Reproducibility Practice in High Performance Computing: Community Survey Results

Beth Plale
Indiana University Bloomington

Tanu Malik
School of Computing, DePaul University

Line Pouchard
Brookhaven National Laboratory

***Abstract*—The integrity of science and engineering research is grounded in assumptions of rigor and transparency on the part of those engaging in such research. HPC community effort to strengthen rigor and transparency take the form of reproducibility efforts. In a recent survey of the SC conference community, we collect information about the SC Reproducibility Initiative practices. We present the survey results in this paper. Results show that the reproducibility initiative practices have contributed to higher levels of awareness on the part of SC conference technical program participants, and hint at contributing to greater scientific impact for the published papers of the SC conference series. Stringent point-of-manuscript-submission verification is problematic for reasons we point out, as are inherent difficulties of computational reproducibility in HPC. Future effort should better decouple the community educational goals from goals that specifically strengthen a research work's potential for long-term impact through reuse 5-10 years down the road.**

Keywords: Transparency, rigor, reproducibility, replicability, open science

■ MOTIVATION

The integrity of science and engineering research is grounded in the practice of rigor and transparency on the part of those participating. Rigor is defined as "the strict application of the scientific method to ensure robust and unbiased experimental design'". Transparency, on the other hand, is the sharing of details about one's research, including study design, operationalization of variables, measurement techniques, and uncertainties [5]. A number of studies have been carried over the last 6 years to measure how well research studies can be reproduced. The studies take the form of selecting a set of published research results, often within a discipline, and attempting to reproduce the primary results using material and products shared publicly (either outright or upon request). These studies by and large illuminate pockets of weakness in the practice of science and engineering (S&E) research.

Concerns brought about by these reproducibility studies raised such broad attention that the US Congress directed the National Science Foundation to fund a National Academies of Science, Engineering, and Medicine (NASEM) study to assess reproducibility and replicability in scientific and engineering research in order to improve the rigor and transparency of S&E research. We use the NASEM study definitions for reproducibility, replicability, as well as for transparency and rigor [5].

Individual research communities are the ones left tasked with translating needs for enhanced rigor and transparency into effective practice for their community. For the high performance computing (HPC) community, the task is fraught with very real technical and policy challenges. There are numerous reasons why results obtained in a complex computational environment cannot attain full *computational reproducibility* (that is, obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis [5]). The scale at which some computational experimentation is carried out makes computational reproducibility at scale prohibitively expensive; results on smaller machines can yield different results. Additionally, software stacks evolve quickly, getting allocations on HPC systems is a competitive process, proprietary software can be involved, and systems get decommissioned every few years, for instance with the DOE Leadership Computing Facilities, among other difficulties [6].

Sustaining a practice of strong rigor and transparency in research is an obligation on a community that does not evaporate in the face of the abundant challenges of high performance computing. Instead *the HPC community must together innovate and educate to ensure the practice of strong rigor and transparency in research by current and future generations of researchers.* This takes the form of innovation in new forms of formal practice to assert rigor and transparency, and in education through training down to each of our laboratories and centers.

We believe that the best course of action within the HPC community is the practice of reproducible science, where authors are rigorous in their research, rigorous in mentoring new researchers, and transparent in their results (manuscript and research artifacts). The current practice is to engage an external party at manuscript submission time to assess artifact functionality from a reuse perspective. Such a form of community-based reproducibility at manuscript submission time must adhere to a formal and itself transparent practice in assessing results. The practice needs the support of the community who are aware of, and can weigh in on, the burden-benefit tradeoffs.

The Supercomputing (SC) series of conferences has taken the lead in community efforts in reproducibility through its SC Reproducibility Initiative of which each of us authors has had a leadership role at one time or another. The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) attracts over 10,000 participants annually to an event that features breaking news, a expansive exhibit floor, and a technical program for high-quality original research, groundbreaking ideas, and compelling insights on future trends in high performance computing, networking, storage, and analysis. The technical program receives over 300 submissions annually, and after an extensive peer review process, selects about 20% for presentation and publication in the SC proceedings which are archived in the ACM Digital Library and IEEE Xplore.

The SC conference began its Reproducibility Initiative in 2015 primarily as an optional practice for authors of accepted papers to describe their experimental framework and results in more detail. The form it took, still practiced today, is for authors to include an Artifact Description (AD) appendix, and more extensive Artifact Evaluation (AE) appendix. The AD appendix allows an external party to determine whether artifacts are available, and the AE appendix provides sufficient detail to support an independent audit. In 2015 authors of only one paper responded to the initiative, and that paper became the source for the SC16 Student Cluster Competition Reproducibility Challenge; it is also the first SC paper to display an ACM badge. By 2017 39 papers had an AD. In 2019 the AD became mandatory [1]. The AD/AE evaluation process is peer-reviewed and provides guidance to the technical program committee, especially if reproducibility of results becomes a critical factor in experimental results.

SURVEY OF THE HPC COMMUNITY

In August 2020, two of us authors surveyed the HPC community on their perceptions of reproducibility and transparency, and on the SC Reproducibility Initiative itself. We sought to understand the impact, perceived burdens, potential benefits of the initiative, and inform its direction.

We present detailed analysis of the survey and follow that in the next section with implications for the HPC community.

Methodology

We used the population of the SC community to generate a purposeful sample. Purposeful sampling is a widely-used technique in qualitative research for the identification and selection of people for their ability to provide information. In August 2020, the survey invitation was sent to 9,949 unique individuals drawn from those who had participated in SC17, SC18, or SC19 technical programs. Registrants for the SC20 technical program were not included as registration was ongoing during the survey response period. The invitation was re-sent once and the survey closed August 31, 2020. Of these recipients 204 self-selected participants responded to at least one question (outside the consent question) and 149 completed the survey.

The survey was conducted under Indiana University protocol #2005780098, “Assessing Reproducibility Initiative of IEEE/ACM Supercomputing Conference,” as an online survey protecting the anonymity of respondents. The questionnaire asked no identifying questions and this paper reports the findings in aggregate.

The survey was organized into 3 sets of questions. The first set addressed the respondent’s role at SC - whether or not they submitted a paper, an appendix; whether or not they were a student.

The second set of questions assessed respondent perceptions and awareness of issues in reproducibility in computer and computational science. In the third set, respondents were asked to evaluate their experience specifically with the SC Artifact Description/Artifact Evaluation (AD/AE) process.

For all questions, a single choice of answer was permitted. For non-demographic questions, respondents were given a Lickert scale (strongly disagree, somewhat disagree, somewhat agree, strongly agree) avoiding ambiguities related to a neutral response.

In the analysis, we use the demographic responses (pertaining to the engagement with the SC as an author, and whether the participant was a student) as explanatory variables and performed an F test and its associated probabilities. We also performed ANOVA (Analysis of Variance) where appropriate to understand if the selected categories had an effect on the relationship between a variable in the demographic response (independent) and the response to a question (dependent variable).

Since responses to the questionnaire was partial, percentages reflect the number of people who responded to that particular question. We aggregated responses for percentages as follows: ”somewhat or strongly agree” are reported here as agreement, and the same for disagreement. For all survey questions, results obtained from those who categorized themselves as students did not differ from the non-students in a statistically significant way so we do not report the test values here. The same is true for the other segmentation groups.

Results

Results break down into four topics: general results, impact on science and engineering, transparency, and technology. General results suggest that after 6 years of the AD/AE initiative at SC20, awareness is high:

- a full 90% of respondents are aware of issues related to reproducibility in computational and computer sciences,
- only 15% think that the concerns about reproducibility in science are exaggerated, and
- only 7% think that the concerns about reproducibility in science do not apply to computer and computational science.
- A full 90% of respondents were satisfied with the SC approach of double blind review for the technical paper coupled with an open-open process for the AD/AE review. The points of interaction appear to have been clearly established and published for authors as respondents agreed with the privacy preserving between the two sides.
- Finally, 76% found the guidelines helpful.

Scientific Impact. The SC Reproducibility Initiative serves both to instill rigorous and transparent community research practice and to enhance the scientific impact of SC conference and workshop technical publications. Our study strove to measure the effect of the latter goal by asking the community to assert a level of agreement with the following statement: *I have*

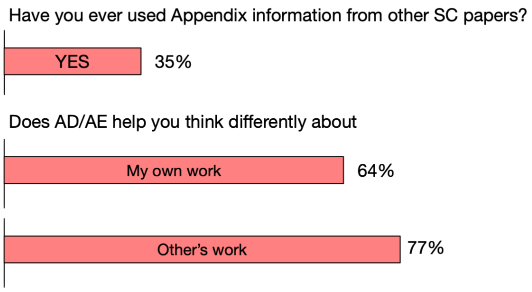

Figure 1. Impact of SC reproducibility effort

used the appendices previously published at SC in the development of my research results. The assumption behind the question is if a researcher has used the appendix information of a published SC paper, even if just to consult it, then the appendices information has contributed to greater overall impact of the published research result. As shown in Figure 1, a full thirty five percent (35%) of the respondents were affirmative in their use of appendix information.

An even greater number of respondents expressed that the SC Reproducibility Initiative appendices requirement has the effect of encouraging researchers to think about reproducibility, as 64% stated that *the content of the appendices makes me think in different ways about publishing my results*, and 77% stated they now think in different ways about results published in other scientific papers.

Together the responses shown in Figure 1 suggest that the SC Reproducibility Initiative appears to be having a positive impact on the quality of the science published by SC.

Transparency. Where authors do not practice transparency in the sharing of their research results, the practice of reproducing those results is largely impossible. A large majority of respondents felt that transparency is the goal of the SC reproducibility activity. That is, a full 75% agreed that it is *”transparency (and not reproducibility) is the goal of the AD/AE forms at SC”*, with the caveat that 20% of respondents separately answered that they do not understand the difference between transparency and reproducibility.

It should be noted that SC20 introduced the term ”transparency” into the vernacular of the SC Reproducibilty Initiative, in part because of the growing number of manuscripts submitted to SC that use artificial intelligence (AI) techniques; AI techniques place additional demands on achieving transparent results. Ad hoc methods (e.g., to tune an AIs' “learning rates”—how much an algorithm corrects itself after each mistake) are sometimes used without justification for why one ad hoc method is better than others.

In addition, authors often report the best accuracy results, results that often cannot be reproduced. Making code available in these cases is insufficient; sharing training models, hyperparameters, data training/test splits is encouraged but difficult due to sizes, lack of appropriate repositories, data privacy, etc. With AI, author attention to enhancing the transparency of their science methodologies is a complement to reproducibility and strengthens the overall rigor of a scientific publication.

Technology. While 21% of respondents believe that as long as their code is published somewhere they do not have to worry about issues of reproducibility, which also means that the vast majority of respondents (over 79%) are aware that publishing code in a repository is insufficient. Too, a full quarter of all respondents who submitted Appendix Descriptions in the years covered by this survey (2017-19), stated that they used containers for their submissions. While not a full answer to reproducibility, containers offer a way to bundle all the products that need to be shared within a single environment wherein the products are known to run. Driven by included scripts, the results can be more easily reproduced or tweaked by a researcher attempting to replicate the work (same scientific question, different data).

DISCUSSION

What can we conclude from the HPC community perceptions of reproducibility efforts? We first acknowledge that the survey response represents a small fraction of the HPC community, and it is likely that those who responded had something, positive or negative, that they wanted to offer. Nevertheless over 200 members of our community spoke to the topic by responding to the survey.

Positive Impacts The survey shows early evidence that the SC reproducibility initiative is having a positive impact on the quality of the research published as part of the SC technical program. The subject of scientific impact could be probed more deeply in a follow-on survey: did the SC appendices information actually make it easier to build off the research result? Did it contribute to enhanced trustworthiness?

Additionally, a full 90% of respondents are aware of the issues of reproducibility. From this measure, one can reasonably conclude that future SC authors are conducting research with the eventual AD/AE reporting in mind. That is, the reporting is having the effect of enhancing the rigor and transparency of science whether it is submitted to SC or elsewhere.

The HPC community is responding favorably to SC21 efforts to carry out rigorous artifact evaluation. SC21 has set up an arrangement with cyberinfrastructure resource providers whereby authors of accepted SC21 technical papers are provided a Virtual Machine (VM) to which they can deposit products that the SC21 AE/AE team then uses to verify the computational reproducibilty of the work. Over 50% of SC21 authors have indicated interest in having the AD/AE team carry out this task on their behalf [7].

Challenges of computational reproducibility. The NASEM report acknowledges the unique challenges that the HPC community faces with computational reproducibility and takes the step of recommending that funding agencies fund exploration of the limits of computational reproducibility in instances in which pure computational (bitwise) reproducibility is not reasonable. The recommendation advocates for *consistent computational results* [that] remain in step with the development of new computational hardware, tools, and methods.

There is an existing model for this expanded form of computational reproducibility already in use in the HPC community. This is, surprisingly enough, the SC annual Student Cluster Challenge (SCC). First developed in 2007, SCC provides an immersive high performance computing experience to undergraduates. Student teams choose to participate in the reproducibility challenge wherein they reproduce results from an accepted paper from the prior year's Technical Program. The students have limited compute resources available to them.

The SCC committee designs the reproducibility challenge experience for the undergraduate students. The committee is very similar to what computational and computer science educators do in the classroom when they provide a hands-on project experience for students: they design a constrained but achievable learning experience.

The SC20 SCC committee in providing a student experience undertook three important steps in consultation with the original author [4]:

- create dataset different from the one used in the author's published paper
- select a subset of metrics in the original paper as criteria upon which successful reproducibility is demonstrated
- interpret/translate results obtained in student constrained compute environments for their suitability in demonstrating reproducibility

By shaping the student experience, the SCC committee has actually created a nice replicability opportunity (per the NASEM study which defines *replicability* as the obtaining of consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data [5]).

While it is not the obligation of the original author to create a new dataset, it should certainly fall to the author to identify the critical metrics that demonstrate replication, and interpret the results for environments other than the original. Performance evaluation papers will offer a host of measures, some more critical than others. Which of these measures uniquely asserts a result as reproducible? And how much leeway in a result can occur when the next researcher doesn't have the same size environment, or exact tool stack?

Replicability of kind enabled by the SCC is one demonstrated form of certifying the validity of a result, thus helping alleviate the wicked problem limitations to reproducibility facing HPC.

Sustainability. Are the content and processes that the HPC community uses for reproducibility verification right for reuse of today's research results 5 years from now? That is, what information should be included in an SC22 paper, appendix, or shared material, that will make it easier to build upon by a graduate student who comes across the paper in 2027? Environments will have changed, making today's point-of-manuscript-submission verification only marginally helpful if not meaningless.

Reproducibility efforts by the HPC community today are largely focused on fairly stringent point-of-manuscript-submission reproducibility verification, where reproducibility evaluation has the potential to impact accept/reject decisions. Stringent gating is problematic for two reasons: i) it is too late in the publication process to be effective as a learning tool, and ii) the demands for rigor versus the high labor obligation are difficult to balance. To the former, while point-of-manuscript-submission reproducibility certainly gets the attention of researchers, is too late in the research process to be fully effective as an educational tool. The final steps of manuscript submission are not teaching moments as the stress of the deadline diminishes the intended learning outcome. Voluntary reproducibility, such as incorporated in SC21 reproducibility initiative through request for evaluation of the Artifact Evaluation (AE) appendix, removes the stress factor and could thus be seen as a positive form of community education if accompanied by abundant educational materials [7].

When reproducibility evaluation results are used in the decision to accept or deny manuscripts, the reproducibility process itself is subject to the same obligations as the original science: that of highest levels of rigor and transparency. It is difficult to balance sustained high levels of rigor with a process that inherently has high manual overhead and low intellectual value.

CONCLUSION

Our survey has shown that the SC community reproducibility effort has contributed to higher levels of awareness by past and future SC authors, suggesting greater practiced rigor and transparency. It hints at greater impact of the published papers of the SC conference series, suggesting a fruitful area for followup. Stringent point-of-manuscript-submission verification is problematic for a number of reasons that we have discussed; it needs to evolve. Given the difficulties of computational reproducibility in HPC, the community needs more experimentation in reproducibility/replicability content and processes. As remarked by a couple survey respondents, there needs to be better norms around attribution. Authors practicing transparency by sharing their code should have full confidence that peers will provide a proper citation to the original author. Finally, future forms of reproducibility should better decouple the community educational goals from goals that are designed to strengthen a research work's potential for greater scientific impact 5 years down the road.

The dataset and instrument for this survey is available at [3]. All identifying information has been removed. Free text entries referencing an individual or a role that could be traced to an individual have been redacted.

ACKNOWLEDGMENTS

Our thanks to JangDong Seo of the Indiana Statistical Consulting Center at Indiana University who carried out the statistical analysis of the survey results. Our thanks to Stephen Harrell, SC20 Student Cluster Competition chair for discussions, and Irene Qualters, Los Alamos National Laboratories, SC20 Technical Chair for suggesting the survey.

This manuscript has been authored in part by employee(s) of Brookhaven Science Associates, LLC under Contract No. DESC0012704. Malik acknowledges the following federal support NSF CNS-1846418, NSF ICER-1639759, NSF ICER-1661918, and the DoE BSSw Fellowship.

■ REFERENCES

1. Lorena A. Barba, ”Trustworthy computational evidence through transparency and reproducibility”, Computing in Science and Engineering, 23(1):58-64 (2021), IEEE Computer Society.
2. Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Administration and policy in mental health and mental health services research, 42(5), 533-544.
3. Pouchard, L. and B. Plale, ”Dataset: Assessing Reproducibility Initiative of IEEE/ACM Supercomputing Conference”, Indiana University protocol #2005780098, August 2020. https://doi.org/forthcoming
4. B. Plale and S. L. Harrell, Transparency and Reproducibility Practice in Large-Scale Computational Science: A Preface to the Special Section, IEEE Trans. Parallel Distributed Systems, 32(11), pp. 2607–2608, 2021
5. “Reproducibility and Replicability in Science: a Consensus Study Report”, H. V. Fineberg, Committee Chair, Na-

tional Academies of Science, Engineering, and Medicine, 2019

6. Pouchard, L., Baldwin, S., Elsethagen, T., Jha, S., Raju, B., Stephan, E., Tang, L. and Van Dam, K.K., 2019. Computational reproducibility of scientific workflows at extreme scales. The Int'l Journal of High Performance Computing Applications, 33(5), pp.763-776.
7. Malik, T., Artifact Description/Artifact Evaluation: A Reproducibility Bane or a Boon, 2021. Keynote at the Practical-Reproducible Evaluation of Computer Systems (P-RECS), held in conjunction with High Performance Distributed Computing (HPDC), 2021.

Beth A Plale is the Michael A and Laurie Burns McRobbie Bicentennial Professor of Computer Engineering at Indiana University Bloomington and Executive Director of the Pervasive Technology Institute. Plale received a PhD in computer science from the Watson School of Engineering at State University of New York Binghamton and carried out postdoctoral studies at Georgia Institute of Technology. Plale's research interests are in research infrastructure, open science and reproducibililty, and AI accountability. Plale is a member of IEEE. Contact Plale at plale@iu.edu.

Tanu Malik is an Associate Professor at the School of Computing, DePaul University. At DePaul, she directs the Data Systems and Optimization Laboratory. Tanu received her PhD in Computer Science from the Johns Hopkins University and was a Fellow at The University of Chicago. Her research interests span topics in data provenance, database systems, distributed systems, and cyber-infrastructure for scientific data management. Tanu received the 2019 NSF CAREER award for her work on computational reproducibility. She was also the 2019 Better Scientific Software Fellow. She is a member of ACM and IEEE. She can be reached at tanu.malik@depaul.edu

Line C. Pouchard is Applications Architect and Senior Researcher at the Computational Science Initiative, Brookhaven National Laboratory, Department of Energy, where she leads multidisciplinary teams to create new approaches for data management, curation, and scientific discovery at Extreme Scales. Her current research focuses on provenance in workflows, computational reproducibility, and enabling explainable and interpretable AI, applied to domains of interest to the DOE, including Scientific User Facilities. A recent award includes Best Paper at the In Situ Infrastructures for Enabling Extreme-scale Analysis and Visualization (ISAV 2020) workshop. Contact Pouchard at pouchard@bnl.gov.