

Evaluating Data Journeys: Climategate, Synthetic Data and the Benchmarking of Methods for Climate Data Processing



Wendy S. Parker

Abstract This chapter concerns the benchmarking of methods used to process data in climate science. It explores the nature and value of benchmarking in this context by examining an ongoing initiative – the International Surface Temperature Initiative (ISTI) – that is developing a public databank of temperature observations as well as a system for benchmarking the methods that databank users employ to further process the data. Interestingly, the benchmarking system will make use of “synthetic data” generated with the help of computer simulation models. It is argued here that the benchmarking system has crucial scientific and gatekeeping roles to play in the context of ISTI. It is further suggested that, once we appreciate how synthetic data are to be produced and used by ISTI, we uncover yet another variety of what Paul Edwards (*A vast machine: computer models, climate data, and the politics of global warming*. MIT Press, Cambridge, MA, 2010) has described as “model-data symbiosis” in the practice of climate science.

1 Introduction

In November 2009, email exchanges among climate scientists were taken without authorization from servers at the U.K.’s Climatic Research Unit and made public on the Internet. Dubbed “Climategate” in blogs and popular media, the contents of the emails gave rise to allegations of fraud and scientific misconduct on the part of climate scientists and called attention to an ongoing struggle between climate scientists and climate contrarians over data access. Several independent reviews exonerated climate scientists of the charges of fraud and misconduct but did fault them in one significant respect: for being insufficiently open and transparent in their dealings with contrarian requests for information, including Freedom of Information requests for raw data used to estimate changes in global mean surface temperature over land (see e.g. Russell et al. 2010).

W. S. Parker (✉)

Department of Philosophy, Centre for Humanities Engaging Science and Society (CHESS) & Institute for Data Science (IDAS), Durham University, Durham, UK
e-mail: wendy.parker@durham.ac.uk

© The Author(s) 2020

S. Leonelli, N. Tempini (eds.), *Data Journeys in the Sciences*,
https://doi.org/10.1007/978-3-030-37177-7_10

191

The International Surface Temperature Initiative (ISTI) was launched in 2010, in the wake of the Climategate episode, and seeks to promote transparency and openness in the process of producing temperature change estimates (Thorne et al. 2011). Spearheaded by leading climate data experts in the UK and around the world, ISTI is working to construct a comprehensive, publicly-accessible global databank of historical surface meteorological observations taken over land, providing data at monthly, daily and even sub-daily resolutions. This is a substantial undertaking.¹ It involves not only obtaining observational data from numerous sources around the world, but also getting the data and any available metadata into a common format and then merging the data records with the aim of maximizing station coverage and data quality while minimizing duplication. Release of the first version of the merged data, focused on monthly mean temperatures, occurred in June 2014 (Rennie et al. 2014), with an updated release in October 2015. These releases included data from over 30,000 observing stations worldwide, several times the number typically used in estimating global surface temperature changes over land.

In addition, ISTI intends to develop a set of benchmarking tests for users who generate “data products” from the databank (see also Tempini, [this volume a, b](#) on “derivative datasets”). These products include reconstructions of the evolution of global and regional temperature over time, from which trends and other changes are often calculated. Arriving at such data products requires the application of quality control and “homogenization” algorithms to data in the databank. *Homogenization* is a process that aims to remove jumps and trends in station time series that are due to non-climatic factors, e.g. because an instrument is replaced with a new one, a building is constructed nearby, or the timing of observations changes. In the envisioned benchmarking tests, users would apply their algorithms to synthetic data that contain deliberately-introduced artefacts (known as “inhomogeneities”) that are not known to the users in advance. The idea is to test how well the different homogenization methodologies work by checking their performance on data that are like real climate data in many important respects, but for which the “true” underlying climate signal is known (Willett et al. 2014). ISTI hopes to host all data products developed using the databank on its website, along with information about benchmarking performance for the generating methodologies (Thorne et al. 2011).

This chapter discusses and reflects upon the data journeys envisioned by ISTI, with special attention to the accompanying benchmarking scheme. As outlined further in Sect. 2, these journeys include the traveling of temperature data from a source or holder, through a processing and merging procedure by ISTI, followed by subsequent quality control and homogenization processes undertaken by third parties, which deliver “data products”. We will see that, given methodological decisions along the way, only some data will make the full journey. Section 3 turns to ISTI’s envisioned benchmarking scheme, explaining how its synthetic data are to be produced with the help of simulation models that serve as analogues to the real

¹It is also largely unfunded. Progress has been somewhat slower than desired, in part because participating researchers are largely volunteering their time (with in-kind support from some of their institutions).

world. The benchmarking scheme and its synthetic data are, in a sense, “external to” the envisioned data journeys, but it is argued that they are far from ancillary components of the ISTI project. On the contrary, benchmarking has crucial roles to play, not only in advancing the scientific goals of the project but also by serving an important gatekeeping function in the complex and politicized context of climate change research. Section 4 contends that the proposed use of synthetic data in ISTI’s benchmarking scheme constitutes a distinctive variety of what Paul Edwards (2010) has called “model-data symbiosis” in the practice of climate science. Finally, Sect. 5 offers some concluding remarks.

2 Data Journeys Envisioned by ISTI

Today, there are thousands of land-based weather stations around the world making regular observations of temperature, pressure, humidity and other weather conditions, often overseen by national meteorological services. It was not always so, of course. Regular observations of temperature began at a few sites in Western Europe in the seventeenth century (Camuffo and Bertolin 2012), but it was not until the mid-nineteenth century that coordinated networks of land-based observing stations began to emerge; they expanded rapidly in the twentieth century (Fleming 1998, Ch. 3). In recent decades, there have been major efforts to locate and bring together records of these past surface observations in support of climate change research (e.g. Menne et al. 2012). These ongoing efforts require international cooperation and involve significant “data rescue” activities, including imaging and digitizing of paper records.

ISTI’s envisioned journeys for surface temperature data – from individual records held by sources to data products of use in regional and global climate change research – are conceptualized in terms of six stages (Thorne et al. 2011). Paper records from observing stations, as well as digital images of those records, are what ISTI call “Stage 0” data. Many of the data obtained by ISTI in constructing their databank, however, are Stage 1 data: “digitized data, in their native format, provided by the contributor” (Rennie et al. 2014, 78). In the simplest case, Stage 1 data might have been produced from Stage 0 data by typing into a computer file what is shown on a paper record.² In other cases, Stage 1 data already reflect substantial processing by the contributor. For instance, many of the Stage 1 data obtained by ISTI had already been subjected to quality control and homogenization algorithms by their contributors; though “raw” data are preferable for the databank, these are not what some sources are willing or able to provide, whether for practical or proprietary reasons.

²That person might have translated or transformed the original data record into a preferred format of her own, so it seems that the “native format” here should be understood as whatever format the contributor to ISTI provides.

At Stage 2, data are converted by ISTI from their native format – units, temporal resolution, etc. – to a common format that also includes some metadata. The conversion to a common format sometimes involves averaging, e.g. in order to convert hourly data to daily or monthly average values. The metadata at Stage 2 indicate not only such things as the station’s ID, latitude, longitude and elevation, but also whether the data have undergone quality control or homogenization by the contributor, how a daily or monthly average value was calculated from observations (if this was necessary), and the mode of transmission from contributor to ISTI (*ibid.*, 79). The documentation accompanying the first release of ISTI data indicates that some 58 source collections were converted to Stage 2 data (see Table 1 for a snapshot). Many of these data collections were obtained from national meteorological services, universities and research stations.

At Stage 3, the data sources are prioritized and then subjected to a merge algorithm, with the aim of maximizing station coverage and data quality while minimizing duplication. In the merge performed for monthly data, ISTI chose to give higher priority to sources “that have better data provenance, extensive metadata, come from a national weather or hydrological service, or have long and consistent periods of record” (Rennie et al. 2014, 82). The highest priority source – in ISTI’s case the Global Historical Climatology Network – Daily (GCHN-D) dataset, which contains on the order of a billion observational records (Durre et al. 2010) – becomes the starting point for building the merged dataset.

The merge algorithm then works through the remaining data sources according to their priority. Each record provided by a source is a candidate station. The algorithm first compares the record to a list of stations with known issues in their data or metadata; this list was generated using another algorithm that looks for signs of problems, such as an undocumented shift in units, or flipping the sign of the station’s longitude, etc. If the record/candidate station is not withheld (“blacklisted”) following this comparison, the merge algorithm continues, trying to determine whether the candidate station is unique or matches an existing station. This is a non-trivial task, given that different data sources can use different names for the same station, can represent latitude and longitude with different precision, etc. ISTI describes the merge algorithm as employing a “quasi-probabilistic approach” that “attempts to mimic the decisions an expert analyst would make manually” (Rennie et al. 2014, 81). It involves comparing features of the metadata of station records, and in some cases of the temperature data themselves, and then assigning scores on a set of metrics. Depending on whether those scores pass particular thresholds, the station records are either withheld, added to the dataset as new stations, or merged with records for existing stations (see Fig. 1). The merge algorithm is made available on the ISTI website, and ISTI emphasizes that users can change the threshold settings to produce alternative merged datasets, as ISTI did themselves (see Rennie et al. 2014, Table 12).

In ISTI’s analysis, their “databank” project encompasses the journeys of data from Stage 0 to Stage 3. The final two stages of the envisioned journeys are left to users of the databank; since the databank is publicly available, in principle these users might be anyone. At Stage 4, quality control procedures are applied to Stage 3

Table 1 Partial list of sources of temperature data that were converted to Stage 2 data

Name	Source	Time scale	Raw/QC/homogenized	TMAX	TMIN	TAVG
Antarctica	SCAR Reader Project	Monthly	Raw	N	N	Y
Antarctica (AWS)	Antarctic Meteorological Research Center	Daily	Raw	Y	Y	N
Antarctica (Palmer Station)	Antarctic Meteorological Research Center	Daily	Raw	Y	Y	Y
Antarctica (South Pole Station)	Antarctic Meteorological Research Center	Monthly	Raw	Y	Y	Y
Arctic	IARC/Univ of Alaska Fairbanks	Monthly	Homogenized	N	N	Y
Argentina	National Institute of Agricultural Technology (INTA)	Daily	Raw	Y	Y	N
Australia	Australia Bureau of Meteorology	Daily	Homogenized	Y	Y	Y
Brazil	INPE, Nat. Institute for Space Research	Daily	Raw	Y	Y	N
Brazil-In met	INMET	Daily	Raw	Y	Y	N
Canada	Environment Canada	Monthly	Homogenized	Y	Y	Y
Canada	Environment Canada	Monthly	Raw	Y	Y	Y
Central Asia	NSIDC	Monthly	Homogenized	Y	Y	Y
Channel Islands	States of Jersey Met	Daily	Raw	Y	Y	N
Colonial Era Archives	Griffith	Monthly	Raw	Y	Y	N
CRUTEM4	UKMO	Monthly	Homogenized	N	N	Y
East Africa	Univ. of Alabama Huntsville	Monthly	Raw	Y	Y	Y
Ecuador	Inst. Nacional De Met E Hidrologia	Daily	Raw	Y	Y	N
Europe/N. Africa	European Climate Assessment (Daily, Non-Blended)	Daily	Raw	Y	Y	Y

Source: Rennie et al. (2014, Table 1)

data. It turns out that the GCHN-D data, which form the starting point for constructing the ISTI monthly merged dataset, have already been subjected to quality control by the U.S. National Center for Environmental Information (NCEI).³ The procedure there involves 19 automated tests designed to detect duplicate data, climatological outliers and spatial, temporal and internal inconsistencies; a small number of problematic data (well under 1%) are consequently excluded (Durre et al. 2010).

³This was formerly called the National Climatic Data Center (NCDC).

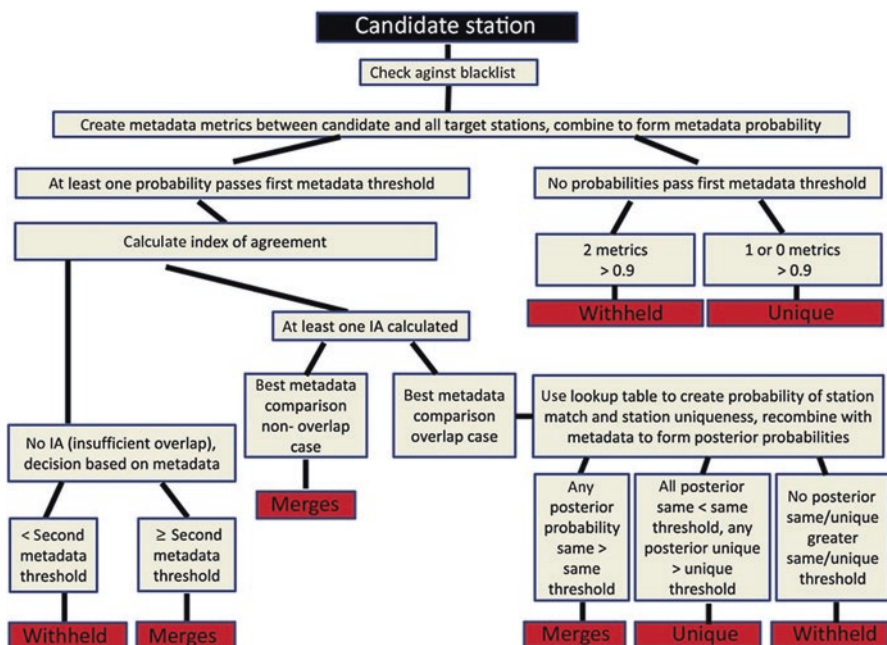


Fig. 1 Workflow for ISTI merge algorithm. (Source: Rennie et al. 2014, Fig. 5)

Many other sources in the ISTI databank, however, have not been subjected to quality control (as their metadata communicates), and it is up to users to address this.

Stage 5 data have, in addition, been homogenized. That is, the data at Stage 5 have been subjected to further processing to try to remove jumps and trends in station time series due to non-climatic factors. When station metadata are available (e.g. reporting a shift in instrument location), this can aid homogenization, but often such metadata are not available. Many homogenization methods thus are statistical methods that compare station records to those of neighbouring stations or of reference stations, identifying and correcting for inhomogeneities based on expected relationships among the records (see e.g. Costa and Soares 2009; Venema et al. 2012). There is substantial uncertainty about how best to identify and correct for inhomogeneities; statistical methods for doing so, for instance, can plausibly employ any of a number of approaches and assumptions. Table 2 summarizes features of several different homogenization algorithms. Even without going into the technical details, one can see that there are differences in what data are compared to (comparison), in how data are searched for potential inhomogeneities (search), and in the form of tests used to identify the presence of inhomogeneities (criterion); there are also differences in how corrections are applied to data once an inhomogeneity has been detected (not shown in Table 2). Attempting to correct for inhomogeneities is particularly important when data will be used to quantify changes in

Table 2 Homogenization algorithms differ in a number of respects

Method	Comparison		Detection		References
	Comparison	Time step	Search	Criterion	
MASH	Multiple references	Annual, parallel monthly	Exhaustive	Statistical test (MLR)	Szentimrey (2007, 2008)
PRODIGE	Pairwise, human synthesis	Annual, parallel monthly	DP	Penalized likelihood	Caussinus and Mestre (2004)
USHCN	Pairwise, automatic synthesis	Serial monthly	HBS	Statistical test (MLR)	Menne et al. (2009)
AnClim	Reference series	Annual, parallel monthly	HBS, moving window	Statistical test	Štepanek et al. (2009)
Craddock	Pairwise, human synthesis	Serial monthly	Visual	Visual	Craddock (1979) and Brunetti et al. (2006)
RhstestV2	Reference series or absolute	Serial monthly	Stepwise	Statistical test (modified Fisher)	Wang (2008)
SNHT	Reference series	Annual	HBS	Statistical test (MLR)	Alexandersson and Moberg (1997)
Climatol	Reference series	Parallel monthly	HBS, moving window	Statistical test	Guijarro (2011)
ACMANT	Reference series	Annual, joint seasonal	DP	Penalized likelihood	Domonkos et al. (2011)

Source: Venema et al. (2012, Table 1)

climate, since trends in the data introduced by non-climatic factors can be of similar size to the changes expected due to increased greenhouse gas emissions.

In contrast to the “data” of Stages 0–3, ISTI refers to Stage 4 and 5 results as “data products” (Thorne et al. 2011). It may be tempting to think that this shift in terminology reflects a substantive change, with later-stage data being, for instance, somehow more heavily processed. This is not really the case, however. As noted above, even some Stage 1 data held by ISTI have been subjected to quality control and homogenization by their sources (see Table 1 above). Thus, while Stage 4 and Stage 5 data will in fact reflect some additional processing by users, similar processing efforts will have already been made with respect to some of the data at earlier stages. ISTI’s distinction between “data” and “data products” primarily marks the boundary of ISTI’s control; results generated by third parties using ISTI’s databank are “data products”.

3 Evaluating Data Journeys: Benchmarking and Its Importance

ISTI scientists hope that users of the databank will develop multiple, independent data products for a given region and period. They hope, for instance, that a variety of reconstructions of global and regional temperature evolution over the twentieth century will be developed, where users apply their own preferred methods for quality control and homogenization to Stage 3 data. Such independent estimates, it is thought, could help to shed light on the extent to which there is uncertainty about temperature trends and other quantities commonly derived from such reconstructions: “Multiple products are the only conceivable way to get even a simple estimate of the structural (methodological choices) uncertainty; we need to attack the problem from many different *a priori* assumptions to create an ensemble of estimates” (Thorne et al. 2011, ES44). Although there are various climate data products already in existence, “quality assurance information is sparse, documentation quality is mixed, and different source data choices and methods can make meaningful inter-comparison hard” (*ibid*). One reason that quality assurance information is sparse is that it is difficult to produce such information in a reliable way. Climate scientists do not have access to the true evolution of regional and global temperatures, nor to some known-to-be-accurate estimates, against which data products can be evaluated.

Benchmarking exercises are now emerging as one approach to learning about the reliability of methodologies used in generating climate data products – that is, in evaluating particular parts of climate data journeys. In very general terms, a benchmark can be understood as “a test or set of tests used to compare the performance of alternative tools or techniques” (Sim et al. 2003). The most ambitious benchmarking exercise to date in climate science is the COST-HOME (European Cooperation in Science and Technology – Advances in Homogenization Methods of Climate Series) project. COST-HOME developed a benchmark dataset and published it online, allowing anyone to attempt to homogenize it and submit data products for evaluation (see Venema et al. 2012). The COST-HOME benchmark dataset included three different types of data, but most contributors focused on the “surrogate data” portion, which was considered the “most realistic” of the three types (*ibid.*, 92). These surrogate data, which represented conditions at a number of small networks of observing stations, were produced with the help of statistical methods, such that they reproduced important statistical features of real homogenized data, such as their “distribution, power spectrum and cross spectra”; several known types of inhomogeneities and other “data disturbances” were then added, and the task for participants was to recover the homogenous surrogate data (*ibid.*). Importantly, those homogenous data were not disclosed to participants until after a deadline for submission of data products. Twenty-five submissions were received, based on 13 different homogenization methods, including some manual methods (*ibid.*). These were evaluated on a variety of metrics that measure similarities between the submitted data product and the homogeneous surrogate data (i.e. “truth”).

ISTI envisions a benchmarking scheme that is similar to that of COST-HOME in some respects. Participants submitting data products for evaluation will not know in advance the “true” underlying data to which inhomogeneities were added. In addition, the benchmarking exercise will be open to all. In fact, ISTI “strongly advocates” that anyone producing Stage 5 data products from the databank take part in benchmarking exercises (Willett et al. 2014). But there are also some differences. Rather than data for small networks of stations, ISTI plans to construct global benchmark datasets, representing what they refer to as “analog inhomogeneous worlds” (ibid.; Thorne et al. 2011), i.e. analogues to the inhomogeneous data collected in the real world. In addition, the construction of these benchmarks will begin not from homogenized real data, but from computer simulations from global climate models.⁴ These simulation results, which include values of temperature on a regular grid, will be interpolated to a set of 30,000+ stations analogous to those in the databank (Willett et al. 2014). Inhomogeneities will then be added to these “analog-clean worlds”, to produce “synthetic data”. The inhomogeneities are intended to be “physically plausible representations of known causes of inhomogeneity (e.g. station moves, instrument malfunctions or changes, screen/shield changes, changes to observing practice over time, and local environment changes)” (ibid., 192). See Fig. 2 for a depiction of some of the ways in which the benchmarking exercise mirrors the analysis of the “real” ISTI databank data.

ISTI highlight several positive features of their envisioned simulation-based approach to the generation of benchmarking datasets. Time series of temperature values from a climate model will be free from inhomogeneities, so the “true” climate signal will be known. In addition, the data will include “globally consistent variability”, including coherent variability associated with events like El Nino – Southern Oscillation (ENSO). Moreover, it will be possible to generate inhomogeneous worlds with different levels of background climate change, since climate models can be run under a variety of scenarios in which greenhouse gas concentrations are rising rapidly, held constant, etc.; at least some information then can be obtained about how the skill of different homogenization algorithms varies, if at all, with the level of background climate change.

ISTI proposes to provide ten inhomogeneous worlds/synthetic datasets in a given benchmarking cycle, each based on a different simulation, with the cycle of analysis and evaluation repeating roughly every 3 years (ibid.). The aim is for these different worlds to incorporate inhomogeneities with a range of frequencies and magnitudes, seasonality, and geographical pervasiveness (e.g. when a whole network changes observing practices at once). Participants would submit their homogenized benchmark data for evaluation by ISTI. The results of this assessment as well as “truth”

⁴These climate models incorporate both basic physical theory (from fluid dynamics, thermodynamics, etc.) and some simplified/idealized representations of small-scale processes; the latter are necessary in part because limited computational power constrains the resolution at which the climate system can be represented. The knowledge on which the models are based, including the theoretical knowledge, is of course empirical, but the climate models are not data-driven models obtained by fitting curves to observations.

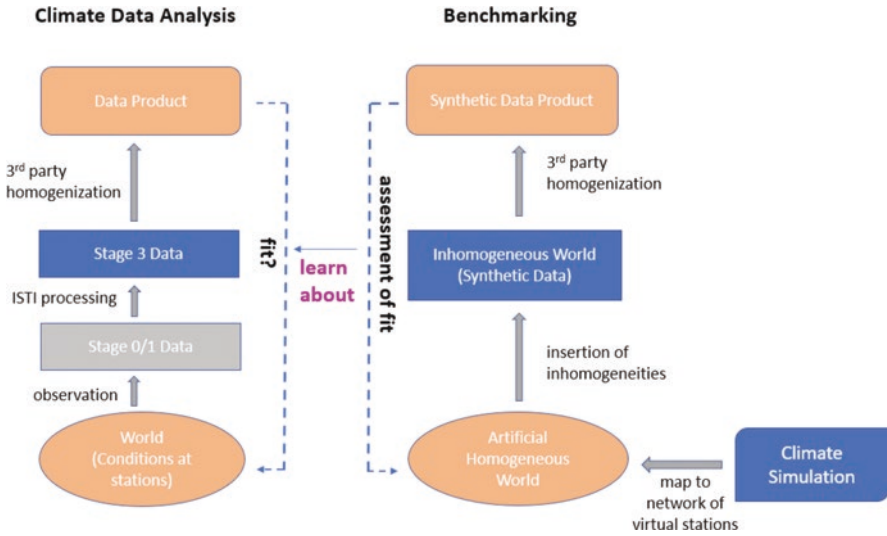


Fig. 2 Envisioned benchmarking of homogenization algorithms. ISTI’s analogue worlds allow for testing of homogenization algorithms in cases where “truth” is known. The aim is to learn about how these algorithms are likely to perform on real data where similar inhomogeneities are present but truth cannot be known

for the ten cases – i.e. the clean analog worlds produced by sampling/interpolating simulation results – would subsequently be unveiled. The cycle would then repeat.

ISTI’s envisioned benchmarking system is intended to support three important scientific goals of the ISTI project: quantification of the potential structural uncertainty of a given climate data product; objective intercomparison of such data products; and advancing homogenization algorithm development (Willett et al. 2014, 192). These are discussed here in reverse order.

The benchmarking scheme aims to support homogenization algorithm development by helping developers to learn more about the strengths and weaknesses of their algorithms – which sorts of inhomogeneities they are good at detecting and correcting, which they are not, etc. In further support of this goal, ISTI plans to provide some “open benchmarks” for which “truth” is also immediately available, so that participants can conduct some preliminary tests of their algorithms before submitting to the evaluation exercise. But the task of homogenizing data for which “truth” is not known to algorithm developers remains very important, since for these cases developers cannot optimize their algorithms to specific features of known inhomogeneities in the data; such optimization can make an algorithm a good performer on that particular dataset, even though it might perform poorly on datasets with somewhat different (but still plausible) inhomogeneity profiles.

It is important to recognize that, insofar as what is learned via ISTI’s benchmarking exercises leads to changes in homogenization algorithms, data journeys of the *future* that involve the application of those algorithms will be somewhat different too.

Reconstructions of the evolution of global and regional surface temperatures since pre-industrial times will be produced again and again as new observations are made and additional past data are rescued and digitized; with homogenization algorithms that are changed in light of past benchmarking exercises, those reconstructions will be somewhat different than they otherwise would have been. Thus, while the sort of benchmarking exercises envisioned by ISTI can be considered “external” to data journeys involving real data, they can influence those journeys by prompting adjustments to homogenization algorithms whose application constitutes part of the journey.

Second, the benchmarking scheme supports the goal of objective and meaningful intercomparison of climate data products, such as reconstructions of global temperature change over the twentieth century. As noted earlier, for some types of data product there already have been multiple products developed by different scientific groups, but it is often difficult to compare the quality of these products, in part because they are constructed from somewhat different source data and in part because there can be no appeal to “truth” to settle the matter. In the benchmarking exercise, participants will all be starting from the same synthetic dataset; differences in their performance will be attributable to differences in their processing methodologies. Moreover, performance on the synthetic data will be objectively assessable, since for these data “truth” is known. Learning about such performance can be useful not only for homogenization algorithm developers (as just noted above), but also for users of climate data products. For instance, if such evaluation reveals that some homogenization algorithms are particularly good at correcting for some types of errors that are, for a user’s intended application, particularly important to avoid, users can choose to work with data products generated with those homogenization algorithms. (In effect, users would then be selecting data products on inductive risk grounds, informed by what is learned via benchmarking activities.) This is just one important way in which the ISTI project can support climate-related research, including research intended to inform societal decision making (often called “climate services”).

Finally, and relatedly, the benchmarking exercise supports ISTI’s goal of providing information about uncertainties associated with climate data products, in particular uncertainties stemming from the process of homogenization. One of the potential benefits of an open-access observational databank is that multiple, independent groups can use the databank to construct data products for the same regions and periods; since there are uncertainties about how best to carry out that construction process, especially in the homogenization step, and since different groups will make somewhat different methodological choices in the face of that uncertainty (see Sect. 2 and Table 2), the products generated by the different groups can, in principle, sample current scientific uncertainty about past conditions in a particular region/period. This is analogous to the way in which a set of forecasts from different weather prediction models can, in principle, sample current scientific uncertainty about tomorrow’s weather conditions. But just as there may be weather prediction models that have strong biases in particular regions – and whose forecasts for those regions we thus wouldn’t want to take at face value – so can there be homogenization methods that have particular strengths and weaknesses that (if known) should

affect how we interpret their results. By helping to reveal those strengths and weaknesses, the benchmarking exercise can aid the interpretation of the set of data products generated, including whether their face-value spread should be considered a lower bound on current uncertainty.

Closely related to this is another important, beneficial function that the benchmarking scheme can serve, though it is not often emphasized by ISTI: a gate keeping function. When it comes to the generation of data products using the databank, ISTI explicitly encourages “contributions from non-traditional participants” (Thorne et al. 2011, ES44). They recognize the possibility of “useful insights from people tackling the problem by thinking “outside the box”” (ibid.). But while this is indeed a potential benefit of an open-access databank, there is also the risk that users with insufficient expertise, political motivations, and so on will decide to generate their own data products, e.g. their own reconstructions of global temperature change over the twentieth century. Such data products may, either unintentionally or intentionally, give a highly misleading picture of the evolution of past climate conditions. For example, suppose that a homogenization algorithm effectively guaranteed that temperature reconstructions would show very little twentieth century warming, almost regardless of the data; the worry arises that such a reconstruction would be touted in sceptical blogs, newspapers, etc. and would add further confusion to public discussion of climate change. If those generating the reconstruction were to participate in ISTI’s benchmarking exercises, however, it might be revealed that their methodologies were highly flawed, in the sense that they did not recover anything like the “truth” in the benchmark cases. The benchmarking system thus could provide “a way of separating the wheat from the chaff” (Stott and Thorne 2010, 159) when it comes to data products generated from the ISTI databank. Of course, anyone might refuse to participate in ISTI’s benchmarking exercises, but this refusal could itself constitute reasonable grounds for questioning the reliability of data products that differ markedly from those produced by others.

Thus, far from being an ancillary component of the ISTI project, synthetic data have crucial roles to play alongside “real” climate data when it comes to learning about past climate change; without synthetic data, and the accompanying benchmarking scheme, some of the primary scientific goals of the ISTI project would be in jeopardy. This does not mean, of course, that there are no limits to what benchmarking can achieve. The kinds of benchmarking exercises envisioned by ISTI can only gauge the performance of homogenization algorithms with respect to the particular inhomogeneities inserted into the synthetic data; even if an algorithm were to consistently and perfectly recover the “truth” in benchmarking exercises, this would be no guarantee that it performs similarly well on real climate data, since there is no guarantee that the inhomogeneities in the latter are fully encompassed by the inhomogeneity types present in the benchmark data. There may be types of inhomogeneities in actual climate data that go beyond those that current scientists have good reason to believe are sometimes present. Moreover, though the use of synthetic data generated with the help of simulation models has the attractive features discussed above, it is also true that simulation results (and synthetic station data interpolated from them) may lack some spatial and temporal characteristics of

real climate data, due to limitations of the climate models used (e.g. their omissions, simplifications, etc.). The ISTI benchmarking team suggests checking empirically whether synthetic data display key statistical properties of real climate data (e.g. levels of correlation among data for nearby stations, station autocorrelation, etc.), using real data that are thought to be of relatively high quality (Willett et al. 2014, 191).

4 Another Variety of Model-Data Symbiosis

In his insightful analysis of the development of modern meteorology and climate science, Paul Edwards (1999, 2010) argues that we find in these domains a kind of symbiosis between models and data – a mutually beneficial but mutually dependent relationship. Computer models of the atmosphere and climate system, he points out, are *data-laden* to a certain extent: in addition to equations from fundamental physical theory, they require various “semi-empirical parameters” that are derived (in a loose sense) from observations. At the same time, weather and climate data are often *model-filtered*. Here he has in mind several kinds of models.

Most striking is the use of computer simulation models in a process known as “data assimilation”. A weather forecast from a computer simulation model provides a first-guess estimate of the atmospheric state, which is then updated in light of available observations to arrive at a revised, best-guess estimate of the state; this best-guess estimate then serves as the initial conditions for the next set of forecasts from the weather model. The same sort of technique has been used retrospectively in climate science, to generate long-term gridded datasets from gappy, irregular historical observations. These “reanalysis” datasets complement the kinds of climate data products described in previous sections of this paper (Parker 2016). When it comes to those data products, Edwards notes that what might be called “intermediate models” – which include models of instrument behaviour, techniques for quality control and many other methods (1999, 450) – are essential to their production; he explicitly notes their use in the process of homogenization.

ISTI’s benchmarking scheme employing synthetic data illustrates yet another variety of model-data symbiosis in climate science, once again involving computer simulation models. Here, however, simulation models are used not to *fill in gaps in datasets* (as they in effect are used in data assimilation) but rather to help *evaluate the quality of datasets/data products*, by helping to assess the strengths and weaknesses of some of the methods used in the production of those datasets/data products. An understanding of the quality is in turn important for using the datasets effectively for various purposes, including for the evaluation of computer simulation models themselves. Indeed, one of ISTI’s stated motivations for constructing an open-access observational databank that includes not just monthly but daily data, is that sub-monthly data are needed for studies of changes in climate extremes, like floods and heatwaves, as well as for evaluating today’s climate models’ ability to

simulate such extremes. Thus, we have climate models assisting in the evaluation of climate data products, so that those climate data products in turn can assist in the evaluation of climate models – a mutually beneficial, but mutually dependent relationship.

5 Concluding Remarks

ISTI is a major effort to promote transparency and openness in the management of surface temperature data, one which has the potential not only to help circumvent the kinds of skirmishes over access to climate data that have occurred in the recent past but also to provide better insight into the uncertainties associated with existing estimates of changes in temperature since pre-industrial times. Its success in the latter, however, depends not only on users actually generating data products that reflect a range of different methodological choices, but also on there being a means of ensuring that these products are of sufficient quality. While still under development, an ingenious benchmarking scheme, involving tests of data processing algorithms on synthetic data, is meant to serve as one important way of gauging the quality of user-generated data products. Far from being an ancillary component of the ISTI project, the benchmarking system has crucial roles to play, not only in advancing the scientific goals of the project but also by serving an important gatekeeping function in the complex and politicized context of climate change research.

The use of synthetic data in benchmarking efforts like that envisioned by ISTI also illustrates a distinctive variety of Edwards' model-data symbiosis in climate science. While he calls attention to cases in which computer simulation models have been used to help fill in gaps in observational data, the envisioned use of synthetic data in benchmarking exercises would involve simulation models aiding the process of evaluating climate datasets, including their attendant uncertainties. These datasets in turn are to be used for, among other purposes, evaluating climate models themselves. Once again, we find climate models and climate data standing in a mutually beneficial but mutually dependent relationship.

References

- Alexandersson, H., and A. Moberg. 1997. Homogenization of Swedish temperature data.1. Homogeneity test for linear trends. *International Journal of Climatology* 17: 25–34.
- Brunetti, M., M. Maugeri, F. Monti, and T. Nanni. 2006. Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series. *International Journal of Climatology* 26: 345–381.
- Camuffo, Dario, and Chiara Bertolin. 2012. The Earliest Temperature Observations in the world: The Medici Network (1654–1670). *Climatic Change* 111 (2): 335–363. <https://doi.org/10.1007/s10584-011-0142-5>.

- Caussinus, H., and O. Mestre. 2004. Detection and correction of artificial shifts in climate series. *Applied Statistics* 53: 405–425.
- Costa, Ana Cristina, and Amílcar Soares. 2009. Homogenization of Climate Data: Review and New Perspectives Using Geostatistics. *Mathematical Geosciences* 41 (3): 291–305. <https://doi.org/10.1007/s11004-008-9203-3>.
- Craddock, J.M. 1979. Methods of comparing annual rainfall records for climatic purposes. *Weather* 34: 332–346.
- Domonkos, P., R. Poza, and D. Efthymiadis. 2011. Newest developments of ACMANT. *Advances in Science and Research* 6: 7–11. <https://doi.org/10.5194/asr6-7-2011>,
- Durre, Imke, Matthew J. Menne, Byron E. Gleason, Tamara G. Houston, and Russell S. Vose. 2010. Comprehensive Automated Quality Assurance of Daily Surface Observations. *Journal of Applied Meteorology and Climatology* L49: 1615–1633. <https://doi.org/10.1175/2010JAMC2375.1>.
- Edwards, Paul N. 1999. Global Climate Science, Uncertainty and Politics: Data-Laden Models, Model-Filtered Data. *Science as Culture* 8 (4): 437–472. <https://doi.org/10.1080/09505439909526558>.
- . 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.
- Fleming, James R. 1998. *Historical Perspectives on Climate Change*. New York: Oxford University Press.
- Guijarro, J.A. 2011. *User's guide to climatol. An R contributed package for homogenization of climatological series*. State Meteorological Agency, Balearic Islands Office, Spain: Report. available at: <http://webs.ono.com/climatol/climatol.html>.
- Menne, Matthew J., Imke Durre, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston. 2012. An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology* 29: 897–910. <https://doi.org/10.1175/JTECH-D-11-00103.1>.
- Menne, M.J., C.N. Williams Jr. and R.S. Vose. 2009. The U.S. historical climatology network monthly temperature data, version 2. *B. Am. Meteorol. Soc.* 90: 993–1007. <https://doi.org/10.1175/2008BAMS2613.1>.
- Parker, Wendy S. 2016. Reanalyses and Observations: What's the Difference? *Bulletin of the American Meteorological Society* 97 (9): 1565–1572. <https://doi.org/10.1175/BAMS-D-14-00226.1>.
- Rennie, Jared J., Jay H. Lawrimore, Byron E. Gleason, et al. 2014. The International Surface Temperature Initiative Global Land Surface Databank: Monthly Temperature Data Release Description and Methods. *Geoscience Data Journal* 1: 75–102. <https://doi.org/10.1002/gdj3.8>.
- Russell, Muir, et al. 2010. *The Independent Climate Change Email Review*. Available at <http://www.cce-review.org/index.php>. Accessed 10 Sept. 2017.
- Sim, Susan E., Steve Easterbrook, and Richard C. Holt. 2003. Using Benchmarking to Advance Research: A Challenge to Software Engineering. *Proceedings of the Twenty-fifth International Conference on Software Engineering*, pp. 74–83. doi: <https://doi.org/10.1109/ICSE.2003.1201189>.
- Štěpánek, P., P. Zahradníček, and P. Skalák. 2009. Data Quality Control and Homogenization of the Air Temperature and Precipitation Series in the Czech Republic in the Period 1961–2007. *Advances in Science and Research* 3: 23–26.
- Stott, Peter, and Peter Thorne. 2010. How best to log local temperatures? *Nature* 465: 158–159.
- Szentimrey, T. 2007. Manual of homogenization software MASHv3.02. *Hungarian Meteorological Service* 65.
- . 2008. *Development of MASH homogenization procedure for daily data*. Proceedings of the fifth seminar for homogenization and quality control in climatological databases, Budapest, Hungary, 2006. WCDMP-No. 71, pp. 123–130.

- Tempini, Niccolò. this volume-a. The Reuse of Digital Computer Data: Transformation, Recombination and Generation of *Data Mixes* in Big Data Science. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer.
- Tempini, Niccolò. this volume-b. Visual Metaphors: Howardena Pindell, Video Drawings, 1975. In *Data Journeys in the Sciences*, ed. Sabina Leonelli and Niccolò Tempini. Cham: Springer..
- Thorne, Peter, et al. 2011. Guiding the Creation of a Comprehensive Surface Temperature Resource for Twenty-First-Century Climate Science. *Bulletin of the American Meteorological Society* 92: ES40–ES47. <https://doi.org/10.1175/2011BAMS3124.1>.
- Venema, Victor K.C., et al. 2012. Benchmarking Homogenization Algorithms for Monthly Data. *Climate of the Past* 8: 89–115. <https://doi.org/10.5194/cp-8-89-2012>.
- Wang, X.L.L. 2008. Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or F test. *Journal of Applied Meteorology and Climatology*. 47: 2423–2444.
- Willett, Kate, et al. 2014. A Framework for Benchmarking of Homogenisation Algorithm Performance on the Global Scale. *Geoscientific Instrumentation Methods and Data Systems* 3: 187–200. <https://doi.org/10.5194/gi-3-187-2014>.

Wendy S. Parker is Associate Professor of Philosophy at Durham University, where she also codirects the Centre for Humanities Engaging Science and Society (CHESS) and the Institute for Data Science (IDAS). She received her PhD in History and Philosophy of Science from the University of Pittsburgh in 2003. Her research focuses on the methodology and epistemology of contemporary science – especially questions related to modelling, evidence, explanation and values – with a particular focus on climate science and meteorology. Her papers have been published in a range of philosophical and scientific journals.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

