Atmospheres)* **112**(D24): D24S08, DOI: 10.1029/2006JD008229.

Rougier JC, Sparks S, Aspinall W, Cornell S, Crosweller S, Edwards T, Freer J, Hill L, Hincks T. 2010. SAPPUR: NERC Scoping Study on Uncertainty and Risk in Natural Hazards Summary and recommendations (revised version), Bristol Environmental Risk Research Centre (BRISK), University of Bristol, UK. Available from http://www.bris.ac.uk/brisk/sappur/.

Russell M, Boulton G, Clarke P, Eyton D, Norton J. 2010. Report of the Independent Climate Change E-mails Review. Available from http://www.cce-review.org/.

Sahu S, Gelfand A, Holland D. 2009. Fusing point and areal level space-time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**: 77–103.

Schmidt GA. 2009. Spurious correlations between recent warming and indices of local economic activity. *International Journal of Climatology* **29**(14): 2041–2048.

Simmons AJ, Willett KM, Jones PD, Thorne PW, Dee DP. 2010. Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: inferences from reanalyses and monthly gridded observational data sets. *Journal of Geophysical Research (Atmospheres)* **115**: D011110, DOI: 10.1029/2009JD012442.

Smith TM, Reynolds RW, Peterson TC, Lawrimore J. 2008. Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880-2006). *Journal of Climate* **21**: 2283–2296, DOI: 10.1175/2007JCLI2100.1.

Thorne PW, Parker DE, Christy JR, Mears CA. 2005. Uncertainties in climate trends – lessons from upper-air temperature records. *Bulletin of the American Meteorological Society* **86**(10): 1437–1442, DOI: 10.1175/BAMS-86-10-1437.

Thorne PW, Brohan P, Titchner HA, McCarthy MP, Sherwood SC, Peterson TC, Haimberger L, Parker DE, Tett SFB, Santer BD, Fereday DR. 2011a. A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *Journal of Geophysical Research (Atmospheres)* **116**: D12116, DOI: 10.1029/2010JD015487.

Thorne PW, Willett KM, Allan RJ, Bojinski SJ, Christy JR, Fox N, Gilbert S, Jolliffe I, Kennedy JJ, Kent E, Klein Tank A, Lawrimore J, Parker DE, Rayner N, Simmons A, Song L, Stott PA, Trewin B. 2011b. Guiding the creation of a comprehensive surface temperature resource for 21st century climate science. *Bulletin of the American Meteorological Society* **92**: ES40–ES47, DOI: 10.1175/2011BAMS3124.1.

Tollefson J. 2010. An erosion of trust?. *Nature* **466**(7302): 24–26, DOI: 10.1038/466024a.

Trenberth KE, Jones PD, Ambenje P, Bojariu R, Easterling D, Klein Tank A, Parker D, Rahimzadeh F, Renwick JA, Rusticucci M, Soden B, Zhai P. 2007. Observations: surface and atmospheric climate change. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds). Cambridge University Press: Cambridge, United Kingdom and New York, NY, USA.

Venema VVC, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, Vertacnik G, Szentimrey T, Stepanek P, Sahradnicek P, Viarre J, Müller-Westermeier G, Lakatos M, Williams CN, Menne MJ, Lindau R, Rasol D, Rustemeier E, Kolokythas K, Marinova T, Andresen L, Acquaotta F, Fratianni S, Cheval S, Klancar M, Brunetti M, Gruber C, Prohom Duran M, Liksol T, Esteban P, Brandsma T. 2012. Benchmarking homogenization algorithms for monthly data. *Climate of the Past* **8**: 89–115, DOI: 10.5194/cp-8-89-2012.

Vose RS, Wuertz D, Peterson TC, Jones PD. 2005. An intercomparison of trends in surface air temperature analyses at the global, hemispheric, and grid-box scale. *Geophysical Research Letters* **32**(18): L18718 (4 pp.), DOI: 10.1029/2005GL023502.

Williams CN, Menne MJ, Thorne PW. 2012. Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research (Atmospheres)*. DOI: 10.1029/2011JD016761, in press.

Yang C, Chandler RE, Isham VS, Wheater HS. 2006. Quality control for daily observational rainfall series in the UK. *Water and Environment Journal* **20**: 185–193.

Zinser T. 2010. Review of NOAA's efforts to modernize the U.S. Historical Climatology Network. *Technical Report STL-19846*, Office of Inspector General, United States Department of Commerce. Available from http://www.oig.doc.gov/OIGPublications/STL-19846.pdf Accessed on 23 September 2011.