

Reference ID: 11224958566_Maechling

Reference ID: 11224958566_Maechling

Submission Date and Time: 12/16/2019 2:46:54 AM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Consent Statement: "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

Consent answer: I consent to NSF's use and display of the submitted information.

Author Names & Affiliations

Submitting author: Philip Maechling - University of Southern California

Additional authors: None

Contact Email Address (for NSF use only): (hidden)

Research domain(s), discipline(s)/sub-discipline(s)

Computer Science;Geoscience

Title of Response

Data-Focused Cyberinfrastructure Needed for Model Validation

Abstract

For multiple NSF research domains, model validation is the missing link between scientific discovery and broad impact. Research organizations that want to achieve broad impact need model validation computing to achieve that impact. Based on experiences at the Southern California Earthquake Center (SCEC), I believe there is a gap in current NSF research cyberinfrastructure. Arguably, model validation

computing is currently not well supported by NSF cyberinfrastructure. Future data-focused cyberinfrastructure can provide the services and policies to support model validation computing, but this will require both new technical and administrative accommodations.

Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s). Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

In seismic hazard research, the goal is to accurately predict future ground motions across a wide range of ground motion frequencies across time scales from seconds to centuries.

Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s). Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

Technical cyberinfrastructure capabilities needed to support model validation computing include: 1.

Validation processing will require a variety of computing capabilities which are shared with continuous integration, continuous delivery (CI/CD) capabilities. For example, model validation testing may be implemented with CI/CD systems like Jenkins. But on a shared NSF computing system, like a Frontera-type system, there is often no obvious place to run a Jenkins server. Running it on a compute node isn't right. Running a Jenkins server externally, such as at SCEC, requires dedicated computing capabilities at SCEC which may not be available. Also, externally operated servers are likely to face issues such as two factor authentication, as they try to submit jobs. 2. Model validation testing often requires access to observational data. There is a variety of observational data access methods such as REST APIs. However, there is a wide variety of other access methods that should be supported. Users may want to run access software, like OpenDAP (an NCAR access protocol and software suite), on NSF computing resources. Users might want to access observational from compute nodes, as part of their validation tests. But if the observational access is inconsistent, then running data retrieval from the compute nodes would require the data access jobs running on compute nodes until access is completed. Possibly dedicated data transfer nodes, that share a mounted file system with the compute nodes would help users. 3. Validation introduces significant new scheduling requirements, probably requiring a

“meta-scheduler” of some type. In some cases, we re-validate codes after a change to the code, or to an input parameters. In other cases, however, models are re-validated on a regular schedule, such as nightly. At a minimum, a cron-type scheduler, that can submit jobs to slurm or other system scheduler will probably be needed. In some cases, a meta-scheduler that can determine whether “yesterdays” jobs have completed may be required, to determine whether it is time to running today’s jobs. 4.

Verification and validation jobs are nearly always extended heterogenous calculations that follow separate processing steps, such as, “run the forecast model, then compare against observations.” Users will definitely need to construct workflows that implement their validation tests. Tools to support composing and running heterogenous workflows, that include both large parallel, and large serial jobs, will be needed. 5. A concept used in validation testing is a data store called an “oracle.” An oracle represents storage of the correct results against which newly calculated results are compared. Oracle’s can be implemented as relational database or object stores, with some search capabilities. In many cases, validation tests will involve file to file comparison, so the oracle needs to support large file management. Support at the computing center for storage, annotating, discovering, and moving large, file-based, expected results data would help support validation testing. 6. After validation, distribution of tests results will be important. Users will likely want to distribute test results to groups for transparency. The system should support some way to create, and then post, test results on a webpage for distribution. Distribution of other tests results, possibly too large to distribute through a web page, are also needed.

Question 3 (maximum 300 words) – Other considerations. Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

Operational and allocation policies are also needed to support model validation, including: 1. The importance of validation testing may warrant special considerations in the allocation process. Both verification and validation testing tend to be repetitive, involving simulations with well-known solutions. System allocation policies should support simulations justified in terms of “new discoveries”. System allocation policies should also support allocations for verification and validation testing, even if the calculations are not expected to produce new discoveries. 2. Often, the computing requirements for validation are substantial because many models require testing at full-scale after significant changes. Also, the computing time requirements increase if problems are found. 3. There is an operational aspect to validation testing, not commonly found in many scientific simulations. Validation tests may need to run on a regular schedule. This can be supported with technical tools, such as the meta-scheduler discussed above. The operational system aspect of validation testing may also involve system policies, such as giving priority boosts to time-dependent jobs. Model validation testing is a multi-domain need. The model validation cyberinfrastructure capabilities described here would support computational research in the earth sciences. Further discussions with other domains may lead to further suggestions on how data-focused cyberinfrastructure can support model validation processing in the future.

Response to NSF 20-015, *Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research*

Reference ID: 11224958566_Maechling

-- End Submission --