Harvard Data Science Review • Issue 2.4, Fall 2020

Perspectives on Data Reproducibility and Replicability in Paleoclimate and Climate Science

Rosemary Bush, Andrea Dutton[1], Michael Evans, Rich Loft[2], Gavin A. Schmidt[3]

[1]Department of Geoscience, University of Wisconsin-Madison,
[2]Computational & Information Systems Lab, National Center for Atmospheric Research,
[3]NASA Goddard Institute for Space Studies and Center for Climate Systems Research, Columbia University

Published on: Dec 16, 2020

DOI: 10.1162/99608f92.00cd8f85

ABSTRACT

This paper summarizes the current state of reproducibility and replicability in the fields of climate and paleoclimate science, including brief histories of their development and applications in climate science, new and recent approaches towards improvement of reproducibility and replicability, and challenges. Recommendations for addressing those challenges include: development of searchable, auto-updated, interlinked, multi-archive public paleoclimate repositories for raw and processed digital datasets; cross-center standardized code base cases, improved data storage techniques, and a focus on replicability for climate simulation storage and access; and support of the development and community awareness of findable, accessible, interoperable and reusable (FAIR) principles by funding agencies and publishers. This paper is largely based on the May 2018 presentations of a panel of researchers to the Committee on Reproducibility and Replicability in Science, part of the National Academies of Science, Engineering, and Medicine. The commentary and recommendations made here are in alignment with those of its Consensus Study Report on Reproducibility and Replicability in Science (2019).

Keywords: data repositories, global climate models, GCMs, proxy archives, replication, reproducibility, metadata

1. Introduction

Reproducibility and replicability are considered core tenets of the scientific process and central to the creation and testing of scientific theories and models of the way the world works (Goodman et al., 2016; Stodden et al., 2016; National Academies, 2019; Powers and Hampton, 2019). Our collective understanding of the mechanics of Earth's climate, the history of ancient climates (paleoclimates) on Earth, and any predictions we can make of future climates are built from the core principles of the scientific method: hypotheses not disproven by empirical observations that are transparently reported as studies subject to external review and repeated testing, creating a feedback cycle that accrues information and seeks consilience (Wilson, 1999). However, Earth's climate is tremendously complex, meaning that scientific studies of climate and paleoclimate are also increasingly complex. Some recent examples of this growing complexity include the work of the Past Global Changes consortium studying the most recent 2000 years of Earth's climate history (PAGES2k), the World Climate Research Programme's Coupled Model Intercomparison Project, now in its 6th phase (CMIP6), the work of the MARGO international working group on reconstructions of sea surface temperatures (Kucera et al., 2006), the International Ocean Discovery Program (IODP), as well as the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) and NCEP North American Regional Reanalysis (NARR), both hosted by NOAA. Natural systems are also simultaneously dynamic and heterogeneous, such that natural phenomena rarely, if ever, repeat themselves to observational precision (Powers & Hampton, 2019; Fraser et al., 2020). Furthermore, as computational analyses grow in size and complexity they also become more difficult to repeat exactly (Stodden et al., 2016), and scholarly communication increasingly relies on access to digital repositories of data and metadata (Konkol et al., 2020; Nüst and Pebesma, 2020). Although journals such as *Scientific Data* and *Geoscience Data Journal* now exist to increase access to natural science datasets, we must consider what it means when a study cannot be repeated, as is increasingly the case in paleoclimate and climate sciences.

Paleoclimate studies are rarely strictly reproducible, as they often rely on physically finite and unique samples—for example, ice cores from the middle of the Antarctic ice sheet (Petit et al., 1999), seafloor sediment cores at 5 km below the ocean surface (Zachos et al., 2001), and cave speleothems (Wang et al., 2001). Expeditions to collect samples from remote field location are often expensive, logistically challenging, and rarely repeated exactly (for example, see the ongoing work of the IODP), and although many projects work on a split core basis—with half going to the researcher and half staying in a core repository for future use—sample material is consumed in destructive analyses as a matter of routine. In a different way, computer-simulated global climate model (GCM) experiments are difficult to reproduce precisely, as their constituent variables have become so complex and achievable resolution so high that the required calculations push the limits of current supercomputer capabilities (National Research Council, 2012). In addition, they generate a tremendous quantity of data that must somehow be labeled, sorted, and made both accessible and usable to others. These issues of reproducibility and replicability also relate to the recent call made by UNESCO and other international stakeholders for Open Science to solve pressing global challenges.

Therefore this paper seeks to (1) address how the awareness and understanding of reproducibility and replication in climate science have evolved over recent years, (2) identify some specific challenges regarding reproducibility and replicability in paleoclimate and climate sciences and steps being taken to manage those challenges, and (3) make recommendations for better handling of reproducibility and replicability issues in paleoclimate and climate sciences going forward for both individual scientists and institutions. The paper is structured as follows: Section 1 defines the terminology we will use; Section 2 focuses on the unique issues in paleo-climate research; Section 3 focuses on the very different challenges faced by researchers using Global Climate Models (GCMs); Section 4 discusses the more general intersections of multiple data types in climate science; Section 5 provides some suggestions and recommendations for improvements, and Section 6 concludes. This paper is based largely on the presentations of a panel of researchers on May 9, 2018, convened by the Committee on Reproducibility and Replicability in Science, part of the National Academies of Science, Engineering, and Medicine: Michael N. Evans, Gavin A. Schmidt, Rich Loft, and Andrea Dutton. An earlier draft of this paper served to inform the National Academies of Sciences, Engineering, and Medicine Consensus Study Report on Reproducibility and Replicability in Science (2019).

1.1. Definitions

As has been noted in other fields (e.g., Goodman et al., 2016; Powers and Hampton, 2019), the terms 'reproducibility' and 'replicability' are defined and used differently, sometimes interchangeably (Bollen et al., 2015; Konkol et al., 2018). This is due in part to the origins and histories of opposing definitions for the terms in experimental versus computational sciences (Drummond, 2009; Plesser, 2018). Here the terms are defined following the recently published National Academies Report, *Reproducibility and Replicability in Science* (National Academies, 2019). *Reproducibility* refers to the computational ability to duplicate results using the same materials, data, methods, and/or analytical conditions. *Replicability* is broader in scope and refers to the ability to obtain results using new materials, data, methods, and/or conditions that are consistent with the original study and independently confirm its results. These definitions are at odds with traditional uses of the terms in experimental sciences (Plesser, 2018), but the National Academies Report (2019) notes the focus on computational analysis because of its large and ever-increasing role in science, including climate and paleoclimate research.

Similarly, 'repository' and 'archive' are both used in lay terms to refer to a storage place for knowledge. For the sake of clarity, here *repository* refers to a database in which information is electronically stored, and *archive* is used only in the paleoclimate sense of the term, referring to the physical materials in which paleoclimate information is in some way imprinted (e.g, drill cores of glacial ice or ocean-bottom sediments).

2. Paleoclimate Proxy Data Collection

2.1. Paleoclimate as the Context for Understanding Climate Today

Analyses of archives of geologic materials utilize proxy records of paleoclimate to reconstruct aspects of the climate system such as temperature and sea level (e.g., Kopp et al., 2016; Rohling et al., 2019; Kaufman et al., 2020). These paleoclimate observations and reconstructions serve multiple purposes, providing (1) an important baseline and deep-time historical context for our understanding of climate today and natural variations in the climate system, especially pre-industrial climate conditions, (2) a natural laboratory to study climate dynamics outside of the historical period of direct instrumental measurements, and (3) valuable benchmarks with which to evaluate climate models, allowing researchers to test whether the models that are used to project future changes in climate conditions can replicate conditions observed or reconstructed from the paleoclimate archives (Schmidt, 2010).

Paleoclimate archives can include any physical, chemical, or biological materials in the geologic record from which observations can be made and connected with some environmental parameter in the modern world. Much quantitative paleoclimate data comes from physical, chemical, or isotopic measurements of a geologic archive. Proxy records require first a calibration model that describes the relationship between climate parameter and its proxy observation, e.g. water temperature and Mg/Ca ratio in planktonic foraminifera (Anand et al., 2003; Jones et al., 2009). This requires also constraining some amount of uncertainty inherent in the relationship, which in turn may be propagated through all subsequent interpretations, including the reconstruction of any climate parameters from the proxy observations. If these proxy system models are not explicitly described, it may be difficult for subsequent research to reproduce, revise and improve on the associated reconstructions as additional data and improved uncertainty quantification are developed (e.g., Evans et al., 2013). Second, proxy reconstructions rely on measurements in geologic archives, which includes uncertainty and issues of reproducibility and replicability in the proxy measurements themselves, as well as—and just as importantly—uncertainty in the process by which the archive was formed, and the way in which chronological information is assigned to the proxy observation. The age model applied to a geologic archive can provide not only absolute age estimates but also be used to determine rates of change, which is especially critical to understanding climate dynamics through time (Cronin, 2010). However, any particular age model may provide limited means for assessing reproducibility and replicability, introducing an additional degree of uncertainty to the paleoclimate reconstruction. If only the derived ages are presented in a published study, and the original source data on which the assigned chronology rests are not published, chronological error can become locked into the interpretation once the study is published and persist even after geochronology is revised or altered (Dutton et al., 2017).

2.2. Reproducibility and Replicability in Paleoclimate Studies

A high degree of public scrutiny in recent years has created strong incentives in the climate science community for high standards in the awareness and understanding of issues of reproducibility and replicability in paleoclimate studies (e.g., National Research Council, 2006). The improvement in the state of reproducibility and replicability can be seen in comparing older studies such as Bond et al. (2001) which, while highly-cited, contained very little metadata or mention of data or code availability, to more recent studies such as Abram et al. (2016), which has source data and analytical code that was stored at NOAA at the time of publication and serves as an example of the direction for best practices in reproducibility. Online data storage has become easier and more prevalent, and many journals and granting agencies now encourage or require online data storage with publication. There is extensive work done on synthesis of data and analyses, including work by the Intergovernmental Panel on Climate Change (IPCC) and Past Global Changes (PAGES), two international working groups that support international and interdisciplinary climate and environmental change research collaborations. Over the past two decades, both of these organizations have developed improved strategies to make the inputs to and the products of their synthesis activities available for further research, analysis, reproduction and replication. For example, the Program for Climate Model Diagnosis and Interpretation (PCMDI) was established in 1989, supports the Earth System Grid Federation (ESGF) that hosts CMIP6, which will be the basis for much of the IPCC's current 6th

Assessment Report synthesis of climate change projections. The PAGES community has developed global-scale temperature-sensitive paleoclimate proxy databases (PAGES2k Consortium, 2013, 2017; Kaufman et al., 2020), for which data and diagnostic tools are provided by the National Center for Environmental Information's (NCEI) World Data Center A for Paleoclimatology, with support from the International Science Council (ISC).

In studies of paleoclimate proxies from geologic archive materials, replicability can be tested both within and between individual studies. The majority of paleoclimate studies involve the destructive analysis of finite natural samples, meaning that collected source material can only be measured a finite number of times, which inherently limits the reproducibility of a study. Materials are often sub-sampled for repeated analyses to validate measurement precision, assess the degree of material preservation, and constrain the internal variability of a sample, site and/or proxy observation, interpretation and chronological accuracy (Bradley, 2014). External reproducibility is constrained through the comparison of community-accepted standards distributed widely among labs, e.g., the reporting of stable isotopic composition via standardization to international reference scales established by community and the International Atomic Energy Agency (IAEA), through the use of standard protocols for acquiring and reporting observations (Coplen et al 2002; de Groot, 2004).

Because of the inherent heterogeneity of the natural world as well as the difficulty and expense of accessing remote field sites, field expeditions often attempt to constrain intra-site reproducibility through multiple parallel sample collections, e.g. repeated sediment or ice cores collected from a single site and/or multiple similar sites during one field expedition. One example is that of paired ice core drilling programs 28 kilometers apart at Summit, Greenland by two international consortia (GISP and GRIP) (Johnsen et al., 1992; Cuffey and Clow, 1997), which demonstrated the robustness of much of the record, but highlighted discrepancies in the older part of the cores (Alley et al., 1995) that arose from differential glacial flow complexities. Paleoclimate proxy analyses may also be replicated via comparison of different analytical measurements made on the same material. For example, DeLong et al. (2013) analyzed Sr/Ca ratios of corals through time using three separate sampling transects to confirm consistency of results between individual transects and quantify uncertainty in the results. The study also counted annual rings in the corals and used U-Th dating to separately estimate ages to test the degree of uncertainty in the age model and constrain error across both proxy observation amplitude and chronology. Community-accepted standards provide a basis for data reproducibility, such that, given the constraints on repeated measurements of finite natural samples and the fundamental limitations on the reproducibility of spatially and temporally heterogeneous natural systems, researchers can be confident that if the standard results are similar across different laboratories, then sample results could also be reproduced by another laboratory or instrument if it were to analyze the same sample. In general, practices of reproducibility in sampling and measurement and inter-laboratory calibration using shared standard materials are well-established and are standard operating procedures for most proxy analyses.

Additional efforts to replicate results across paleoclimate studies employ observing networks which are expected to demonstrate correlated variability—expanding the number of records, whether to fill in gaps in time or space with additional samples or to compare multiple proxies for the same climate parameter. For example, Linsley et al. (2015) reconstructed sea surface temperature (SST) anomalies from coral Sr/Ca measurements from Fiji, Tonga and Rarotonga, all sites which were expected to record the same interdecadal pattern of variation. The proxy-based composite record qualitatively increases the fraction of variance explained by direct historical observations of SST anomaly from these sites, relative to that observed from any of the three records individually, by averaging across random error and emphasizing the regional SST signature. The composite record not only demonstrates the reproducibility of the individual data but also smooths some of the errors inherent in them. The availability of the dataset through online repositories makes it then possible for other scientists to build larger networks, improve composite records, and develop spatially resolved reconstructions (PAGES2k Consortium, 2017; Neukom et al., 2019).

2.3. Frontiers in Paleoclimate Data Synthesis and Archiving

New practices for replicating and reproducing paleoclimate data include the increasing availability of open-access, online databases that facilitate comparison and synthesis. There is also continual development of new paleoclimate proxies and refinement of existing proxies, and each proxy has its own level of associated uncertainty that is often dependent on setting and analytical precision. Interdisciplinary collaborations are also increasing—for example, collaborations between geochemists and geophysicists in resolving discrepancies in geochemical proxy-based reconstructions of sea level using geophysical models of glacial isostatic adjustments (Potter & Lambeck, 2004; Dutton et al., 2015) to create a geophysically synchronized 'stack' of multiple proxy records across time that can serve as a more robust test case and comparison for climate models, similar previous proxy record stacks such as that for benthic oxygen isotopes (Lisiecki & Raymo, 2005). For true replication of a study, all its components, including sample materials, chronology and methods of proxy data conversion and interpretation, must be stored in repositories and made freely accessible. New platforms to support the replication of scientific results from raw and processed data include Linked Paleo Data (LiPD), which is a part of the EarthCube-supported Linked Earth (http://linked.earth/) project and provides the flexibility to deal with different paleoclimate sensors, archives, and observations, divided into four broad categories of data and metadata: georeferencing and spatial context, publication record, interpretive proxy system (data) models, and chronology models (McKay & Emile-Geay, 2016; PAGES2k Consortium, 2017). Other examples of publicly stored digital data and metadata in use by the paleoenvironmental research community include observational databases managed by the Neotoma Paleoecology Database and Pangeo, the National Center for Environmental Information, and the World Data Center for Climate at DRKZ. Code sharing is becoming more common, for instance open-source LiPD-linked tools and PRoxY Systems Modeling (PRYSM) tools are available on GitHub (Dee et al., 2015), and code for the calculations in Sherwood et al. (2020) is available in a DOI-linked repository at Zenodo. The NSF Paleoclimate and Paleo Perspectives on Climate Change (P2C2) programs fund much of the relevant, ongoing paleoclimate research in the US, with its goals of generating proxy datasets that can serve as tests for climate models and synthesizing proxy and model data to understand longer-term and higher-magnitude climate system variability not captured by the instrumental record.

A frequent challenge in replicating and synthesizing paleoclimate studies is incomplete reporting or archiving of data or metadata, due in part to a lack of community-wide standards and awareness of the potential utility of metadata for further studies, and a lack of enforcement by reporting journals or funding agencies. One measure to address this challenge is the publication and dissemination of data-reporting standards across the natural sciences (e.g., the National Science Foundation's Data Management Plan requirements within its directorates) and within disciplines. A recent example of the latter within radiometric chronology in paleoclimatology is a report on uranium-thorium (U-Th) dating measurements (Dutton et al., 2017) that prescribes necessary procedures and the data and metadata required for repositories to increase the utility and longevity of study data. Similar guidelines have been devised for geochemistry data and for other sub-disciplines (Deines et al., 2003; Renne et al., 2009). The NSF Data Infrastructure Building Blocks (DIBBs) program supports the development of robust and shared data-centric cyberinfrastructure, such as Whole Tale (Brinckman et al., 2019). DIBBs also supports work such as developing the capability to seamlessly upload uranium measurement data directly from the analytical instrument to online repositories such as those managed by the Interdisciplinary Earth Data Alliance (IEDA) in collaboration with programs such as NSF EarthCube for visualization and analysis. Although this type of standardization enhances the utility of future work, there remains the challenge of incorporating older work published before methods standardization, i.e., "legacy data" (Deevey and Flint, 1959), into relevant online repositories (Liu et al., 2019).

In order to address issues associated with enforcement of standards and best practices, there must be an increase in awareness of needs and benefits, including broadly disseminated best practice reports among journals and agencies, which can then set policies based on those best practices. Additionally, if there is a lack of an international standard for analysis, it can be difficult to transition the community to a new practice or standard. For example, among laboratories that conduct U-Th dating measurements, there is no common standard for Th activity ratios. Many laboratories rely either on individually developed gravimetric standards or a standard known as HU-1, that was distributed to many labs around the globe, but with increased instrument precision, it has been discovered that different aliquots have measureable differences. In other words, instrumental precision is now outstripping the accuracy of the standard being used (Dutton, 2015). Furthermore, there has been refinement in U-Th decay constant values, meaning that ages published using different decay constants cannot be directly compared with one another (Cheng et al., 2013). This is a challenge common to many paleoclimate proxies, for which scientific progress may render age models and reconstructions obsolete. Paleoclimate data are often archived with assigned ages and not with the underlying geochronological data, which may include radiometric dates by sample sequence, tie-point age modeling, inter-date age interpolation, stratigraphic constraints, and dating uncertainty (Cronin, 2009). In the absence of these geochronological metadata, it is difficult to use new information, study replicates, and observing networks to revise age models, degrading the potential utility of earlier results. As a means of addressing this issue for the particular example at hand in U-Th dating, a cyberinfrastructure project was supported by NSF DIBBs, building upon the success of software developed for the U-Pb community through the EarthTime initiative such that the open-source software to which measurements are uploaded will make calculations based on the new decay constants (Dutton et al., 2017). There is also an effort to distribute new analytical standard material to labs around the world. However, database development that brings together multiple proxies remains difficult, and incentives are misaligned where data uploading archiving remains a challenge and a burden to the producer of the data, while it is a benefit to others. Furthermore, there remains the challenge that current data repositories freeze in place ‘as-published’ data that rapidly become obsolete and are not automatically machine-readable.

New initiatives such as the LiPD project are addressing some of these challenges by setting standards and building structures for putting more complete metadata into paleoclimate databases and repositories so that they can be continuously accessed, reassessed and updated. There are also a number of efforts to produce ‘intelligent’ repositories that could update age models, account for uncertainties in age and interpretation, and recalculate syntheses interactively. For example, the new Phantastic project, led by the Smithsonian Institution, aims to build a temperature record for the entire Phanerozoic Eon, 500 million years long, taking into account the considerations listed above. An NSF-funded Research Coordination Network, Improving Reconstructions of Cenozoic pCO2 and Temperature Change, led by Baerbel Hoenisch and Pratigya Polissar, aims to achieve similar objectives for the Cenozoic history of atmospheric CO_2 concentrations. Last, stacked paleoclimate records such as that generated by Lisiecki and Raymo (2005) or Snyder (2016) can successfully synthesize multiple proxy records for a single time interval. All of these initiatives would not be possible without publicly available data and metadata.

3. Global Climate Models

3.1. Reproducibility and Replicability in Climate Modeling

For global climate models (GCMs), computational reproducibility refers to the ability to re-run a model with a given set of initial conditions and produce the same results with subsequent runs. This is achievable within the short time spans and individual locations and is essential for model testing and software debugging, but the dominance of this definition as a paradigm in the field is giving way to a more statistical way of understanding model output. Historically, climate modelers felt that they needed the more rigid definition of bitwise reproduction because the non-linear equations governing Earth systems are chaotic and sensitive to initial conditions. However, this numerical reproducibility is difficult to achieve with the computing arrays required by modern GCMs. Global climate models also have a long history of occurrences in the models that have caused random errors and have never been reproduced, including possible cosmic ray strikes (Hansen et al., 1984) and other reported events in uncontrolled model runs that may or may not be the result of internal model variability or software problems (e.g., Hall & Stouffer, 2001; Rind et al., 2018). Reproducing the conditions that cause these random events is difficult, and our lack of understanding of their effects undermines the scientific conclusions of the model. Features of computer architecture that undermine the ability to achieve bitwise reproducibility include fused multiply-add, which cannot preserve order of operations, memory details, and issues of parallelism when a calculation is divided across multiple processors. Some of these features, e.g., parallelism, are testable and the resulting irreproducibility reduced with additional care. Moreover, the environment in which GCMs are run is fragile and ephemeral on the scale of months to years, as compilers, libraries, and operating systems are continually updated, such that revisiting a 10 year-old study would require an impractical museum of supercomputers or the additional framework of an environment emulator. Retaining bitwise reproducibility will become even more difficult in the near future as nondeterministic machine-learning algorithms and neural networks are introduced (e.g., Chadwick et al., 2011). There is also interest in representing stochasticity in the physical models by harnessing noise inherent within the electronics, and some current devices have mixed or variable bit precision. Last, cosmic ray strikes are a real source of undetected error, where a high-energy particle strikes and alters a single bit in what is known as single-event upset, a phenomenon that is increasingly likely with altitude and is well-studied at Los Alamos, elevation 2230 m (Normand et al., 2010; Sridharan et al., 2013). Therefore, the focus of the discipline has not been on model run reproducibility, but rather on replication of model phenomena observed and their magnitudes, which is performed mostly in organized multi-model ensembles.

One of the main multi-model ensembles is the Coupled Model Intercomparison Project (CMIP), which has evolved through several iterations and is currently at CMIP6. The projects in CMIP are community-driven standardized simulations with common, publicly available outputs, especially via the Earth System Grid Federation (ESGF), headquartered at the US Department of Energy. CMIP has enjoyed complete buy-in from all global modeling groups for over a decade, and has set the standard by which all models are tested and diagnostics produced. One of the major challenges of CMIP is that it exists via an unfunded mandate and relies heavily on donated time and work from modeling groups. Furthermore, the CMIP projects have become increasingly massive and complex, with CMIP6 estimated to generate ~100 PB (petabytes, or ~100 million Megabytes) of data. Across the CMIP ensemble, it is easy to test the replicability and robustness of results and identify common elements across models, but more difficult to interpret those results, e.g., changes in precipitation patterns. Similarly, the resulting data is stored and accessible online, but the capacity to analyze that data remains limited as the ability to perform complex multivariate analyses is limited by bandwidth. There is also no support for archiving derived or intermediate data or analysis code, which limits reproducibility, and as yet little in the way of server-side analytics. New projects, such as Pangeo, which promises scalable sharing of PB-scale datasets, may provide solutions to these problems.

To track and combat error sources, the Community Earth System Model (CESM) at UCAR has in place a series of elements to the climate model pipeline, one of the most central of which is the model 'case,' which is a standard piece of metadata that describes source code versions and tags a model output with the input datasets and other settings. There are also considerations of the run-time configuration and levels of parallelism in distributed components across processors. The compiler and libraries in the computer architecture may be ephemeral, but researchers can process the model output with standard diagnostics to produce very similar results. CESM cases can be thought of as experimental sandboxes, and they can be documented within a database. These calculations and database information need to be shared more widely and made publicly available, but standard diagnostic packages can be attached to a case so that its workflow is also reproducible.

Most GCMs require bit-for-bit reproducibility for restart purposes to maintain functionality of model simulations. However, threading and MPI tasks are different forms of parallelism and component layout, and these can introduce reproducibility issues to various parts of the climate model. In moving towards a statistical approach in dealing with issues of reproducibility, groups such as CESM and the Met Office Hadley Centre for Climate Change in the UK assesses whether data that is changed from the original data due to ephemeral errors during run time is statistically distinguishable from the original. As part of this, CESM has developed an Ensemble Consistency Test (ECM), which begins with an accepted ensemble of data coming from a model that is considered 'correct.' The variability in that model data is quantified, typically involving principal component analysis. New climate simulations may then be created, and the comparison between the leading principal components with those from the 'correct' simulations allows for a statistical discrimination between results that do or do not belong with the original ensemble. Downstream of the model output, users are interested in generating derived products with the simulations, and issues of reproducibility similar to those described above are propagated. This is an area of high need for development of tools, procedures, and repositories to better connect a published artifact to the conditions that produced it.

3.2. Preservation and Analysis of GCM Results

Archiving the very large datasets generated by climate models is necessary but comes with a high cost associated with the maintenance of petabytes of data and storing them in such a way that they are both accessible and usable. Data is rarely deleted, and as computing power increases, future models will only generate more data more quickly. Lossy data compression, in which replication but not reproducibility is preserved, can potentially serve as a means of reducing the cost of preservation and making data more easily accessible. However, it must be determined which information can be safely lost and which must be preserved so that results and scientific progress are not negatively impacted. NCAR has experimented with a tool, known as fpzip, that gives an average 5x compression factor across climate variables, but with favorable statistical similarity scores to the 'true' output. Some of the complications with lossy compression involve the fact that not all variables are equally compressible and not all variables respond in the same way to compression, and so it may be necessary to apply a very complex algorithmic filter such that the correct compression algorithm is applied to each climate variable. Furthermore, it remains unclear how this process can be documented and made reproducible. This is an area of active investigation developing compression metrics to compare compressed and uncompressed data products and assessing differences both visually and quantitatively. Ultimately, the goal of this endeavor is to make the maintenance of these repositories more affordable via the application of statistics and effective emphasis on replication, overriding the much more expensive strong form of reproducibility, i.e., its original, more rigid definition as bitwise reproducibility.

The analytics of reproducibility requires improvement. This must begin with making datasets and associated parameters not just accessible but also easily discoverable in active, intelligent public repositories where data (both raw and derived), tools, and code can be linked and updated. Making data and tools citable, tagging them with DOIs, and referencing them in publications—steps that are easily actionable for many studies—certainly also helps with accessibility (Stodden et al., 2016). There are many standard toolkits, including netCDF Operators (NCO), Community Data Analysis Tools (CDAT), and libraries for code in Python, R, Matlab, and IDL languages that have code repositories as well places like GitHub, as well as example efforts to share open data science tools (Lowndes et al., 2017). However, archiving of analysis code is haphazard and difficult to search, often being associated variously with publication journals, institutions, or with individual researchers as opposed to being stored in a permanent DOI-linked facility. Peer review can also be extended to tools such as Jupyter notebooks, although journal publications alone are not sufficient to capture peer review of modern big data studies and the current developments in the field of big data science. Most GCMs have public releases of frozen codes or snapshots of code, e.g. from CMIP, NCAR, and others, but no GCM is truly open-source and experimental and developmental versions of model code are not always made available, which makes the benefit of publicly known tags for non-public model versions unclear. There is also no standardized repository for specific model versions, although this is an option currently being explored by the NASA Goddard Institute for Space Sciences (GISS).

Downstream analyses of these extremely large model datasets will themselves require parallel computing, which means that all of the challenges in maintaining reproducibility in parallel computing of the original models may be recapitulated to some degree in parallel analyses of model results. Analytic tools must also be made more accessible, and platforms that may be able to assist in this include PanGeo. Last, while operational bitwise reproducibility is necessary for code development, long-term reproducibility is not, and it will be necessary to revisit the paradigm of bitwise reproducibility and why that should be expected from complex computer models of chaotic natural systems, given that experimental studies of other natural systems (e.g., in chemistry) do not expect exact, bit-for-bit reproducibility of their results and rely on statistical estimates for making inference.

4. Interfacing Paleoclimate Proxies and Modern Climate Science

4.1. Climate Science Data Repositories, Their Use for GCMs, and Associated Challenges

Climate science currently has three massive data streams: 1) remote sensing from satellites operated by NASA, NOAA, ESA, Japan, etc., which produce continuous streams of global, multivariate data; 2) weather forecast and hindcast analyses and re-analyses, where highly detailed forecasts are generated every six hours; and 3) coupled GCMs, which are producing as much data as the current supercomputers will allow. Almost all of the raw data from these data streams is available publicly in some form, but there do not exist joint repositories or storage of derived data. This gap in the ability to combine data is major and is hampering the rate of scientific progress in the field, for example in bias estimation, model initialization, and forecast improvement.

An example of operational data products, GISS Surface Temperatures (GISTEMP) was originally developed in 1981 and has undergone continual expansion and improvement since that time. GISTEMP only uses publicly available data, and its analysis code has been available online since 2007. The analysis code was re-coded in a more modern language by an external company after its release by GISS, which demonstrates the benefit of citizen science when code is made available to the public and illustrates practical reproducibility and replicability of its results. GISTEMP recalculates the homogenization every month based on new data; thus as the input data grows and methods change with time, this generates a resulting historical change in the product with time. When examining global mean temperature over time from 1880 to present, there is an increase in noise and uncertainty with increasing age (Lenssen et al., 2019), but our understanding of the progression of global mean temperature over the history of instrumental records is quite robust. That replication is confirmed via comparisons with independently calculated datasets, including reanalyses (Kalnay et al., 1996; Kistler et al., 2001; Compo et al., 2011) and satellite products (Susskind, Schmidt, Lee, & Iredell, 2019).

4.2. Climate Sensitivity from Paleoclimate Records to GCMs

The example considered here is the study of equilibrium climate sensitivity, which estimates the climate temperature response to external radiative forcing after allowing the global climate system to come to equilibrium. Recent research has found that estimates of the equilibrium climate sensitivity vary between consideration of fast feedbacks only (on the order of minutes to months) and of fast as well as slow feedbacks, on the order of years to millennia and including factors such as ocean circulation (Hansen et al., 1985; Rohling et al., 2012). Most studies of recent paleoclimate provide only low-CO_2 climates, but Rohling et al. (2012) examined paleoclimate studies from deeper time that included high-CO_2 climates and found that inclusion of slow feedback mechanisms significantly raised the temperature response of the climate system. With the addition of multiple studies of the same phenomenon, i.e. study replication, the authors also demonstrated the robustness of the results. Equilibrium climate sensitivity estimates are a useful test of climate models, as paleoclimate proxy records provide evidence of climate systems that have not and cannot be directly observed; verified estimates can then applied to model forecasts of future climate.

5. Major Recommendations

5.1. Paleoclimate Data

We support the development of interactive, intelligent, multi-proxy paleoclimate data repositories that are publicly accessible, where datasets are linked to other relevant information (e.g., geospatial information and chronology models) and all datasets are updatable as models change and improvements are made. Data repositories should include native observations and derived estimates, enriched metadata, and the ability to ingest legacy formats, emerging data structures, and analytic code. Repositories should be designed to enable large-scale data synthesis, building on the work of initiatives such as Linked Paleo Data (LiPD), Linked Earth , PRYSM (Dee et al., 2015), and Neotoma Paleoecology. The inclusion of raw as well as derived chronological data is important for updating assigned chronologies as the underlying databases are improved. Repositories should be institutionally supported and may be best arranged along major themes such as specific events or periods (e.g., the Paleocene-Eocene Thermal Maximum, Last Glacial Maximum, Quaternary, or Pliocene) or recurring phenomena (e.g., orbital cycles, Dansgaard-Oeschger variability, or patterns such as El Niño-Southern Oscillation).

5.2. Global Climate Model (GCM) Simulations

We believe development of community standards and archives for GCM simulations should be focused on issues of *replicability* of model results under similar experimental conditions, building on the best practices established by groups such as the Coupled Model Intercomparison Project (CMIP). Accordingly, GCM groups should organize database entries around “cases”, in which the experimental conditions (source code, compilation environment, model configuration, input data sources, etc.) that give rise to specific simulation ensembles are documented, versioned and made publicly available via established repositories. GCM groups, including NCAR, CESM, and GISS, already keep track of specific code bases and configurations for any specific production run using local ‘cases’ that allow for reproducibility of output in the short term, but there are no cross-center standards for this. Because of the sheer magnitude of storage required for these efforts, further work to develop improved data storage, such as lossy compression and loss diagnostics, should be pursued. This will involve further testing and standardization around questions of what data and metadata should be preserved and for how long. Storage costs for large datasets are not negligible (currently around $12K/PB/yr) and the increase in model data output is growing much faster than costs of storage are falling (for example, see the IPCC Task Group on Data Support for Climate Change).

5.3. Incentives and Awareness

Supporting organized, intelligent repositories requires sufficient and long-term funding, as well as the alignment of incentives regarding who does the work and who benefits from it. Replicability carries a cost to individual researchers who would not be conducting *de novo* experiments while replicating existing work. Organization and standardization of repositories should be designed to present the lowest possible barriers to dataset submission for both observations and simulations (Konkol et al., 2020). They should also provide additional and immediate benefits to the contributor by providing digital object identifiers (DOIs) for datasets, code, cases, replication exercises, discussion papers and comments on publications during and following review (Stodden et al., 2016). Journals and funding agencies can improve awareness and adherence by encouraging best practices in which these recommendations reside (Stodden et al., 2018), such as the findable, accessible, interoperable and reusable (FAIR) principles. In fact, this recommendation was recently made in the community decadal survey report for Earth Sciences at the National Science Foundation (NASEM, 2020), and journals published by the American Geophysical Union (AGU), including *Paleoceanography and Paleoclimatology*, have adopted FAIR guidelines.

Standardization and oversight of the application of paleoclimate proxy standards is also important and in different stages of development for different proxies. This requires the dissemination of new standards and information, methods and best practices for applying them, and incentivizing the use and acceptance of globally applied standards and methods. To address issues associated with enforcement of standards and best practices, there must be an increase in awareness, including broadly disseminated best practice reports among journals and agencies, which can then set policies based on those best practices.

6. Conclusions

Paleoclimate records serve multiple purposes in contextualizing, calibrating, and testing modern climate observations and models. Different paleoclimate proxies are at different stages of development, which entails different degrees of error propagation from modern calibration studies, proxy measurements, and timescale measurements, and different stages of standardization among the sub-disciplines. In paleoclimate proxy studies, practices of reproducibility in sample measurement and calibration against accepted standard materials are generally well-established. Observational reproduction within and across paleoclimate sites improves our estimation of amplitude and chronological uncertainties. Practices of error propagation and uncertainty estimation are crucial for determining replication and will continue to benefit from international standardization and development of public repositories that support synthesis activities and meta-analyses. Reproducibility of analyses of large multivariate databases is improving as databases, metadata and analysis codes are being made publicly available. These enriched databases will accelerate progress, for instance in geochronology, as legacy data are refined, geochronometric methods improve, and new, more precise and accurate data are added.

In global climate models, bitwise or computational reproducibility may be relatively straightforward to demonstrate for short simulations and within computing centers. However, distributed computing makes reproducibility difficult, as do subsequent parallel analytics of model results, and may be difficult or impossible for complex models and as analytical methods such as machine learning are introduced. Consequently, global climate modeling groups have de-emphasized bitwise reproducibility in favor of replicability across ensembles generated from similar and analog experimental conditions. As computing speeds and dataset sizes increase, methods of data storage and access will require novel storage algorithms, such as lossy data compression, which must be tested for reproducibility but ultimately will support detection of replicability. Replication is feasible within multi-model ensembles such as are produced by the CMIP initiative. Replication of equilibrium climate sensitivity estimates is an important test of climate projection-based estimates because paleoclimate data may reflect climate system processes operating over long time scales. The limits of code-sharing (Greene and Thirumalai, 2019) and inclusion of innovative new tools and access platforms (Yin et al., 2019) are worth further consideration, but the state of both paleoclimate and climate science would benefit greatly from the development of active repositories of linked data, tools, and code that are easily accessible and searchable, as well as an alignment of incentives for participation in and maintenance of those repositories.

Disclosure Statement

Funding for the initial writing of this manuscript was provided by the National Academies of Science, Engineering, and Medicine.

Acknowledgements

We thank the National Academies for supporting the conference and initial writing of this manuscript, and four anonymous reviewers for their constructive comments and feedback, which greatly improved the manuscript.

References

Abram, N. J., McGregor, H. V., Tierney, J. E., Evans, M. N., McKay, N. P., Kaufman, D. S., … & Phipps, S. J. (2016). Early onset of industrial-era warming across the oceans and continents. *Nature, 536*(7617), 411-418. https://doi.org/10.1038/nature19082

Alley, R. B., Gow, A. J., Johnsen, S. J., Kipfstuhl, J., Meese, D. A., & Thorsteinsson, Th. (1995). Comparison of deep ice cores. *Nature, 373*, 393. https://doi.org/10.1038/373393b0

Anand, P., Elderfield, H., & Conte, M. H. (2003). Calibration of Mg/Ca thermometry in planktonic foraminifera from a sediment trap time series. *Paleoceanography and Paleoclimatology, 18*(2), 1050. https://doi.org/10.1029/2002PA000846

Bollen, K., Cacioppo, J. T., Kaplan, R., Krosnick, J., & Olds, J. L. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science.* Arlington, VA: National Science Foundation.

Bond, G., Kromer, B., Beer, J., Muscheler, R., Evans, M. N., Showers, W., . . . Bonani, G. (2001). Persistent solar influence on North Atlantic climate during the Holocene. *Science*, *294*(5549), 2130-2136. https://doi.org/10.1126/science.1065680

Bradley, R.S. (2014). *Paleoclimatology: Reconstructing Climates of the Quaternary.* New York: Academic Press, 3rd ed, ISBN 978-0-12-386913-5

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., … & Turner, K. (2019). Computing environments for reproducibility: capturing the "Whole Tale". *Future Generation Computer Systems* 94, 854-867. https://doi.org/10.1016/j.future.2017.12.029

Chadwick, R., Coppola, E., & Giorgi, F. (2011). An artificial neural network technique for downscaling GCM outputs to RCM spatial scale. *Nonlinear Processes in Geophysics, 18*(6), 1013-1028. https://doi.org/10.5194/npg-18-1013-2011

Cheng, H., Edwards., R.L., Shen, C.-C., Polyak, V.J., Asmerom, Y., Woodhead, J., ... & Alexander, E.C. (2013). Improvements in ^{230}Th dating, ^{230}Th and ^{234}U half-life values, and U-Th isotopic measurements by multi-collector inductively coupled plasma mass spectrometry. *Earth and Planetary Science Letters 371-372,* 82-91. https://doi.org/10.1016/j.epsl.2013.04.006

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., ... Bessemoulin, P. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society, 137*(654), 1-28. https://doi.org/10.1002/qj.776

Coplen, T.B., Hopple, J.A., Böhlke, J.K., Peiser, H.S., Rieder, S.E., Krouse, H.R., ... & De Bièvre, P. (2002.) *Compilation of minimum and maximum isotope ratios of selected elements in naturally occurring terrestrial materials and reagents: U.S. Geological Survey Water-Resources Investigations Report* 01-4222, 98p. https://doi.org/10.3133/wri014222

Cronin, T. M. (2009). *Paleoclimates: understanding climate change past and present.* Columbia University Press.

Cuffey, K.M. & Clow, G.D. (1997). Temperature, accumulation, and ice sheet elevation in central Greenland through the last deglacial transition. *Journal of Geophysical Research: Oceans, 102*(C12), pp.26383-26396. https://doi.org/10.1029/96JC03981

de Groot, P. (ed.). (2004). *Handbook of Stable Isotope Analytical Techniques, Volume I.* Amsterdam, Netherlands: Elsevier, 1258pp.

Dee, S., Emile-Geay, J., Evans, M., Allam, A., Steig, E., & Thompson, D. (2015). PRYSM: An open-source framework for PRoxY System Modeling, with applications to oxygen-isotope systems. *Journal of Advances in Modeling Earth Systems,* 7(3), 1220-1247. https://doi.org/10.1002/2015MS000447

Deevey, E., & Flint, R. (1959). Preface. *Radiocarbon, 1,* https://doi.org/10.1017/S0033822200020312

Deines, P., Goldstein, S.L., Oelkers, E.H., Rudnick, R.L., & Walter, L.M. (2003). Standards for publication of isotope ratio and chemical data in Chemical Geology. *Chemical Geology 202*(1-2), 1-4. https://doi.org/10.1016/j.chemgeo.2003.08.003

DeLong, K. L., Quinn, T. M., Taylor, F. W., Shen, C.-C., & Lin, K. (2013). Improving coral-base paleoclimate reconstructions by replicating 350 years of coral Sr/Ca variations. *Palaeogeography, Palaeoclimatology, Palaeoecology, 373*, 6-24. https://doi.org/10.1016/j.palaeo.2012.08.019

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science. *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*. Montreal, Canada.

Dutton, A. (2015). U-Th Dating, in *Handbook of Sea Level Research*, eds. Shennan, I., Long, A., & Horton, B. John Wiley & Sons, Ltd., 386-403.

Dutton, A., Carlson, A., Long, A., Milne, G., Clark, P., DeConto, R., . . . Raymo, M. (2015). Sea-level rise due to polar ice-sheet mass loss during past warm periods. *Science, 349*(6244). https://doi.org/10.1126/science.aaa4019

Dutton, A., Rubin, K., McLean, N., Bowring, J., Bard, E., Edwards, R., . . . Sims, K. (2017). Data reporting standards for publication of U-series data for geochronology and timescale assessment in the earth sciences. *Quaternary Geochronology, 39*, 142-149. https://doi.org/10.1016/j.quageo.2017.03.001

Evans, M.N., Tolwinski-Ward, S.E., Thompson, D.M. & Anchukaitis, K.J. (2013). Applications of proxy system modeling in high resolution paleoclimatology. *Quaternary Science Reviews, 76*, 16-28. https://doi.org/10.1016/j.quascirev.2013.05.024

Fraser, H., Barnett, A., Parker, T.H., & Fidler, F. (2020). The role of replication studies in ecology. *Ecology and Evolution 10,* 5197-5207. https://doi.org/10.1002/ece3.6330

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean?. *Science Translational Medicine, 8*(341). https://doi.org/10.1126/scitranslmed.aaf5027

Greene, C. A., & Thirumalai, K. (2019). It's time to shift emphasis away from code sharing. *EOS*, 100, https://doi.org/10.1029/2019EO116357.

Hall, A., & Stouffer, R. J. (2001). An abrupt climate event in a coupled ocean–atmosphere simulation without external forcing. *Nature, 409*(6817), 171. https://doi.org/10.1038/35051544

Hansen, J., Lacis, A., Rind, D., Russell, G., Stone, P., Fung, I., . . . Lerner, J. (1984). Climate sensitivity: Analysis of feedback mechanisms. *Climate processes and climate sensitivity, 5*, 130-163.

Hansen, J., Russell, G., Lacis, A., Fung, I., Rind, D., & Stone, P. (1985). Climate response times: Dependence on climate sensitivity and ocean mixing. *Science, 229*(4716), 857-859. https://doi.org/10.1126/science.229.4716.857

Johnsen, S. J., Clausen, H. B., Dansgaard, W., Fuhrer, K., Gundestrup, N., Hammer, C. U., ... & Steffenon, J. P. (1992). Irregular glacial interstadials recorded in a new Greenland ice core. *Nature, 359*, 311-313. https://doi.org/10.1038/359311a0

Jones, P. D., Briffa, K., Osborn, T., Lough, J., Van Ommen, T., Vinther, B., . . . Mann, M. (2009). High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *The Holocene, 19*(1), 3-49. https://doi.org/10.1177/0959683608098952

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., . . . Woollen, J. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society, 77*(3), 437-472.

Kaufman, D., McKay, N., Routson, C., Erb, M., Dätwyler, C., Sommer, P.S., ... & Davis, B. (2020). Holocene global mean surface temperature, a multi-method reconstruction approach. *Scientific Data,* 7(1), pp.1-13. https://doi.org/10.1038/s41597-020-0530-7

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., . . . Kousky, V. (2001). The NCEP–NCAR 50-year reanalysis: monthly means CD-ROM and documentation. *Bulletin of the American meteorological Society, 82*(2), 247-268.

Konkol, M., Kray, C., & Pfeiffer, M. (2018). Computational reproducibility in geoscientific papers: insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science, 33*(2), 408-429. https://doi.org/10.1080/13658816.2018.1508687

Konkol, M., Nüst, D., and Goulier, L. (2020). Publishing computational research—a review of infrastructures for reproducible and transparent scholarly communication. *Research Integrity and Peer Review* 5, 10. https://doi.org/10.1186/s41073-020-00095-y.

Kopp, R.E., Kemp, A.C., Bittermann, K., Horton, B.P., Donnelly, J.P., Gehrels, W.R., ... & Rahmstorf, S. (2016). Temperature-driven global sea-level variability in the Common Era. *Proceedings of the National Academy of Sciences 113*, E1434–E1441. https://doi.org/10.1073/pnas.1517056113

Kucera, M., Schneider, R., & Weinelt, M. (2006). *MARGO – Multiproxy Approach for the Reconstruction of the Glacial Ocean surface.* Amsterdam, Netherlands: Elsevier Science.

Lenssen, N., Schmidt, G., Hansen, J., Menne, M., Persin, A., Ruedy, R., & Zyss, D. (2019). Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres, 124*, 6307-6326. https://doi.org/10.1029/2018JD029522

Linsley, B. K., Wu, H. C., Dassie, E. P., & Schrag, D. P. (2015). Decadal changes in South Pacific sea surface temperatures and the relationship to the Pacific decadal oscillation and upper ocean heat content. *Geophysical Research Letters, 42*(7), 2358-2366. https://doi.org/10.1002/2015GL063045

Lisiecki, L. E., & Raymo, M. E. (2005). A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography and Paleoclimatology, 20*(1), PA1003. https://doi.org/10.1029/2004PA001071

Liu, Z., J. Wang, S. Pan, & D. Meyer. (2019). Improving reproducibility in Earth science Research. *Eos, 100*, https://doi.org/10.1029/2019EO136216.

Lowndes, J.S.S., Best, B.D., Scarborough, C., Afflerbach, J.C., Frazier, M.R., O'Hara, C.C., ... & Halpern, B.S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution 1,* 0160. https://doi.org/10.1038/s41559-017-0160

McKay, N., & Emile-Geay, J. (2016). Technical note: The Linked Paleo Data framework—a common tongue for paleoclimatology. *Climate of the Past, 12*, 1093–1100. https://doi.org/10.5194/cp-12-1093-2016

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science.* Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

National Academies of Sciences, Engineering, and Medicine. (2020). *A Vision for NSF Earth Sciences 2020-2030: Earth in Time.* Washington, DC: The National Academies Press. https://doi.org/10.17226/25761

National Research Council. (2006). *Surface Temperature Reconstructions for the Last 2000 Years.* Washington, DC: The National Academies Press. https://doi.org/10.17226/11676

National Research Council. (2012). *A National Strategy for Advancing Climate Modeling.* Washington, DC: The National Academies Press. https://doi.org/10.17226/13430

Neukom, R., Steiger, N., Gómez-Navarro, J. J., Wang, J., & Werner, J. P. (2019). No evidence for globally coherent warm and cold periods over the preindustrial Common Era. *Nature, 571*(7766), 550-554. https://doi.org/10.1038/s41586-019-1401-2

Normand, E., Wert, J. L., Quinn, H., Fairbanks, T. D., Michalak, S., Grider, G., ... Johnson, S. (2010). First record of single-event upset on ground, cray-1 computer at Los Alamos in 1976. *IEEE Transactions on Nuclear Science, 57*(6), 3114-3120. https://doi.org/10.1109/TNS.2010.2083687

Nüst, D. & Pebesma, E. (2020). Practical Reproducibility in Geography and Geosciences. *Annals of the American Association of Geographers*. https://doi.org/10.1080/24694452.2020.1806028

PAGES2k Consortium. (2013). Continental-scale temperature variability during the past two millennia. *Nature Geoscience, 6*(5), 339-346. https://doi.org/10.1038/ngeo1797

PAGES2k Consortium. (2017). A global multiproxy database for temperature reconstructions of the Common Era. *Scientific Data, 4*, 170088. https://doi.org/10.1038/sdata.2017.88

Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J. –M., Basile, I., ... & Stievenard, M. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica.

Nature, 399, 429-436. https://doi.org/10.1038/20859

Plesser, H. E. (2018). Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in Neuroinformatics, 11*, 76. https://doi.org/10.3389/fninf.2017.00076

Potter, E.-K., & Lambeck, K. (2004). Reconciliation of sea-level observations in the Western North Atlantic during the last glacial cycle. *Earth and Planetary Science Letters, 217*(1-2), 171-181. https://doi.org/10.1016/S0012-821X(03)00587-9

Powers, S.M. & Hampton, S.E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications, 29*(1), e01822. https://doi.org/10.1002/eap.1822

Renne, P.R., Deino, A.L., Hames, W.I., Heizler, M.T., Hemming, S.R., Hodges, K.V., ... & Wijbrans, J.R. (2009). Data reporting norms for ^{40}Ar/^{39}Ar geochronology. *Quaternary Geochronology* 4(5), 346-352. https://doi.org/10.1016/j.quageo.2009.06.005

Rind, D., Schmidt, G. A., Jonas, J., Miller, R., Nazarenko, L., Kelley, M., & Romanski, J. (2018). Multicentury instability of the Atlantic meridional circulation in rapid warming simulations with GISS ModelE2. *Journal of Geophysical Research: Atmospheres, 123*(12), 6331-6355. https://doi.org/10.1029/2017JD027149

Rohling, E., Sluijs, A., Dijkstra, H., Köhler, P., Van de Wal, R., Von Der Heydt, A., . . . Crucifix, M. (2012). Making sense of palaeoclimate sensitivity. Nature, 491(7426), 683. https://doi.org/10.1038/nature11574

Rohling, E.J., Hibbert, F.D., Grant, K.M., Galaasen, E.V., Irvalı, N., Kleiven, H.F., ... & Schulz, H. (2019). Asynchronous Antarctic and Greenland ice-volume contributions to the last interglacial sea-level highstand. *Nature Communications, 10*(1), 1-9. https://doi.org/10.1038/s41467-019-12874-3

Schmidt, G.A. (2010). Enhancing the relevance of palaeoclimate model/data comparisons for assessments of future climate change. *Journal of Quaternary Science, 25*(1), 79-87. https://doi.org/10.1002/jqs.1314

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., ... & Zelinka, M. D. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics, 58,* e2019RG000678. https://doi.org/10.1029/2019RG000678

Snyder, C. W. (2016). Evolution of global temperature over the past two million years. *Nature, 538*(7624), 226. https://doi.org/10.1038/nature19798

Sridharan, V., Stearley, J., DeBardeleben, N., Blanchard, S., & Gurumurthi, S. (2013). Feng shui of supercomputer memory: positional effects in DRAM and SRAM faults. *SC '13: Proceedings of the*

International Conference on High Performance Computing, Networking, Storage and Analysis, Denver, CO, 2013, pp. 1-11.

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P.A., & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science, 354*(6317), 1240-1241. https://doi.org/10.1126/science.aah6168

Stodden, V., Seiler, J., & Ma, Zhaokun. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences 115*(11), 2584-2589. https://doi.org/10.1073/pnas.1708290115

Susskind, J., Schmidt, G., Lee, J., & Iredell, L. (2019). Recent global warming as confirmed by AIRS. *Environmental Research Letters, 14*(4), 044030. https://doi.org/10.1088/1748-9326/aafd4e

Wang., Y. J., Cheng, H., Edwards, R. L., An, Z. S., Wu, J. Y., Shen, C. –C., & Dorale, J. A. (2001). A high-resolution absolute-dated Late Pleistocene monsoon record from Hulu Cave, China. *Science, 294*(5550), 2345-2348. https://doi.org/10.1126/science.1064618

Wilson, E.O. (1999). *Consilience: The Unity of Knowledge.* New York:Vintage, p. 384

Yin, D., Liu, Y., Hu, H., Terstriep, J., Hong, X., Padmanabhan, A., & Wang, S. (2019).

CyberGIS-Jupyter for reproducible and scalable geospatial analysis. *Concurrency and Computation: Practice and Experience, 31,* e5040. https://doi.org/10.1002/cpe.5040

Zachos, J., Pagani, M., Sloan, L., Thomas, E., & Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science, 292*(5517), 686-693. https://doi.org/10.1126/science.1059412