Trustworthy computational evidence through transparency and reproducibility

Lorena A. Barba[1]

[1]Affiliation not available

January 12, 2021

Abstract

Abstract content goes here

Many high-performance computing applications are of high consequence to society. Global climate modeling is a historic example of this. In 2020, the societal issue of greatest concern, the still-raging COVID-19 pandemic, saw a legion of computational scientists turning their endeavors to new research projects in this direction. Applications of such high consequence highlight the need for building trustworthy computational models.

Emphasizing transparency and reproducibility has helped us build more trust in computational findings. In the context of supercomputing, however, we may ask: how do we trust results from computations that cannot be repeated? Access to supercomputers is limited, computing allocations are finite (and competitive), and machines are decommissioned after a few years. In this context, we might ask how reproducibility can be ensured, certified even, without exercising the original digital artifacts used to obtain new scientific results. This is often the situation in HPC. It is compounded now with greater adoption of machine learning techniques, which can be opaque. The ACM in 2017 issued a Statement on Algorithmic Transparency and Accountability, targeting algorithmic decision-making using data models (ACM U.S. Public Policy Council, 2017). Among its seven principles, it calls for data provenance, auditability, validation and testing. These principles can be applied not only to data models, but to HPC in general. I want to discuss the next steps for reproducibility: how we may adapt our practice to achieve what I call *unimpeachable provenance*, and full auditability and accountability of scientific evidence produced via computation.

An invited talk at SC20

I was invited to speak at SC20 about my work and insights on transparency and reproducibility in the context of HPC. The session's theme was *Responsible Application of HPC*, and the title of my talk was "Trustworthy computational evidence through transparency and reproducibility." At the previous SC, I had the distinction to serve as Reproducibility Chair, leading an expansion of the initiative, which was placed under the Technical Program that year. We moved to make Artifact Description appendices required for all SC papers, created a template and an author kit for the preparation of the appendices, and introduced three new Technical Program tracks in support of the initiative. These are: the Artifact Description & Evaluation Appendices track—with an innovative double-open constructive review process—, the Reproducibility Challenge track, and the Journal Special Issue track, for managing the publication of select papers on the reproducibility benchmarks of the Student Cluster Competition. This year, the initiative was augmented to address issues of transparency, in addition to reproducibility, and a community sentiment study was launched to assess the impact of the effort, six-years in, and canvas the community's outlook on various aspects of it.

Allow me to thank here Mike Heroux, Reproducibility Chair for SC in 2017 and 2018, Michela Taufer, SC19 General Chair—who put her trust in me to inherit the role from Mike—, and Beth Plale, the SC20 Transparency and Reproducibility Chair. I had countless inspiring and supportive conversations with Mike and Michela about the topic during the many months of planning for SC19, and more productive conversations with Beth during the transition to her leadership. Mike, Michela and I have served on other committees and working groups together, in particular, the group that met in July 2017 at the National Science Foundation (convened by Almadena Chtchelkanova) for the Workshop on Reproducibility Taxonomies for Computing and Computational Science. My presentation at that event condensed an inventory of uses of various terms like reproducibility and replication, across many fields of science (Barba, 2017). I then wrote the review article "Terminologies for Reproducible Research," and posted it on arXiv (Barba, 2018). It informed our workshop's report, which came out a few months later as a Sandia technical report (*Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences. (Technical Report) — OSTI.GOV*, 2018). In it, we highlighted that the fields of computational and computing sciences provided two opposing definitions of the terms reproducible and replicable, representing an obstacle to progress in this sphere.

The Association of Computing Machinery (ACM), representing computer science and industry professionals, had recently established a reproducibility initiative, and adopted diametrically opposite definitions to those used in computational sciences for more than two decades. In addition to raising awareness about the contradiction, we proposed a path to a compatible taxonomy. Compatibility is needed here because the computational sciences—astronomy, physics, epidemiology, biochemistry and others that use computing as a tool for discovery—and computing sciences (where algorithms, systems, software, and computers are the focus of study) have community overlap and often intersect in the venues of publication. The SC conference series is one example. Given the historical precedence and wider adoption of the definitions of reproducibility and replicability used in computational sciences, our Sandia report recommended that the ACM definitions be reversed. Several ACM-affiliated conferences were already using the artifact review and badging system (approved in 2016), so this was no modest suggestion. The report, however, was successful in raising awareness of the incompatible definitions, and the desirability of addressing it.

A direct outcome of the Sandia report was a proposal to the National Information Standards Organization (NISO) for a *Recommended Practice Toward a Compatible Taxonomy, Definitions, and Recognition Badging Scheme for Reproducibility in the Computational and Computing Sciences*. NISO is accredited by the American National Standards Institute (ANSI) to develop, maintain, and publish consensus-based standards for information management. The organization has more than 70 members; publishers, information aggregators, libraries and other content providers use its standards. I co-chaired this particular working group, with Gerry Grenier from IEEE and Wayne Graves from ACM; Mike Heroux was also a member. The goal of the NISO Reproducibility Badging and Definitions Working group was to develop a Recommended Practice document—a step before development of a *standard*. As part of our joint work, we prepared a letter addressed to the ACM Publications Board, delivered in July 2019. It described the context and need for compatible reproducibility definitions and made the concrete request that ACM consider a change. By that time, not only did we have the Sandia report as justification, but the National Academies of Sciences, Engineering and Medicine (NASEM) had just released the report *Reproducibility and Replicability in Science* (NASEM, 2019). It was the product of a long consensus study conducted by 15 experts, including myself, and sponsored by the National Science Foundation responding to Congressional decree. The NASEM report put forth its definitions as:

> Reproducibility is obtaining consistent results using the same input data, computational steps, methods and code, and conditions of analysis.
>
> Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

The key contradiction with the ACM badging system resides on which term comprises using the author-created digital artifacts (e.g., data and code). We stated in the NISO working-group letter that if the ACM definitions of reproducible and replicable could be interchanged, the working group could move forward

towards its goal of drafting recommended practices for badging that would lead to wider adoption in other technical societies and publishers. The ACM Publications Board responded positively, and began working through the details on how to make changes to items already published in the Digital Library with the *"Results Replicated"* badge—about 188 items existed at that time that were affected. Over the Summer of 2020, the ACM applied changes to the published Artifact Review and Badging web pages, and added a version number. From version 1.0, we see a note added that, as a result of discussions with NISO, the ACM was harmonizing its terminologies with those used in the broader scientific research community.

All this background serves to draw our attention to the prolonged, thoughtful, and sometimes arduous efforts that have been directed at charting paths for adoption and giving structure to reproducibility and replicability in our research communities. Let us move now to why and how might the HPC community move forward.

Insights on transparent, reproducible HPC research

Deployed barely over a year ago, the NSF-funded Frontera system at the Texas Advanced Computing Center (TACC) came in as the 8th most powerful supercomputer in the world, and the fastest on a university campus. Up to 80% of the available time on the system is allocated through the NSF Petascale Computing Resource Allocation program. The latest round of Frontera allocations (as of this writing) was just announced on October 25, 2020. I read through the fact sheet on the 15 newly announced allocations, to get a sense for the types of projects in this portfolio. Four projects are machine-learning or AI-focused, the same number as those in astronomy and astrophysics, and one more than those in weather or climate modeling. Other projects are single instances spanning volcanology/mantle mechanics, molecular dynamics simulations of ion channels, quantum physics in materials science, and one engineering project in fluid-structure interactions. One could gather these HPC projects in four groups:

1. Astronomy and astrophysics are mature fields that in general have high community expectations of openness and reproducibility. As I'll highlight below, however, even these communities with mature practices benefit from checks of reproducibility that uncover areas of improvement.
2. The projects tackling weather and climate modeling are candidates for being considered of *high consequence to society*. One example from the Frontera allocations concerns the interaction of aerosols caused by industrial activity with clouds, which can end up composed of smaller droplets, and become more reflective, resulting in a cooling effect on climate. Global climate models tend to overestimate the radiative forcing, potentially underestimating global warming: why? This is a question of great consequence for science-informed policy, in a subject that is already under elevated scrutiny from the public. Another project in this cluster deals with real-time high-resolution ensemble forecasts of high-impact winter weather events. I submit that high standards of transparency, meticulous provenance capture, and investments of time and effort in reproducibility and quality assurance are justified in these projects.
3. Four of the winning projects are applying techniques from machine learning to various areas of science. In one case, the researchers seek to bridge the gap in the trade-off between accuracy of prediction and model interpretability, to make ML more applicable in clinical and public health settings. This is clearly also an application of high consequence, but in addition all the projects in this subset face the particular transparency challenges of ML techniques, requiring new approaches to provenance capture and transparent reporting.
4. The rest of the projects are classic high-performance computational science applications, such as materials science, geophysics, and fluid mechanics. Reproducible-research practices vary broadly in these settings, but I feel confident saying that all or nearly all those efforts would benefit from prospective data management, better software engineering, and more automated workflows. And their communities would grow stronger with more open sharing.

The question I have is: how could the merit review of these projects nudge researchers towards greater transparency and reproducibility? Maybe that is a question for later, and a question to start with is how

could support teams at cyberinfrastructure facilities work with researchers to facilitate their adoption of better practices in this vein? I'll revisit these questions later.

I also looked at the 2019 Blue Waters Annual Report, released on September 15, 2020, with highlights from a multitude of research projects that benefitted from computing allocations on the system. Blue Waters went into full service in 2013 and has provided over 35 billion core-hour equivalents to researchers across the nation. The highlighted research projects fall into seven disciplinary categories, and include 32 projects in space science, 20 in geoscience, 45 in physics and engineering, and many more. I want to highlight *just one* out of the many dozens of projects featured in the Blue Waters Annual Report, for the following reason. I did a word search on the PDF with Zenodo, and that project was the *only* one listing Zenodo entries in the "Publications & Data Sets" section that ends each project feature. One other project (in the domain of astrophysics) mentions that data is available through the project website and in Zenodo, but doesn't list any data sets in the report. Zenodo is an open-access repository funded by the European Union's Framework Programs for Research, and operated by CERN. Some of the world's top experts in running large-scale research data infrastructure are at CERN, and Zenodo is hosted on top of infrastructure built in service of what is the largest high-energy physics laboratory of the world. Zenodo hosts any kind of data, under any license type (including closed-access). It has become one of the most used archives for open sharing of research objects, including software.

The project I want to highlight is "Molten-salt reactors and their fuel cycles," led by Prof. Kathryn Huff at UIUC. I've known Katy since 2014, and she and I share many perspectives on computational science, including a strong commitment to open-source software. This project deals with modeling and simulation of nuclear reactors and fuel cycles, combining multiple physics and multiple scales, with the goal of improving design of nuclear reactors in terms of performance and safety. As part of the research enabled by Blue Waters, the team developed two software packages: Moltres, described as a first-of-its-kind finite-element code for simulating the transient neutronics and thermal hydraulics in a liquid-fueled molten-salt reactor design; and SaltProc: a Python tool for fuel salt reprocessing simulation. The references listed in the project highlight include research articles in the Annals of Nuclear Energy, as well as the Zenodo deposits for both codes, and a publication about Moltres in the Journal of Open Source Software, JOSS. (As one of the founding editors of JOSS, I'm very pleased.) It is possible, of course, that other projects of the Blue Waters portfolio have also made software archives in Zenodo or published their software in JOSS, but they did not mention it in this report and did not cite the artifacts. Clearly, the research context of the project I highlighted is of *high consequence*: nuclear reactor design. The practices of this research group show a high standard of transparency that should be the norm in such fields. Beyond transparency, the publication of the software in JOSS ensures that it was subject to peer review and that it satisfies standards of quality. JOSS reviewers install the software, run tests, and comment on usability and documentation, leading to quality improvements.

Next, I want to highlight the work of a group that includes CiSE editors Michela Taufer and Ewa Deelman, posted last month on arXiv (Brown et al., 2020)[6]. The work sought to directly reproduce the analysis that led to the 2016 discovery of gravitational waves, using the data and codes that the LIGO collaboration had made available to the scientific community. The data had previously been re-analyzed by independent teams using different codes, leading to *replication* of the findings, but no attempt had yet been made at reproducing the original results. In this paper, the authors report on challenges they faced during the reproduction effort, *even* with availability of data and code supplementing the original publication. A first challenge was the lack of a single public repository with all the information needed to reproduce the result. The team had the cooperation of one of the original LIGO team members, who had access to unpublished notes that ended up being necessary in the process of iteratively filling in the gaps of missing public information. Other highlights of the reproduction exercise include: the original publication did not document the precise version of the code used in the analysis; the script used to make the final figure was not released publicly (but one co-author gave access to it privately); the original documented workflow queried proprietary servers to access data, which needed to be modified to run with the public data instead. In the end, the result—the statistical significance of the gravitational-wave detection from a black-hole merger—*was reproduced*, but not

independently of the original team, as one researcher is co-author in both publications. The message here is that even a field that is mature in its standards of transparency and reproducibility needs checks to ensure that these practices are sufficient or can be improved.

Science policy trends

The National Academies study on Reproducibility and Replicability in Science was commissioned by the National Science Foundation under Congressional mandate, with the charge coming from the Chair of the Science, Space, and Technology Committee. NASEM reports and convening activities have a range of impacts on policy and practice, and often guide the direction of federal programs. NSF is in the process of developing its agency response to the report, and we can certainly expect to hear more in the future about requirements and guidance for researchers seeking funding.

The recommendations in the NASEM report are directed at all the various stakeholders: researchers, journals and conferences, professional societies, academic institutions and national laboratories, and funding agencies. Recommendation 6-9, in particular, prompts funders to ask that grant applications discuss how they will assess and report uncertainties, and how the proposed work will address reproducibility and/or replicability issues. It also recommends that funders incorporate reproducibility and replicability in the *merit-review criteria* of grant proposals. Combined with related trends urging for more transparency and public access to the fruits of government-funded research, we need to be aware of the shifting science-policy environment.

One more time, I have a reason to thank Mike Heroux, who took time for a video call with me as I prepared my SC20 invited talk. In his position as Senior Scientist at Sandia, 1/5 of his time is spent in service to the lab's activities, and this includes serving in the review committee of the internal Laboratory Directed Research & Development (LDRD) grants. As it is an internal program, the Calls for Proposals are not available publicly, but Mike told me that they now contain specific language asking proposers to include statements on how the project will address transparency and reproducibility. These aspects *are discussed* in the proposal review and are a factor in the decision-making. As community expectations grow, it *could happen* that between two proposals equally ranked in the science portion the tie-break comes from one of them better addressing reproducibility. Already some teams at Sandia are performing at a high level, e.g., they produce an Artifact Description appendix for every publication they submit, regardless of the conference or journal requirements.

We don't know if or when NSF might add similar stipulations to general grant proposal guidelines, asking researchers to describe transparency and reproducibility in the project narrative. One place where we see the agency start responding to shifting expectations about open sharing of research objects is the section on results from prior funding. NSF currently requires here a listing of publications from prior awards, and *"evidence of research products and their availability, including ... data [and] software."*

I want to again thank Beth Plale, who took time to meet with me over video and sent me follow-up materials to use in preparing my SC20 talk. In March 2020, NSF issued a "Dear Colleague Letter" on Open Science for Research Data, with Beth then acting as the public access program director. The DCL says that NSF is expanding its Public Access Repository (NSF PAR) to accept metadata records, leading to data discovery and access. It requires research data to be deposited in an archival service and assigned a Digital Object Identifier (DOI), a global and persistent link to the object on the web. A grant proposal's Data Management Plan should state the anticipated archive to be used, and include any associated cost in the budget. Notice this line: *"Data reporting will initially be voluntary."* This implies that it will later be mandatory! The DCL invited proposals aimed at growing community readiness to advance open science. At the same time, the Office of Science and Technology Policy (OSTP) issued a Request for Information early this year asking what could Federal agencies do to make the results from research they fund publicly accessible. The OSTP sub-committee on open science is very active. An interesting and comprehensive response to the OSTP RFI comes from the MIT Libraries. It recommends (among other things):

- Policies that default to open sharing for data and code, with opt-out exceptions available [for special cases]...
- Providing incentives for sharing of data and code, including supporting credentialing and peer-review; and encouraging open licensing.
- Recognizing data and code as "legitimate, citable products of research" and providing incentives and support for systems of data sharing and citation...

The MIT Libraries response addresses various other themes like responsible business models for open access journals, and federal support for vital infrastructure needed to make open access to research results more efficient and widespread. It also recommends that Federal agencies provide incentives for documenting and raising quality of data and code, and also "promote, support, and require effective data practices, such as persistent identifiers for data, and efficient means for creating auditable and machine readable data management plans."

To boot, the National Institutes of Health (NIH) just announced on October 29 a new policy on data management and sharing. It requires researchers to plan prospectively for managing and sharing scientific data openly, saying: *"we aim to shift the culture of research to make data sharing commonplace and unexceptional."*

Another setting where we could imagine expectations to discuss reproducibility and open research objects is proposals for allocation of computing time. For this section, I need to thank John West, Director Of Strategic Initiatives at the Texas Advanced Computing Center (and CiSE Associate EiC), who took time for a video call with me on this topic. We bounced ideas about how cyber-infrastructure providers might play a role in growing adoption of reproducibility practices. Currently, the NSF science proposal and the computing allocation proposal are awarded separately. The Allocation Submission Guidelines discuss review criteria, which include: intellectual merit (demonstrated by the NSF science award), methodology (models, software, analysis methods), research plan and resource request, and efficient use of the computational resources. For the most part, researchers have to show that their application scales to the size of the system they are requesting time on. Interestingly, the allocation award is not tied to *performance*, and researchers are not asked to show that their codes are optimized, only that they scale and that the research question is feasible to be answered in the allocated time. The responsible stewardship of the supercomputing system is provided for via a close collaboration between the researchers and the members of the supercomputing facility. Codes are instrumented under the hood with low-overhead collection of system-wide performance data (in the UT facility, with TACC-Stats) and a web interface for reports.

I see three opportunities here: 1) workflow-management and/or system monitoring could be extended to also supply automated provenance capture; 2) the expert staff at the facility could broaden their support to researchers to include advice and training in transparency and reproducibility matters; and 3) cyber-infrastructure facilities could expand their training initiatives to include essential skills for reproducible research. John floated other ideas, like the possibility that some projects be offered a bump on their allocations (say, 5% or 10%) to engage in R&R activities; or, more drastic perhaps, that projects may not be awarded allocations over a certain threshold *unless* they show commitment and a level of maturity in reproducibility.

Next steps for HPC

The SC Transparency and Reproducibility Initiative is one of the innovative, early efforts to gradually raise the expectations and educate a large community about how to address it and why it matters. Over six years, we have built community awareness, and buy-in. This year's community sentiment study shows frank progress: 90% of the respondents are aware of the issues around reproducibility, and only 15% thought the concerns are exaggerated. Importantly, researchers report that they *are* consulting the artifact appendices of technical papers, signaling impact. As a community, we are better prepared to adapt to raising expectations from funders, publishers, and readers.

The pandemic crisis has unleashed a tide of actions to increase access and share results: the Covid-19 Open Research Dataset (CORD-19) is an example (Wang et al., 2020); the COVID-19 Molecular Structure and Therapeutics Hub at MolSSI is another. Facing a global challenge, we as a society are strengthened by facilitating immediate public access to data, code, and published results. This point has been made by many in recent months, but perhaps most eloquently by Rommie Amaro and Adrian Mulholland in their Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19—signed by more than a hundred researchers from around the world (Amaro & Mulholland, 2020). It says: *"There is an urgent need to share our methods, models, and results openly and quickly to test findings, ensure reproducibility, test significance, eliminate dead-ends, and accelerate discovery."* Then it follows with several commitments:

1. to making results available quickly via pre-prints;
2. to make available input files, model-building and analysis scripts (e.g., Jupyter notebooks), and data necessary to reproduce the results;
3. to use open data-sharing platforms to make available results as quickly as possible;
4. to share algorithms and methods in order to accelerate reuse and innovation; and
5. to apply permissive open-source licensing strategies.

Interestingly, these commitments are reminiscent of the pledges I made in my Reproducibility PI Manifesto (Barba, 2012) eight years ago!

One thing the pandemic instantly provided is a *strong incentive* to participate in open science and attend to reproducibility. The question is how much will newly adopted practices persist once the incentive of a world crisis is removed.

I've examined here several issues of *incentives* for transparent and reproducible research. But social epistemologists of science know that so-called Mertonian norms (for sharing widely the results of research) are supported by both economic and ethical factors—incentives and norms—in close interrelation. Social norms require a predominant normative expectation (for example, sharing of food in a given situation and culture). In the case of open sharing of research results, those expectations are not prime, due to researchers' sensitivity to credit incentives. Heesen (Heesen, 2017) concludes: "Give sufficient credit for whatever one would like to see shared ... and scientists will indeed start sharing it."

In HPC settings, where we can hardly ever reproduce results (due to machine access, cost, and effort), a vigorous alignment with the goals of transparency and reproducibility will develop a blend of incentives and norms, will consider especially the applications of high consequence to society, and will support researchers with infrastructure (human and cyber). Over time, we will arrive at a level of maturity to achieve the goal of trustworthy computational evidence, not by actually exercising the open research objects (artifacts) shared by authors (data and code), but by a research process that ensures unimpeachable provenance.

References

Statement on algorithmic transparency and accountability. (2017). Online http://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf. valuehttp://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

Science Reproducibility Taxonomy. (2017). https://doi.org/10.6084/m9.figshare.5248273./articles/presentation/Science$_R$*eproducibility*$_T$*axonomy*/5248273/1

Terminologies for Reproducible Research. (2018). https://arxiv.org/abs/1802.03311. https://arxiv.org/abs/1802.03311v1

(2018). https://doi.org/10.2172/1481626. https://doi.org/10.2172/1481626

Reproducibility and Replicability in Science. (2019). National Academies Press. https://doi.org/10.17226/25303

Reproducing GW150914: the first observation of gravitational waves from a binary black hole merger. (2020). https://arxiv.org/abs/2010.07244. https://arxiv.org/abs/2010.07244v1

CORD-19: The COVID-19 Open Research Dataset. (2020). https://arxiv.org/abs/2004.10706. https://arxiv.org/abs/2004.10706v4

A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. (2020). https://doi.org/10.1021/acs.jcim.0c00319. https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c00319

Reproducibility PI Manifesto. (2012). http://doi.org/10.6084/m9.figshare.104539./articles/presentation/Reproducibility$_P I_M anifesto$/104539/1

Communism and the Incentive to Share in Science. (2017). *Philosophy of Science*, *84*(4), 698–716. https://www.journals.uchicago.edu/doi/abs/10.1086/693875