This paper was commissioned for the Committee on Reproducibility and Replicability in Science, whose work was supported by the National Science Foundation and the Alfred P. Sloan Foundation. Opinions and statements included in the paper are solely those of the individual author, and are not necessarily adopted, endorsed, or verified as accurate by the Committee on Reproducibility and Replicability in Science or the National Academies of Sciences, Engineering, and Medicine.

Perspectives on Reproducibility and Replication of Results in Climate Science

FINAL DRAFT for July 17, 2018

drafted by Rosemary T. Bush (Northwestern University)

Summary

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 some of the challenges for the future with recommendations for addressing those challenges. This paper is based largely on the presentations of a panel of researchers on May 9, 2018: Michael Evans (Associate Professor in the Department of Geology and Earth System Science Interdisciplinary Center, University of Maryland), Gavin Schmidt (Director of the Goddard Institute for Space Studies, National Aeronautics and Space Administration), Rich Loft (Director of the Technology Development Division, National Center for Atmospheric Research), and Andrea Dutton (Assistant Professor in the Department of Geological Sciences, University of Florida). See Appendix A for the agenda from this meeting.

Major Recommendations

Paleoclimate data archives

Support the development of interactive, intelligent, multi-proxy paleoclimate data archives 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. Archives should include analytic code and raw as well as derived data, develop capacity for incorporating legacy data, and allow for large-scale data synthesis, building on the work of current archives such as Linked Paleo Data (LiPD), Linked Earth, PRYSM, and Neotoma Paleoecology. The inclusion of raw as well as derived chronology data is important to address the challenge of outdated chronologies becoming locked in to paleoclimate proxy records. Archives should be institutionally supported and may be best arranged along major themes such as the Paleocene-Eocene Thermal Maximum (PETM) or Last Glacial Maximum.

Reproducibility and data archives for global climate models (GCMs)

With the challenges of numerical reproducibility, and sources of random error such as cosmic ray strikes and other aberrations, bitwise reproducibility beyond what is necessary for debugging is difficult at best and of increasingly limited utility. Development of community standards and archives for GCMs are largely focused on issues of replicability of model results, building on the archives and best practices established by groups such as Coupled Model Intercomparison Project (CMIP). To 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. 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.

Because preserving the entire run time environment for GCMs is not cost-effective, there should exist standardized repositories of snapshots of GCM codes, configurations, and input files. This information should be linked to raw and derived datasets, and will preserve valuable information on ephemeral GCM computing environments. Code repositories currently exist in places like GitHub and Jupyter, but archiving of analysis code is haphazard and difficult to search, being associated variously with publication journals, institutions, or with individual researchers. Most GCMs have public releases of frozen codes or snapshots of code, e.g. from CMIP and NCAR, but no GCM is truly open-source and experimental and developmental versions of model code are not made available. This makes the benefit of publicly known tags for non-public model versions unclear. There is currently no standardized archive for specific model versions, although this is an option being explored by the NASA Goddard Institute for Space Sciences (GISS).

GCM archives will have to handle the tremendous amount of data generated, which makes re-analysis of archived raw data very challenging. This challenge might be addressed in part through continued development of data-handling methods such as lossy compression, However, currently it is unclear how this process can be documented and made reproducible, and this is an area of active investigation developing compression metrics to compare compressed and uncompressed data products to assess differences both visually and quantitatively.

Communication and incentives

Supporting organized, intelligent archives requires sufficient 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 reproducing existing work, but at the same time it carries a clear benefit to the broader community by increasing the robustness our understanding of climate systems, especially for applications of GCMs. Incentives are misaligned where data archiving remains a challenge and a burden to the producer of the data, while it is a benefit to others. Some of the challenges of aligning incentives would be alleviated with more organized and standardized centers and repositories for model and experiment replications, variations, code, and derived data, as well as easier routes to publication of comments and replications of existing work. Making data and tools citable and tagged with DOIs and referencing them in publications may help with accessibility.

Standardization and oversight of the application of 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.

1. Definitions of reproducibility and replicability

As has been noted in other fields, the terms reproducibility and replicability are defined and used differently across the climate science disciplines. Here, the terms are defined as follows: ***reproducibility*** refers to the 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.

2. Paleoclimate proxy data collection

2.1. Paleoclimate as the context for understanding climate today

Analyses of archives of geologic materials utilize paleoclimate proxy records to reconstruct aspects of the climate system such as temperature and sea level, and these paleoclimate archives 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 calibrate climate models, allowing researchers to test whether the models that are used to project future changes in climate conditions can reproduce conditions observed in the paleoclimate records.

Paleoclimate proxies can include any physical, chemical, or biological materials in the geologic record that can be correlated with some environmental parameter in the modern world, and much of the quantitative paleoclimate data comes from chemical or isotopic measurements of a geologic archive. Proxy records require first a calibration to identify and constrain the

relationship between climate parameter and proxy material, e.g. water temperature and Mg/Ca ratio in planktonic foraminifera (Anand, Elderfield, & Conte, 2003). This requires also constraining some amount of uncertainty inherent in the correlation, which in turn is propagated through all subsequent interpretations. Second, proxy reconstructions rely on measurements of geologic archive materials, which includes uncertainty and issues of reproducibility and replicability in the proxy measurements themselves, as well as—and just as importantly—uncertainty in the geochronology or age model used to date the proxy record. The age model applied to a geologic archive can provide not only absolute ages but also rates of change, which is especially critical to understanding climate dynamics through time. However, any particular age model is subject to its own separates issues with reproducibility and replicability and introduces an additional degree of uncertainty to the paleoclimate reconstruction; if the original source data are not published, or if only the derived ages are presented in a study, this age-based error can become locked into the interpretation once the study is published and persist even after geochronology is revised or altered. The NSF Paleo Perspectives on Climate Change (P2C2) Program funds much of the relevant, ongoing paleoclimate research, with the 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.

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. 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 no 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 archived at NOAA at the time of publications and serves as an example of the direction for best practices in reproducibility. There is extensive work done on synthesis of data and analyses, including work by the Intergovernmental Panel on Climate Changes (IPCC) and Past Global Changes (PAGES), which is an international working group that supports international and interdisciplinary research collaborations. Although important, much of this synthesis work is largely unfunded, in part due to its international nature. Online data archiving has become easier and more prevalent; in many cases, online archives are required with publication for several journals and granting agencies.

In studies of paleoclimate proxies from geologic archive materials, reproducibility and replicability can be tested both within and between individual studies. The majority of paleoclimate studies involve the destructive analysis of natural samples, meaning that collected source material can only be measured a finite number of times, which 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 or site. External variability is constrained through the comparison of community-accepted standards distributed widely among labs, e.g., stable isotope standards distributed by the International Atomic Energy Agency (IAEA). These standard measurements are reported with the sample data.

Because of the inherent chaos and 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 from a site. Paleoclimate proxy samples 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 reproducibility of each individual transect 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 multiple dimensions. 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.

Replication efforts in paleoclimate studies are typically folded into studies expanding the number of records, whether to fill in gaps in time or space with additional samples, or on comparing multiple proxies for the same climate parameter. Demonstrating inter-site replicability, Linsley et al. (2015) examined reconstructed sea surface temperature anomalies between three different sites in the Pacific. These proxy-based reconstructions qualitatively agree with historical records, but there is a high degree of random error in each individual proxy record, which is made transparent and accessible as the source data is archived with the journal.

The composite record not only demonstrates the replicability of the individual data but also smooths some of the error inherent in them.

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. 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 (Dutton et al., 2015; Medina-Elizalde, 2013; Potter & Lambeck, 2004) 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 (Lisiecki & Raymo, 2005). For true replication of a study, of all its components, including chronology and methods of proxy data conversion and interpretation, must be archived and made accessible. New platforms to support this include Linked Paleo Data (LiPD), which is a part of the EarthCube-supported Linked Earth project and provides the flexibility to deal with the peculiarities of paleoclimate proxies, divided into four broad categories: georeferencing and spatial context, publication record, proxy observations and systems models, and chronology and age models (Emile-Geay et al., 2017; McKay & Emile-Geay, 2016). Other examples of publicly archived digital data and metadata in use by the scientific community include databases managed by the Neotoma Paleoecology Database and the Linked Earth group, and code is available via

GitHub in open-source LiPD-linked tools and PRoxY Systems Modeling (PRYSM) tools (Dee et al., 2015).

However, there is frequently incomplete reporting or archiving of data or metadata, due in part to a lack of community-wide standards as well as a lack of enforcement by reporting journals or funding agencies. One measure to address this challenge is the publication and dissemination of discipline standards. A recent example of this 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 archives to increase the utility and longevity of study data. The NSF Data Infrastructure Building Blocks (DIBBs) program supports the development of robust and shared data-centric cyberinfrastructure, supporting 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. While 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, into relevant online archives.

In order 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. 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. Many laboratories have a standard that was originally distributed, but with increased instrument precision, it was

discovered that different dissolution techniques introduced differences in accuracies between the laboratory standards. Furthermore, there has been refinement in the U-Th decay constant values, meaning that ages published using different decay constants cannot be directly compared with one another. This is a challenge common to many paleoclimate proxies, which by their nature do not necessarily have stable relationships to time, as age models for proxy datasets are often preliminary and subject to change and update. Paleoclimate data archiving is often based on geochronology ages, which means that obsolete age data is locked in to the dataset, becomes propagated, and is difficult to remediate, hampering the utility of older studies. As a means of addressing this issue in U-Th dating, Andrea Dutton and colleagues are combining cyberinfrastructure supported by NSF DIBBs with modeling from EarthTime such that the open-source software to which measurements are uploaded will make calculations based on the new decay constants. 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, down-core data archives are easier to perform than age-based archives, and there remains the challenge that current archives 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 better metadata into paleoclimate archives so that they can be continuously reassessed and updated. Recent discussion from the Past Global Changes 2k (PAGES 2k) Network, which integrates paleoclimate datasets from the last 2000 years, highlighted the tension between the need to have studies be reproducible and replicable without

abusing early career scientists with investments in small parts of the larger collaborations (Kaufman, 2018). There are also a number of efforts to produce "intelligent" archives that could update age models, account for uncertainties in age and interpretation, and recalculate syntheses interactively. 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) can successfully synthesize multiple proxy records for a single time interval.

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., in press). 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. 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. Retaining bitwise reproducibility will become even more difficult in the near future as machine-learning algorithms and neural networks are introduced. 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, and by mapping errors in model output, researchers have been able to reconstruct the path of a particle as it passed through the memory of a supercomputer stack. 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 of data. Across the CMIP ensemble, it is easy to test the reproducibility 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 archived 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 no server-side analytics.

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 diags to reproduce 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.

CESM requires 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, CESM 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 principle component analyses, and then new runs are created with the changed results, and the comparison between the two datasets 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 model data, and issues of reproducibility similar to those described above are propagated. This is an area of high need for developing tools, procedures, and archives for handling reproducibility in order to better connect a published artifact to its original model input.

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 increasing, future models will only generate more data more quickly. Thus, lossy data compression can potentially serve as a means of reducing the cost of preservation and making data more easily accessible, but it must be determined which information can be safely lost and

which must be preserved so that results are not negatively impacted. NCAR has experimented with a tool, known as fpzip, that gives an average 5x compression factor across climate variables and favorable statistical scores. 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, and 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 to make the maintenance of these archives more affordable via the application of statistics, overriding the mathematically more feasible but 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 archives 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 may also help with accessibility. There are many archives and repositories for code as well as standard toolkits, including netCDF Operators (NCO), Community Data Analysis Tools (CDAT), and libraries for code in Python, R, Matlab, and IDL languages. Standard code repositories exist in places like GitHub and Jupyter, but archiving of analysis code is haphazard and difficult to search, being associated variously with publication journals, institutions, or with individual researchers. 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 made available, which makes the benefit of publicly known tags for non-public model versions unclear. There is also no standardized archive for specific model versions, although this is an option currently be 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 will be recapitulated 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, 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.

4. Interfacing paleoclimate proxies and modern climate science

4.1. Climate science data archives, 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 archives

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.

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. GISTEMP recalculates the homogenization every month based on new data; thus as the input data 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, but our understanding of the progression of global mean temperature over the history of instrumental records is quite robust. That robustness is confirmed via comparisons with independently calculated datasets, and our understanding of global mean temperature is both robustly replicable and reproducible.

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 (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 reproduction, 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. 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 replicability in sample measurement and calibration against accepted standard materials are generally well-established. Practices of error propagation and constraining uncertainty impact study reproducibility and require international assessment and oversight. Observational replication within and across paleoclimate sites improves our estimation of amplitude and chronological uncertainties. 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.

Paleoclimate data and metadata archiving has improved greatly over time. However, issues of data archiving, which is especially challenging for legacy data, and large-scale synthesis are being addressed by ongoing database development projects and international working groups, but are beset with challenges regarding funding sources and misaligned incentives regarding who does the work and who benefits from it. The reliance of paleoclimate data archives and publications on geochronology means that obsolete dates and age models can become locked into paleoclimate datasets and propagated through the community, sometimes for decades after the original age model was revised.

In global climate models, bitwise or computational replication may be relatively easy in short time scales and within computing centers, but is difficult to expand, especially as methods such as machine-learning are introduced. Global climate computer models have moved away from bitwise reproducibility as computing speeds and dataset sizes increase, and as methods of data storage and access incorporate lossy data compression, which must be tested for its own reproducibility challenges. Reproducibility is feasible within multi-model ensembles such as CMIP; while reproductions test the robustness of model results, interpretation is a challenge where models diverge. Distributed computing presents its own reproducibility challenges due to the difficulty of tracking the impact of threading and other decision processes inherent in parallelization, and these same reproducibility challenges will necessarily apply to subsequent parallel analytics of model results. Similarly to paleoclimate proxy research, the state of GCM science would benefit greatly from the development of active archives 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 archives.

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.

Anand, P., Elderfield, H., & Conte, M. H. (2003). Calibration of Mg/Ca thermometry in planktonic foraminifera from a sediment trap time series. *Paleoceanography, 18*(2).

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.

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.

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.

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), 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.

Emile-Geay, J., McKay, N. P., Kaufman, D. S., Von Gunten, L., Wang, J., Anchukaitis, K. J., . . . Evans, M. N. (2017). A global multiproxy database for temperature reconstructions of the Common Era. *Scientific data, 4*, 170088.

Hall, A., & Stouffer, R. J. (2001). An abrupt climate event in a coupled ocean–atmosphere simulation without external forcing. *Nature, 409*(6817), 171. doi: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*, 130-163.

Kaufman, D. S. (2018). Open-paleo-data implementation pilot–the PAGES 2k special issue. *Climate of the Past, 14*(5), 593.

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.

Lisiecki, L. E., & Raymo, M. E. (2005). A Pliocene□Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography, 20*(1).

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

Medina-Elizalde, M. (2013). A global compilation of coral sea-level benchmarks: implications and new challenges. *Earth and Planetary Science Letters, 362*, 310-318.

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.

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.

Appendix A: From the agenda for the Paleoclimate Panel:

Perspectives on Reproducibility and Replication of Results in Climate Science

Each panelist will have 15 minutes to address the set of questions below followed by 5 minutes of questions form the committee

• How has the awareness and understanding about reproducibility and replication in climate science evolved over recent years?
• Are there specific challenges regarding reproducibility that you have encountered or are aware of? Identify specific steps that are being taken, either by you or by others, to ameliorate these issues.
• Highlight historical and potential new approaches to reproducing and replicating climate science research using examples such as paleoclimate data to test models and estimate uncertainties.

Michael Evans, Associate Professor, University of Maryland
Gavin Schmidt, Director of the Goddard Institute for Space Studies, National Aeronautics and Space Administration
Rich Loft, Director of the Technology Development Division, National Center for Atmospheric Research (NCAR)
Andrea Dutton, Assistant Professor, University of Florida