 doi:10.20944/preprints202001.0176.v1

Beyond Modeling: A Roadmap to Community Cyberinfrastructure for Ecological Data-Model Integration

Istem Fer[1,*], Anthony K. Gardella[2,3], Alexey N. Shiklomanov[4], Shawn P. Serbin[5], Martin G. De Kauwe[6,7], Ann Raiho[8], Miriam R. Johnston[9], Ankur Desai[10], Toni Viskari[1], Tristan Quaife[11], David S. LeBauer[12], Elizabeth M. Cowdery[2], Rob Kooper[13], Joshua B. Fisher[14], Benjamin Poulter[15], Matthew J. Duveneck[16], Forrest M. Hoffman[17,18], William Parton[19], Joshua Mantooth[20], Eleanor E. Campbell[21], Katherine D. Haynes[22], Kevin Schaefer[23], Kevin R. Wilcox[24], Michael C. Dietze[2]

[1] Finnish Meteorological Institute, P.O. Box 503, 00101 Helsinki, Finland [2] Department of Earth and Environment, Boston University, Boston, MA 02215, USA [3] School for Environment and Sustainability, University of Michigan, Ann Arbor, MI 48109, USA [4] Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD 20740, USA [5] Brookhaven National Laboratory, Environmental and Climate Sciences Department, Upton, NY, 11973, USA [6] ARC Centre of Excellence for Climate Extremes, Sydney, NSW 2052, Australia. [7] Climate Change Research Centre, University of New South Wales, Sydney, NSW 2052, Australia. [8] Fish, Wildlife, and Conservation Biology Department, Colorado State University, Fort Collins, CO 80523, USA [9] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA [10] Department of Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, 1225 W Dayton St, Madison, WI 53706, USA [11] UK National Centre for Earth Observation and Department of Meteorology, University of Reading, Reading, RG6 6BB, UK [12] College of Agriculture and Life Sciences, University of Arizona, Tucson, AZ 85721, USA [13] NCSA (National Center for Supercomputing Applications), University of Illinois at Urbana Champaign, Urbana, IL, 61801-2311 USA [14] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA, 91109, USA [15] NASA Goddard Space Flight Center, Biospheric Sciences Lab., Greenbelt, MD 20771, USA [16] Harvard Forest, Harvard University. 324 North Main Street, Petersham, MA 01366, USA [17] Computational Earth Sciences Group and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6301, USA [18] Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN 37996, USA [19] Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA [20] The Fulton School at St. Albans, St. Albans, MO 63073, USA [21] Earth Systems Research Center, University of New Hampshire, Durham, NH 03824, USA [22] Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523, USA [23] National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309, USA [24] Ecosystem Science and Management, University of Wyoming, WY 8207, USA

*Corresponding author: istem.fer@fmi.fi, Finnish Meteorological Institute, P.O. Box 503, 00101 Helsinki, Finland

Abstract

In an era of rapid global change, our ability to understand and predict Earth's natural systems is lagging behind our ability to monitor and measure changes in the biosphere. Bottlenecks in our ability to process information have reduced our capacity to fully exploit the growing volume and variety of data. Here, we take a critical look at the information infrastructure that connects modeling and measurement efforts, and propose a roadmap that accelerates production of new knowledge. We propose that community cyberinfrastructure tools can help mend the divisions between empirical research and modeling, and accelerate the pace of discovery. A new era of data-model integration requires investment in accessible, scalable, transparent tools that integrate the expertise of the whole community, not just a clique of 'modelers'. This roadmap focuses on five key opportunities for community tools: the underlying backbone to community cyberinfrastructure; data ingest; calibration of models to data; model-data benchmarking; and data assimilation and ecological forecasting. This community-driven approach is key to meeting the pressing needs of science and society in the 21st century.

Keywords: community cyberinfrastructure, accessibility, reproducibility, interoperability, models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

Introduction

In an era of rapid environmental change, the scientific community is deeply invested in understanding and predicting nature's dynamics (Eyring et al. 2016a; Dietze et al. 2018; Rineau et al., 2019). Thankfully, recent decades have seen an explosion of environmental data globally due to advancements in a diverse range of field, lab and genomic technologies; a growth in network observatory and citizen science; a diverse constellation of remote sensing platforms; and information technology (and changes in perceptions of data sharing) that are delivering these data to us faster than ever before (LaDeau et al. 2017; Bond-Lamberty, 2018; Farley et al., 2018; Exbrayat et al., 2019; Shiklomanov et al., 2019a; Schimel et al., 2019). Process-based models play a critical role in translating data into mechanistic understanding of natural systems, as they provide us with the ability to synthesize knowledge about how the world works, reformulate that knowledge across organizational, spatial and temporal scales, and generate testable predictions from alternative hypotheses (Fisher et al., 2014; Medlyn et al., 2015; Norby et al., 2016; Rogers et al., 2017; Lovenduski and Bonan, 2017; Braghiere et al. 2019). Yet, having more data than ever before, we have not seen comparable progress in our capacity to forecast natural systems with our process-based models. For example, model projections out to the year 2100 do not agree on whether the land will be a carbon sink or source in response to climate change, and this has not changed despite years of apparent model improvement (Friedlingstein et al., 2006, 2014; Arora et al., 2019). The goal of this paper is to better characterize the bottlenecks that have obstructed the rates at which new information has been translated into knowledge and then integrated into models, and to lay out a roadmap for how to overcome these bottlenecks.

A more predictive global change science needs to be based on models that can capture important processes rather than merely reproducing the data (Medlyn et al., 2015; Walker et al., 2018; Bonan and Doney, 2018). Thus, modeling efforts should be geared towards generating hypotheses that are testable against data. Most of the modeling activities, though, are too limited in scope: conducted by a small subset of the research community ('modelers') using a small non-representative fraction of the data generated by the community (e.g. models are more likely to be constrained by standardized, high-volume high-level observational data than experimental manipulations or studies focused on low-level process details). As a result, most of the modeling activities end up comparing a single model (or sometimes a group of models) to an individual site or data type. This is in direct contrast with the incredibly diverse range of data generated by ecology as a discipline. More importantly, this approach frequently fails to actively engage the expertise in the non-modeler community, who are often better positioned to tell which observations can inform models and what processes are critical to capture.

Until modeling tools become more accessible, people generating the data will not be able to play a more active role in its ingest, interpretation, and hypothesis testing. This bottleneck not only impacts the pace and the quantity, but also the quality of the modeling efforts. The challenge of accessibility is further exacerbated by another bottleneck: uniqueness of models in terms of their operation. Since individual model components are technically different, and thus, have to be managed differently, modelers keep reinventing wheels around them instead of redirecting their productivity to automation or construction of reusable tools for the ecology community. The realm of modeling will remain elusive to non-modelers if each model has a unique learning curve. This lack of interoperability also hinders efforts to replicate findings in other studies and to perform regular model-based ecological analyses. The more reusable modeling tools become, the more accessible and reproducible they will get, as learned from genetics (Afgan et al., 2016) and atmospheric sciences (Eyring et al., 2019). Only then will it be possible to replicate and build on the work of others, and provide action upon information faster while pushing the frontiers of our understanding.

We argue that solution to these model-data bottlenecks lies in developing and supporting community-level cyberinfrastructure. Cyberinfrastructure refers to the computational environment where we can operate on data, simulate natural phenomena, perform model evaluation and scientific interpretation. In a good cyberinfrastructure, key modeling activities take place seamlessly. These activities include but are not limited to: i) obtaining and processing of data (data ingest), ii) estimating model parameters through statistical comparisons between models and real-world observations (calibration), iii) evaluating and comparing performance skills through standardized and repeatable multi-model tests (evaluation and benchmarking), and iv) combining model predictions with multiple observations to update our understanding of the state of the system (data assimilation).

In the following sections, we first present key features of a community cyberinfrastructure, next discuss specific challenges and solutions for each modeling activity, then finish by exploring implications for environmental management, policy and education. We provide specific recommendations for the modeler and developer community, the measurement community and the broader community throughout each section. Given the background of the authors, who came together during a workshop held by Predictive Ecosystem Analyzer (PEcAn) Project (pecanproject.org), many of the examples are focused on terrestrial carbon cycle and ecosystem models, but the principles highlighted are general across different systems and processes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

Cyberinfrastructure

> *Models and the tools associated with them currently live in silos. Instead, we advocate adopting common cyberinfrastructure tools that are accessible, reproducible, interoperable, scalable, and community-driven.*

Making models accessible

There should be few things more repeatable in science than running a deterministic model. In practice, running a process-based simulation model is often fraught with roadblocks to any new user or developer. Retrieving the source code of a model (which itself is not always possible) does not guarantee one can compile and run simulations. Even then, accessibility is not just an issue of getting the model to run or understanding model equations. A broader technical barrier exists in terms of the abilities required to perform complex analyses. A model-based ecological analysis rarely relies on just running a model once. For this reason, it is becoming more and more common to have "helpers" that accompany the models to reduce the barrier in their use (Duursma, 2015; Metcalfe et al. 2018; Bagnara et al., 2019). These accompanying codes are usually scripts, functions, or packages (e.g. R, Python) that provide a certain level of abstraction by taking control of the data stream in and out of the models, supporting more advanced visualization and analyses. In addition, they enable basic automation by specifying the steps performed in a modeling workflow, and repeating the sequence as needed. However, because the community has been slow to adopt community standards for model inputs and outputs, model helpers are usually specific to the model they are written for (sometimes they are not even compatible with a different version of the same model). Tackling this issue for individual models leads to redundant efforts and inhibits economies of scale that could be gained by sharing informatics tools and analytical workflows across communities (e.g. the Protocol for the Analysis of Land Surface Models (PALS), Abramowitz 2012; PEcAn, LeBauer et al., 2013; Earth System Model Evaluation Tool (ESMValTool), Eyring et al., 2016b; and International Land-Model Benchmarking (ILAMB), Hoffman et al. 2017).

Considering use cases for a community, rather than individual models or one-off projects, would facilitate designs of abstraction layers for working with models and data that are easier to standardize and generalize, and would result in fewer *ad hoc* solutions. Abstraction starts by determining the important tasks involved, planning and documenting purpose of each. Then, these individual actions would be modularized (Fig 1) to i) isolate the tasks and make internal modifications to their implementation without altering the overall behaviour of the system possible; ii) allow the reuse of the module package outside of the system independently; and iii) allow users to swap in/out alternative

algorithms/modules and customize their workflow for their own needs. Workflows orchestrate the smooth and continuous coordination of the modules to achieve the overall modeling activity successfully, to ensure that its replication is possible, and to reduce costs of manual operation. Indeed, automated workflows not only help save time but also avoid simple errors, such as having the output of one run overwrite a previous run. Such situations can arise easily during more advanced analyses that may require numerous runs if the user has to manually update a large number of settings that need to be changed every time the model is run. Overall, besides a massive reduction in redundant efforts, such common infrastructure will foster innovation and create an incentive for developers to make better, more sophisticated algorithms that have gone through more extensive testing (rather than one off scripts) as there is a larger community that can more easily adopt them.

Repeatability and transparency

Recording the full history of an analysis to enable versioning, repeatability, and transparency is known as provenance. Various approaches have been used to record provenance in model experiments, from recording run history in repositories (Medlyn et al., 2016) or in files that support rich metadata (e.g. Network Common Data Form) to utilizing a database infrastructure to track model inputs, settings, and simulation results. However, for larger workflows executing multiple ensembles, models, or experiments, embedding provenance in files can be limited, and tracking history with a database is preferred. Without workflow automation, it is hard to imagine a functioning system where researchers record every run they do in a database, since each modeling activity would involve a panoply of model-specific formats and approaches. Provenance tracking is crucial for i) the transparency of the modeling activity to ensure its reliability, and ii) replication of these analyses to build upon them. The latter particularly carries practical importance for complex cyclic workflows such as environmental forecasting where programs would run periodically for a long time, and automatic scheduling and restarting of jobs are crucial (Oliver et al., 2019). Full provenance of the overall workflow requires creating permanent records of the key metadata about each activity: exact versions of data sets and tools, parameters and settings, outputs produced, where to find them, as well as any logs and messages reported along the way (Dietze, 2017). The workflow and provenance tracking system themselves should also be version controlled to ensure a fully reproducible record (Ellison, 2010; Piccolo and Frampton, 2016).

The use of databases to track workflows provide a tremendous amount of transparency and repeatability if the database and workflow are both open community resources. Shared databases, whether centralized or distributed, create the possibility for globally unique identifiers being assigned to every model run, much like the accession numbers used in GenBank to keep track of genetic data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

(Clark et al., 2016). Such IDs could be used in publications (for an example, see Fer et al., 2018), providing readers a pointer to the full model output and the metadata required to repeat a model run. One goal is to work towards making reproduction of model runs as easy as accessing a DOI.

Interoperability

Even if all these components are successfully brought together, they would inevitably be written for specific versions of operating systems, computer languages, compilers, and software libraries, all reflecting the available resources to their developers at the time. The portability of models and cyberinfrastructure constitutes a roadblock itself for new users because current systems emerged from autonomous efforts by individuals and institutions to fill specific goals, often making model outputs effectively unverifiable.

Community cyberinfrastructure should be accessible to users without having to deal with obscure system requirements and dependencies. Fortunately, modern virtualization technologies offer a number of tools to make that possible. Virtual machines dissolve the boundaries that arise from having to work with a specific operating system as they allow users to run a whole operating system within another operating system, complete with the required software and all its dependencies. We recommend developer communities adopt recent light-weight containerization systems, such as Docker (www.docker.com) and Singularity (singularity.lbl.gov) that are even easier to install, set up, upgrade, and scale up with new locations to run the models. Containerization improves the portability of existing infrastructures so that they can be run reliably in new computing environments (for an example see FAIRifying eWaterCycle project https://www.ewatercycle.org). Combining containerized tools with cloud-based virtual services would provide an ultimate solution for interoperability, reproducibility and scalability as they allow remote operation, freeing the users from their local machine constraints altogether.

Data ingest

> *Current pipelines for processing the data that are needed to run and constrain models are cumbersome and inefficient. We advocate human- and machine-friendly community-scale approaches to foster effective discovery and reuse of both data and software.*

Data acquisition

Data play a critical role in the modeling activity: as covariates, drivers, and initial conditions; to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

constrain model parameters and states; and to test process representations and predictions (Dietze, 2017). However, linking data and models is not a trivial task. Due to their sheer volume and diversity, data can be difficult to locate and obtain, even if they are open access. Distributing data from repositories with accessible human-centric interfaces does not solve the problem, as sifting through large volumes of data manually is practically impossible. Even when the data is machine accessible, restrictive licenses and cumbersome web interfaces often limit automation of data download.

To make data discoverable and accessible, we recommend data producers use general-data repositories with publicly available Application Programming Interfaces (APIs) that provide machine-readable services (Hart et al., 2016). APIs are interfaces between data providers (servers) and data seekers (clients) that enable automated search and programmatic interaction with the data. For example, repositories within the DataOne federation are jointly searchable, allowing developers to leverage one set of tools for many sources. As we invest in community cyberinfrastructure, efforts can be focused on optimizing this exchange where problems can be solved once between widely-used common tools, rather than many times between every server and client.

Modeling activities often involve a subset of input or output data (e.g., a specific region or period) for which it makes more sense to take the computations to where the data are rather than bringing data to the computation, especially when the time to transfer data exceeds the time to process it (e.g., Google Earth Engine - earthengine.google.com). Thus, we recommend developers of these data services to expand their utility and flexibility of data access to move from a paradigm of having to download, expand and operate on data, to one where online services can subset, transform, or perform basic operations on the data. The valuable effort towards making the Coupled Model Intercomparison Project Phase 6 (CMIP6) data archive available on Google Cloud (cloud.google.com) where an ensemble of open source tools can perform large-scale computations on the datasets (https://pangeo.io) is a great example of this, and will make a substantial difference for groups who have limited local resources for such interactions with such large datasets.

Data standards

Assuming the data are obtained, the next hurdle becomes harmonizing and processing of the data. As data are generated by different individuals and organizations, often non-standardized terminologies and metrics are used during data collection and processing. Data are further stored in unique formats that may not match the unique formats models require. Even if the formats are standardized within communities (e.g. remote sensing data in hierarchical data format, flux data as comma-separated values file), there is still a considerable variation across communities whereas the same file format

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

could have been just as functional. While reducing the proliferation of both data and model formats would alleviate this in the long term, in the short-term tackling this problem at the individual level rather than the community level is inefficient (Fig 2, top panel).

Using standard data pipelines can remedy the redundant efforts put into building custom tools for specific needs. For example, consider the common problem of providing inputs for a model. If there are *n* sources of data available in the community and *m* models being used, then having every team work independently requires producing *m x n* converters. On the other hand, if a common pipeline with internal standards is used, this is reduced to an *m* + *n* problem: *n* conversions to a community standard, *m* conversions to model-specific formats (Fig 2, bottom panel). This also extends beyond data conversions to the development of tools and analyses. For example, if input data need to be extracted, downscaled, debiased, gap-filled, or have their uncertainties estimated, each of these steps does not need *m x n* variants but rather just one tool that can be applied to the standard. We recommend the ecological community leverage existing efforts in harmonizing data for these standard formats such as the Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP) standards (Huntzinger et al., 2016), the Climate and Forecast (CF) convention (Eaton et al., 2017), and ontologies that provide standardized vocabulary and semantic framework that is needed for large-scale integration of heterogeneous ecological data (Cooper et al., 2013; Walls et al., 2014; Stucky et al., 2018). In due course, the growing cyberinfrastructure community will facilitate centralized discussions of data harmonization for emerging needs.

When using standards, we not only write less code, but also debug fewer lines with more people. Hence, the more these tools are used and tested, the less error-prone they will become. More importantly, this approach is much more scalable. We incur a small cost of effort for every new data stream added (i.e. write only one processor that converts data format to the intermediate standard), and get access to *m* models (Fig 2, bottom panel). Vice versa, for each new model added, we only need to write one processor that converts the intermediate standard to the model format, and the new model is immediately able to access *n* data streams. We recommend modeling community to adopt this scalable approach that will allow us to tackle ever bigger problems.

Calibration

> *Current calibration tools are not set up to work with process-based models, which hinders the application of more advanced calibration techniques. We advocate adopting statistical and computational solutions developed and generalized by domain experts under community cyberinfrastructure.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

There are a number of important statistical and computational challenges in using data to constrain model parameters, also known as calibration (Dietze et al. 2013; MacBean et al. 2016). Calibration is a persistent problem in process-based modeling activities, as these models can have large numbers of parameters. Some parameters may be directly informed by ecological trait data (e.g., specific leaf area, turnover rates) for which meta-analysis informatics tools can pull values together based on selection of traits from open-access, machine-readable, curated databases (LeBauer et al. 2013, 2018; Shiklomanov et al. 2019b). However, a non-negligible portion of these parameters are not (easily) measurable or directly estimable, therefore there is a need to estimate these values indirectly using numerical methods (Hartig et al 2012). While doing this, we recommend the community treat parameters as distributions to reflect the incompleteness of our knowledge and quantify our initial uncertainty about their values (priors) (LeBauer et al 2013). This way, we can understand how much parameters contribute to the overall model uncertainty and target them in calibration. Next, using the Bayesian approach, we can update our understanding about these parameter values. The product of the Bayesian estimation then becomes the multivariate probability distribution. This distribution, known as the posterior, accounts for the uncertainties in all the targeted parameters and their covariances, rather than providing a single point estimate that would come from traditional optimization approaches.

Standard tools for Bayesian algorithms such as BUGS (http://www.openbugs.net), JAGS (http://mcmc-jags.sourceforge.net/), NIMBLE (r-nimble.org) and STAN (mc-stan.org), that are otherwise used widely to fit models, are not set up to work with external ‘black box’ models. Process-based models cannot simply be “plugged-into” these tools and are often too complicated to be re-implemented in the native language of these software. In addition, current off-the-shelf Bayesian approaches generally have limited support or no functionality to re-read their own outputs (posteriors) as new inputs (priors) which is critical for iterative updating of the calibration analysis as new data become available. Native implementations of these numerical techniques require non-trivial amount of computational and statistical expert knowledge. As a result, models are frequently used uncalibrated (or hand-tuned) due to lack of available tools. But when they are calibrated, it is rarely done with multi-criteria (i.e. fitting is usually done using one data source only) which ignores the fact that these models are sum of different sub-components and there are covariances between these subcomponents that might be essential for the overall model performance. Even when a model is calibrated for one setting (e.g. for a site or for a period of time), it does not guarantee good performance at another setting (e.g. at a new site or a new period) because there is variability and heterogeneity in natural systems. While calibrating models with data aggregated from all settings ignores this heterogeneity and variability, more flexible techniques such as hierarchical Bayesian calibration can account for it (Dietze, 2017). Hierarchical Bayesian calibration can also provide directions for model development, but there are even fewer available tools

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

for their standard implementation with external models. Assessment of uncalibrated (or naively calibrated) models can cause poor calibration to be mistaken for inadequate model structure or mask real problems with the model structure (Luo et al., 2016), hindering overall progress in model development.

Within a community cyberinfrastructure, the challenges of developing advanced calibration tools only need to be faced by experts in statistical and computational domains. Software alternatives for calibration of 'black-box' models are becoming increasingly available. The R package BayesianTools (Hartig et al., 2017) and ecological model-data informatic toolboxes such as PEcAn Ecosystem Analyzer (PEcAn) (Fer et al., 2018) and EcoPAD (Huang et al., 2019) are promising examples. PEcAn also has a hierarchical Bayesian calibration scheme implemented to account for site to site variability and heterogeneity. For the experts in statistical and computational domains, we recommend augmenting and improving hierarchical tools for the community that can account for all kinds of ecological variability and heterogeneity. We further recommend model developers to develop their models with their coupling to calibration workflows in mind. When that is not the case, for example, when the model's runtime prohibits its calibration through computationally expensive algorithms, community tools can also provide statistical solutions which are typically more advanced than an average user can implement (Fer et al., 2018).

When these solutions are implemented and generalized around the aforementioned data ingest, model execution pipelines, they can seamlessly link multifarious data with multiple models. As the calibration workflows are tracked and recorded by the shared databases, results from one analysis can readily be used by a subsequent analysis elsewhere without having to repeat it. Investing in such standardization and generalization will not only allow a wider audience to adopt these methods as common practices, but also foster progress on developing novel calibration techniques rather than spending energy on debugging and solving similar issues.

Model intercomparison and benchmarking

> *Benchmarking a suite of models with varying levels of process representation, complexity, and scale is logistically challenging. We advocate for community cyberinfrastructure that allows current model intercomparison projects to become persistent benchmarks, and domain experts to take the lead in setting up and performing these benchmarks.*

Models, at their essence, formalize the hypotheses about how natural systems work (Walker et al., 2018). Comparing models to data allows hypothesis testing and model evaluation (Fisher et al., 2014;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

Best et al., 2015). To verify progress, and assess the tradeoffs between model parsimony and complexity, results from the previous model versions can be set as the performance benchmark while models change through time (Luo et al., 2012; Best et al., 2015). Although this is a regular practice for some of the modeling groups, it is still remarkably rare to put an ecological model through all its past assessment exercises with performance expectations every time it is updated because such a workflow has not been automated (Kelley et al., 2013; Best et al., 2015). Significant costs associated with an automated benchmarking workflow involve importing the benchmark data (e.g. deciphering file formats, ensuring consistent variable names and units), aligning it with model outputs (e.g. matching or aggregating over timesteps and subsetting, reconciling quantities across different scales), performing benchmarking analysis (e.g. using appropriate scoring across different model variables and a suite of metrics), and converting those activities into persistent benchmarks. While clearly challenging, an example of a benchmarking framework that has attempted to address these issues is the ILAMB framework (Collier et al., 2018). However, even within this system a number of challenges remain, including how to specify expectations of performance for the models and deal with data sets and metrics that are incomplete or inconsistent with each other, among numerous others (Hoffman et al., 2017; Collier et al., 2018).

Assessment of model skills and representations are also beneficial when multiple models are considered to identify shortfalls or generalizability of different approaches (Schwalm et al., 2019), and to specify areas where mechanistic understanding is incomplete (Tuomi et al. 2008). Model intercomparison projects (MIPs) have also been constructed to estimate the range of possible responses to future change or processes at large scales and to identify commonalities, divergence, and uncertainties across models (Friedlingstein et al. 2006, 2014; Palosuo et al. 2012; Todd-Brown et al. 2013; Huntzinger et al., 2017; Arora et al., 2019). While MIPs are scarce in the rest of ecology, they are relatively common for terrestrial ecosystem models (Warszawski et al., 2014; De Kauwe et al., 2014; Rollinson et al., 2017; Huntzinger et al., 2017; Müller et al., 2019. Also see Hoffman et al. 2017, Free-Air Carbon Dioxide Enrichment (FACE) model-data synthesis project https://facedata.ornl.gov/facemds/ and references therein). However, MIP benchmarking has all the challenges of single model benchmarking, and many more.

Traditional MIPs are indeed logistically challenging to coordinate (Fig 3, top panel). Modeling groups need to minimize the differences in their results due to differences in their setups (e.g. a common driver dataset might need temporal downscaling for some models and aggregation for others, without introducing artificial differences due to these schemes). Likewise, some groups may join MIPs with uncalibrated models whereas others calibrate theirs (see *Calibration* section), making it hard to compare “apples to apples” (Seidel et al. 2018). Furthermore, due to the cost of running and managing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

a MIP, including the potentially large storage requirements, model output requests are typically kept to a minimum (e.g. either a select set of outputs or limited temporal scales, such as monthly or annual) which limits the ability to aid the interpretations with additional outputs. For example, MIPs largely focus on single model realizations which can lead to falsely overconfident decisions about model performances. More importantly, it is hard to expand an existing MIP as new data, new models, and new metrics arrive. For example, FACE-MIP would benefit from repeating it with the latest generation of CMIP6 land schemes (De Kauwe et al., 2014; Hoffman et al., 2017). Likewise, the teams participating in the Arctic-Boreal Vulnerability Experiment (ABoVE) stated the need for a community cyberinfrastructure to support continuous benchmarking as models advance and missing datasets or uncertainties are identified (Fisher et al., 2018).

Many of the benefits of community cyberinfrastructure are particularly valuable for MIPs and MIP benchmarking: standardization of both inputs and outputs, automated workflows (including calibration), provenance tracking and troubleshooting are already included in the process of embedding each individual model in the system (Fig 3, bottom panel). Within the community cyberinfrastructure, it follows naturally to propagate uncertainty and generate ensembles of model outputs, for additional variables and processes, as and when they are needed. We recommend that the community move towards model benchmarking that accounts for model and benchmark uncertainty and leverage this information when computing model scores (e.g. benchmarking that takes into account the uncertainty bounds in model and observation to compute a score based on overlap probability). Moreover, once a model is added to this framework, it becomes trivial to add its alternative versions, benchmark against existing MIPs and seamlessly feedback to future model developments. We further recommend model developers to enable functionality that allows direct comparison to observations when possible. For example, including a radiative transfer module in models would allow direct comparison of model predictions with reflected spectral radiance as measured by the satellites (Huang et al., 2019). This way, instead of using modeled remote sensing data products (e.g. leaf area index, soil moisture) whose uncertainties are harder to determine, researchers can directly track uncertainties in the observations. Such examples could be extended to other types of standard observations. Bringing models to data, rather than the other way around, may eventually reduce artificial inconsistencies between data sets that stem from additional manipulations for making data and models match. Community cyberinfrastructure would facilitate compilation of such standard data sets that our models need to be able to reproduce repeatedly. Within or in addition to existing frameworks such as ILAMB and PEcAn, interactive environments (e.g. Rstudio/Jupyter) would allow users to perform more extensive analyses with pre-loaded and aligned models and data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

Who sets up the benchmarks?

Hoffman et al. (2017) stated that "Developing metrics that make appropriate use of observational data remains a scientific challenge that should be addressed through synthesis activities in collaboration with the modeling and observational communities." Indeed, the inaccessibility of current modeling practices to non-modelers frequently, and inadvertently, excludes domain experts from the process of confronting models with data and knowledge. Even before the challenges of running models, it is nearly impossible for empiricists to keep abreast of which models exist, their most updated version, and their respective strengths and weaknesses (Schwalm et al., 2019). The balance cannot be restored without concurrently increasing modeling literacy and lowering the technical barrier for the modeling activities (Seidl, 2017).

Through community cyberinfrastructure, domain experts can easily participate in or lead a MIP. For example, with input/output standardization and data harmonization, the person leading the MIP no longer needs to be concerned with multiple file formats and model-specific terminology. As cyberinfrastructure automates tedious activities associated with a MIP such as aligning model outputs with data, calibration, and uncertainty propagation, experts can focus on their analysis rather than the logistics, making modeling activities more relevant for their science. We further recommend developers encode model structural characteristics as traceable metadata. Although there are preliminary examples of this (e.g. MsTMIP encoding presence and absence of process representations, Huntzinger et al., 2016), standards need to be developed by the community to provide information about key structural characteristics of the models. As a result, process representations that repeatedly perform below-average across multiple MIPs can be considered rejected hypotheses (Schwalm et al., 2019). In time, the community cyberinfrastructure would serve as a central database, allowing users to discover new models and to evaluate their updated versions with minimal technical barriers.

Putting data-model comparisons into the hands of the data generators and disciplinary experts, rather than the “modeler” minority, is absolutely critical to scaling up and addressing the bottleneck that only a small fraction of the data collected by ecologists (often under the argument of improving projections) ever makes its way into models. By centralizing these comparisons into databases, community cyberinfrastructure then allows the modeling minority to focus on learning from their colleagues and improving models, rather than the status quo where the overwhelming majority of their time is spent on mundane informatics issues.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

Data assimilation and ecological forecasting

> *Establishing automated data assimilation and forecasting pipelines is much costlier than a single lab can afford. We advocate adopting a community approach that allows regular updating and assessment of ecological forecasts, and forecasting many different aspects of ecology.*

To move ecology and environmental science up to speed with the pace of global change, the nature of the relationship between ecological models and data has to be reconsidered. While most ecological forecasts are long-term (e.g. 2100 projections), there is much to be learned from making short-term forecasts that can be tested and updated as new observations become available (Fox et al., 2009; Dietze et al., 2018; Huang et al, 2019). Data assimilation methods allow formal fusion of data and modeled states to assess and update forecasts.

Not unlike calibration, data assimilation methods also require advanced statistical and computational expertise. Ecological models and data frequently violate the standard statistical assumptions that are foundational to assimilation algorithms developed in other disciplines (Dietze 2017; Raiho, 2019). As mentioned earlier, standard Bayesian tools tend not to read their own outputs as inputs, which is particularly important for iterative forecasting. Likewise, process-based models are not often developed with data assimilation in mind, and often lack the ability to restart exactly from a state they themselves write out, which is a key feature required by data assimilation algorithms. Making a forecast operational also requires a higher level of repeatability and efficient scheduling of cyclic workflows where large number of jobs are executed at regular intervals and each forecast cycle depends on previous ones (Oliver et al., 2019). Among the tasks required, open archiving, community standards, and a full uncertainty accounting and propagation have proven to be prohibitively difficult (White et al., 2019). Overall, the breadth of expertise and investment of resources needed to set up a forecasting pipeline using state-of-the-art data assimilation methods often exceeds the limits of individualistic efforts (White et al., 2019).

Community-level development of automated data assimilation and forecasting pipelines provides a key economy of scale and builds upon many of the features already discussed: informatics tasks of gathering and processing new data, managing the execution of analytical workflows, and publicly archiving and reporting results. These features are integral to the vision for such an infrastructure and could then be coupled to, and build upon, existing community tools for workflow scheduling (Cylc, Oliver et al. 2019) and data assimilation (Data Assimilation Research Testbed, Fox et al., 2018; PEcAn, Viskari et al., 2015, Raiho, 2019; Land Surface Data Toolkit, Arsenault et al. 2018). These

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

existing community tools are often searching for opportunities to collaborate with a wide array of teams to expand the capabilities of their systems. In order to improve and augment ecological forecasts, more collaborations between ecologists and data assimilation scientists need to be formed for which community cyberinfrastructure can provide a practical platform. Similar to what the atmospheric science community has achieved, ecological and environmental science communities will not only make ecology more relevant to the society but will also transform basic ecological science and theory by investing in data assimilation and forecasting tools.

Conclusion

Scientists, managers, and policy makers all increasingly rely on models to understand the impact of decisions on ecological processes (Arneth et al. 2014; Pongratz et al., 2018; Smith et al., 2019). Yet, currently it takes too long to “turn the crank” on our deductive machinery that we are rarely able to effectively employ our models for decision making. If we want to bring the time frames associated with model-data integration in line with the pressing needs of managers, policymakers, and society more broadly, community cyberinfrastructure is the engine to do this.

Admittedly, there is an initial cost placed on the community to develop this infrastructure, and on individuals to generalize their approaches for broad-ranging use as they get involved with the cyberinfrastructure development. It becomes worthwhile to invest the time and the effort if these tools are used by the wider community, as it ultimately accelerates the communication and productivity on the whole (Leprevost et al., 2014; Bond-Lamberty et al., 2016). Contributing data, code, and methods to shared community tools also ends up being faster in the long run by preventing time loss from having to correct and repeat erroneous practices as it increases the overall quality of the job done (Easterbrook, 2014). Making this a team effort and extending tools to new people will allow us to benefit from others’ developments and to free our time for intellectual productivity, which will lead to new scientific discoveries.

As long as common cyberinfrastructure tools are Open Source licensed, their uptake and support by the community will allow their ongoing maintenance and development. For this to work, developers need to adhere sound development practices at all times (Leprevost et al., 2014; Oliver et al. 2019). As an example, 55 developers, with a core of fewer than 20, have contributed to the community tool PEcAn (https://github.com/PecanProject/pecan) to date. New code additions are reviewed by peers and are not merged in unless documented thoroughly, for both developers and users. Every submission goes through a set of automatic tests to verify that the code is working as expected. Significant revisions and additions are discussed and agreed upon in online and offline meetings. For

keeping the community tools up and running, all these aspects are directly addressed in the way they are developed and taught. Ultimately, it is important that funding agencies support the sustainability of these community tools as critical components of the community's collective scientific infrastructure in a similar way they do with the physical infrastructure (field stations, sensor networks, satellites) and data repositories.

To build a community where more people play a more active role in confronting models with data, there will be a need for significant changes in how all students are trained in informatics, statistics and modelling. Community tools not only provide scalable solutions but also a scaffold for bringing diverse groups together and a platform for training. Just as the training required to drive a car is very different from the training required by a mechanic to repair a car or by an engineer to design a car, the training required to use community cyberinfrastructure tools will be different, and more accessible, from that required for those who build them.

The goal of community cyberinfrastructure is to facilitate and accelerate the rapid proliferation and creation of knowledge about the biosphere. New communities of model users should not have to wait until they learn how to navigate complex computational architectures before gaining access to the latest information on ecosystem science. If the barriers to entry to use the latest models and data are lowered, then decisions will be made with better information, and scientific problems are solved more quickly. Just as Geographic Information Systems (GIS) has become standardized among a multitude of domains such as environmental management and consulting, so could model-based ecological analysis and forecasting.

Process-based models, though imperfect, are our only window into the future functioning of ecosystems under global change. The next generation of ecological models will need to ingest increasingly diverse data to inform and test new process representations and scaling approaches, allow rapid detection and explanation of global change patterns, and even possibly allow them to be prevented. Their application and integration in operational forecasting and decision support tools will not get any easier. This need is now more pressing than ever. To achieve ecological model-data integration in a way that is transparent, easily communicable, and scales up to the size and diversity of the ecological community, we must invest in community cyberinfrastructure.

References (79)

Abramowitz G. 2012. Towards a public, standardized, diagnostic benchmarking system for land surface models. *Geosci. Model Dev.* **5**: 819–827. doi:10.5194/gmd-5-819-2012.

Afgan E, Baker D, van den Beek M, et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update, *Nucleic Acids Research*. **44:** W3–W10, https://doi.org/10.1093/nar/gkw343

Arneth A, Brown C, and Rounsevell M. 2014. Global models of human decision-making for land-based mitigation and adaptation assessment. *Nature Clim Change* **4:** 550–557 doi:10.1038/nclimate2250

Arora VK, Katavouta A, Williams RG, et al. in review, 2019. Carbon-concentration and carbon-climate feedbacks in CMIP6 models, and their comparison to CMIP5 models. *Biogeosciences Discuss*. https://doi.org/10.5194/bg-2019-473

Arsenault KR, Kumar SV, Geiger JV, et al. 2018. The Land surface Data Toolkit (LDT v7.2) – a data fusion environment for land data assimilation systems. *Geosci. Model Dev.* **11:** 3605–3621, https://doi.org/10.5194/gmd-11-3605-2018

Bagnara M, Gonzalez RS, Reifenberg S, Steinkamp J, Hickler T, Werner C, Dormann CF, Hartig F. 2019. An R package facilitating sensitivity analysis, calibration and forward simulations with the LPJ-GUESS dynamic vegetation model. *Environmental Modelling & Software.* **111:** 55-60. https://doi.org/10.1016/j.envsoft.2018.09.004

Best MJ, Abramowitz G, Johnson HR, et al. 2015. The Plumbing of Land Surface Models: Benchmarking Model Performance. *J. Hydrometeor.* **16:** 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Bonan GB, and Doney SC. 2018. Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. Science. **359**. 10.1126/science.aam8328

Bond-Lamberty B, Smith AP, Bailey V. 2016. Running an open experiment: transparency and reproducibility in soil and ecosystem science. *Environ. Res. Lett*. **11:** 084004. doi: 10.1088/1748-9326/11/8/084004

Bond-Lamberty B. 2018. Data sharing and scientific impact in eddy covariance research. *Journal of Geophysical Research: Biogeosciences*, **123:** 1440–1443. https://doi.org/10.1002/2018JG004502

Braghiere RK, Quaife T, Black E, He L, and Chen JM. 2019. Underestimation of global photosynthesis in Earth System Models due to representation of vegetation structure. *Global Biogeochemical Cycles.* 33. https://doi.org/10.1029/2018GB006135

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2016. GenBank. *Nucleic Acids Res*. **44:** D67–D72. doi:10.1093/nar/gkv1276

Cooper L, Walls RL, Elser J, et al. 2013. The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.* **54:** e1. doi:10.1093/pcp/pcs163

Collier N, Hoffman FM, Lawrence DM, et al. 2018. The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*. **10:** 2731– 2754. https://doi.org/10.1029/2018MS001354

De Kauwe MG, Medlyn BE, Zaehle S, et al. 2014. Where does the carbon go? A model–data intercomparison of vegetation carbon allocation and turnover processes at two temperate forest free-air CO_2 enrichment sites. *New Phytol*. **203:** 883-899. doi:10.1111/nph.12847

Dietze MC, Lebauer D, and Kooper R. 2013. On improving the communication between models and data. *Plant, Cell & Environment,* **36**: 1575-1585.

Dietze MC. 2017. Ecological Forecasting(Princeton Univ Press, Princeton).

Dietze MC, Fox A, Beck-Johnson LM, et al. 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proc Natl Acad Sci*. **115**: 1424–1432. doi:10.1073/pnas.1710231115

Duursma R. 2015. Maeswrap: Wrapper Functions for MAESTRA/MAESPA. R package version 1.7. https://CRAN.R-project.org/package=Maeswrap

Easterbrook, S. 2014. Open code for open science? *Nature Geosci.* **7:** 779–781 doi:10.1038/ngeo228

Eaton B, Gregory J, Drach B, et al. 2017. Netcdf Climate and Forecast (CF) metadata conventions. http://cfconventions.org/

Ellison AM. 2010. Repeatability and transparency in ecological research. *Ecology*. **91:** 2536-2539. doi:10.1890/09-0032.1

Exbrayat, JF, Bloom AA, Carvalhais N, et al. 2019. Understanding the Land Carbon Cycle with Space Data: Current Status and Prospects. *Surv Geophys*. **40:** 735. https://doi.org/10.1007/s10712-019-09506-2

Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, and Taylor KE. 2016a. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev*. **9:** 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Eyring V, Righi M, Lauer A, et al. 2016b. ESMValTool (v1.0) – A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. Geosci. Model Dev. **9:** 1747–1802. doi:10.5194/gmd-91747-2016.

Eyring V, Cox PM, Flato GM, et al. 2019 Taking climate model evaluation to the next level. *Nature Clim Change* **9:** 102–110. doi:10.1038/s41558-018-0355-y

Farley SS, Dawson A, Goring SJ, Williams JW. 2018. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*. **68:** 563–576. https://doi.org/10.1093/biosci/biy068

Fer I, Kelly R, Moorcroft PR, Richardson AD, Cowdery EM, and Dietze MC. 2018. Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation, *Biogeosciences*. **15:** 5801–5830. https://doi.org/10.5194/bg-15-5801-2018

Fisher JB, Huntzinger DN, Schwalm CR, and Sitch S. 2014. Modeling the terrestrial biosphere. *Annual Review of Environment and Resources*. **39:** 91-123 https://doi.org/10.1146/annurev-environ-012913-093456

Fisher JB, Hayes DJ, Schwalm CR, et al. 2018. Missing pieces to modeling the Arctic Boreal puzzle. *Environ. Res. Lett.* **13:** 020202 https://doi.org/10.1088/1748-9326/aa9d9a

Fox A, Williams M, Richardson AD, et al. 2009. The REFLEX Project: Comparing Different Algorithms and Implementations for the Inversion of a Terrestrial Ecosystem Model against Eddy Covariance Data. *Agricult. Forest Meteorol*. **149:** 1597–1615.

Fox A, Hoar TJ, Anderson JL, et al. 2018. Evaluation of a data assimilation system for land surface models using CLM4.5. *Journal of Advances in Modeling Earth Systems*. **10:** 2471– 2494. https://doi.org/10.1002/2018MS001362

Friedlingstein P, Cox P, Betts R, Bopp L, von Bloh W, et al. 2006. Climate-carbon cycle feedback analysis: Results from the C4MIP model intercomparison. *J. Clim.* **19**: 3337–53

Friedlingstein P, Meinshausen M, Arora VK, Jones CD, Anav A, et al. 2014. Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *J. Clim.* **27**: 511–26

Hart EM, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, et al. 2016. Ten Simple Rules for Digital Data Storage. *PLoS Comput Biol.* **12:** e1005097. https://doi.org/10.1371/journal.pcbi.1005097

Hartig F, Dyke J, Hickler T, Higgins S, O'Hara R, Scheiter S, and Huth A. 2012. Connecting dynamic vegetation models to data – an inverse perspective. *Journal of Biogeography.* **39**: 2240-2252.

Hartig F, Minunno F, Paul S. 2017. BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics. R package version 0.1.3

Hoffman FM, Koven CD, Keppel-Aleks G, et al. 2017. International Land Model Benchmarking (ILAMB) 2016 Workshop Report, DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, doi:10.2172/1330803.

Huang J, Ma H, Sedano F, et al. 2019. Evaluation of regional estimates of winter wheat yield by assimilating three remotely sensed reflectance datasets into the coupled WOFOST–PROSAIL model. *European Journal of Agronomy*. **102:** 1-13. https://doi.org/10.1016/j.eja.2018.10.008

Huang Y, Stacy M, Jiang J, et al. 2019. Realized ecological forecast through an interactive Ecological Platform for Assimilating Data (EcoPAD, v1.0) into models. *Geosci. Model Dev.* **12:** 1119–1137. https://doi.org/10.5194/gmd-12-1119-2019

Huntzinger DN, Schwalm CR, Wei Y, et al.. 2016. NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs (version 1) in Standard Format. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1225.

Huntzinger DN, Michalak AM, Schwalm C. et al. 2017. Uncertainty in the response of terrestrial carbon sink to environmental drivers undermines carbon-climate feedback predictions. *Sci Rep* **7:** 4765. doi:10.1038/s41598-017-03818-2

Kelley DI, Prentice IC, Harrison SP, Wang H, Simard M, Fisher JB, and Willis KO. 2013. A comprehensive benchmarking system for evaluating global vegetation models. *Biogeosciences*. **10:** 3313–3340, https://doi.org/10.5194/bg-10-3313-2013

LaDeau SL, Han BA, Rosi-Marshall EJ, et al. 2017. The Next Decade of Big Data in Ecosystem Science. *Ecosystems* **20:** 274–283. doi:10.1007/s10021-016-0075-y

LeBauer DS, Wang D, Richter KT, Davidson CC, and Dietze MC. 2013. Facilitating feedbacks between field measurements and ecosystem models. *Ecol. Monogr.* **83:** 133–154. https://doi.org/10.1890/12-0137.1

LeBauer D, Kooper R, Mulrooney P, Rohde S, Wang D, Long SP, and Dietze MC. 2018. BETYdb: a yield, trait, and ecosystem service database applied to second-generation bioenergy feedstock production. *GCB Bioenergy*. **10:** 61-71. doi:10.1111/gcbb.12420

Leprevost FV, Barbosa VC, Francisco EL, Perez-Riverol Y and Carvalho PC. 2014. On best practices in the development of bioinformatics software. *Front. Genet*. **5**: 199. doi: 10.3389/fgene.2014.00199

Lovenduski NS, and Bonan, GB. 2017. Reducing uncertainty in projections of terrestrial carbon uptake. Environmental Research Letters.**12** http://dx.doi.org/10.1088/1748-9326/aa66b8

Luo Y, Randerson JT, Abramowitz G, et al. 2012. A framework for benchmarking land models, Biogeosciences. **9:**3857–3874, https://doi.org/10.5194/bg-9-3857-2012

Luo Y, Ahlström A, Allison SD, et al. 2016. Toward more realistic projections of soil carbon dynamics by Earth system models, *Global Biogeochem. Cycles*. **30:** 40– 56. doi:10.1002/2015GB005239.

MacBean N, Peylin P, Chevallier F, Scholze M, and Schürmann G. 2016. Consistent assimilation of multiple data streams in a carbon cycle data assimilation system. *Geoscientific Model Development.* **9:** 3569-3588.

Medlyn B, Zaehle S, De Kauwe M, et al. 2015. Using ecosystem experiments to improve vegetation models. *Nature Clim Change* **5:** 528–534. doi:10.1038/nclimate2621

Medlyn B, De Kauwe M, Zaehle S, et al. 2016. Using models to guide field experiments: a priori predictions for the CO2 response of a nutrient- and water-limited native Eucalypt woodland. Global Change Biology. Zenodo. http://doi.org/10.5281/zenodo.47282

Metcalfe P, Beven K., and Freer J. 2018. dynatopmodel: Implementation of the Dynamic TOPMODEL Hydrological Model. R package version 1.2.1. https://CRAN.R-project.org/package=dynatopmodel

Müller C, Elliott J, Kelly D, et al. 2019. The Global Gridded Crop Model Intercomparison phase 1 simulation dataset. *Sci Data* **6:** 50. doi:10.1038/s41597-019-0023-8

Norby RJ, De Kauwe MG, Domingues TF, et al. 2016. Model–data synthesis for the next generation of forest free-air CO_2 enrichment (FACE) experiments. *New Phytol.* **209:** 17-28. doi:10.1111/nph.13593

Oliver H, Shin M, Sanders S, et al. 2019. Workflow automation for cycling systems. Computing in Science & Engineering. **21:** 7-21. doi: 10.1109/MCSE.2019.2906593

Palosuo T, Foereid B, Magnus S, et al. 2012. A multi-model comparison of soil carbon assessment of a coniferous forest stand. *Environmental Modelling & Software*. **35:** 38-49. https://doi.org/10.1016/j.envsoft.2012.02.004.

Piccolo SR, and Frampton MB. 2016. Tools and techniques for computational reproducibility, *GigaScience*. **5** https://doi.org/10.1186/s13742-016-0135-4

Pongratz J, Dolman H, Don A, et al. 2018. Models meet data: Challenges and opportunities in implementing land management in Earth system models. *Glob Change Biol*. **24:** 1470– 1487. https://doi.org/10.1111/gcb.13988

Raiho A. 2019. Seeing the Trees through the Forest: Understanding Community Ecology's Influence on Long Term Ecosystem Dynamics. PhD Thesis. University of Notre Dame.

Rineau F, Malina R, Beenaerts N. et al. 2019. Towards more predictive and interdisciplinary climate change ecosystem experiments. *Nat. Clim. Chang.* **9:** 809–816 doi:10.1038/s41558-019-0609-3

Rogers A, Medlyn BE, Dukes JS, etal. 2017. A roadmap for improving the representation of photosynthesis in Earth system models. *New Phytol*. **213:** 22-42. doi:10.1111/nph.14283

Rollinson CR, Liu Y, Raiho A, et al. 2017. Emergent climate and CO_2 sensitivities of net primary productivity in ecosystem models do not agree with empirical data in temperate forests of eastern North America. *Glob Change Biol*. **23:** 2755-2767. doi:10.1111/gcb.13626

Schimel D, Schneider FD, and JPL Carbon and Ecosystem Participants. 2019. Flux towers in the sky: global ecology from space. New Phytol, 224: 570-584. doi:10.1111/nph.15934

Schwalm CR, Schaefer K, Fisher JB, et al. 2019. Divergence in land surface modeling: linking spread to structure. *Environ. Res. Commun.* **1:** 111004 https://doi.org/10.1088/2515-7620/ab4a8a

Seidel SJ, Palosuo T, Thorburn P, and Wallach D. 2018. Towards improved calibration of crop models – Where are we now and where should we go? *European Journal of Agronomy*. **94:** 25-35, https://doi.org/10.1016/j.eja.2018.01.006

Seidl R. 2017. To model or not to model, that is no longer the question for ecologists. *Ecosystems*. **20:** 222. https://doi.org/10.1007/s10021-016-0068-x

Shiklomanov AN, Bradley BA, Dahlin KM, et al. 2019a. Enhancing global change experiments through integration of remote sensing techniques. *Front Ecol Environ*. **17:** 215– 224, doi:10.1002/fee.2031

Shiklomanov AN, Cowdery EM, Bahn M, et al. 2019b. Does the leaf economic spectrum hold within plant functional types? A Bayesian multivariate trait meta-analysis. *Ecological applications*. In press. https://doi.org/10.1002/eap.2064

Smith P, Soussana J-F, Angers D, et al. 2019. How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Glob Change Biol*. **00:** 1– 23. https://doi.org/10.1111/gcb.14815

Stucky BJ, Guralnick R, Deck J, Denny EG, Bolmgren K, and Walls R. 2018. The Plant Phenology Ontology: A New Informatics Resource for Large-Scale Integration of Plant Phenology Data. *Frontiers in Plant Science*. **9:** 517. https://doi.org/10.3389/fpls.2018.00517

Todd-Brown KEO, Randerson JT, Post WM, Hoffman FM, et al. 2013. Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences*. **10:** 1717–1736, https://doi.org/10.5194/bg-10-1717-2013.

Tuomi M, Vanhala P, Karhu K, Fritze H, and Liski J. 2008. Heterotrophic soil respiration - Comparison of different models describing its temperature dependence. *Ecological Modelling*. **211:** 182-190.

Viskari T, Hardiman B, Desai AR, and Dietze MC. 2015. Model-data assimilation of multiple phenological observations to constrain and predict leaf area index. *Ecological Applications.* **25:** 546-558. doi:10.1890/14-0497.1

Walker AP, Ye M, Lu D, De Kauwe MG, Gu L, Medlyn BE, Rogers A, and Serbin SP. 2018. The multi-assumption architecture and testbed (MAAT v1.0): R code for generating ensembles with dynamic model structure and analysis of epistemic uncertainty from multiple sources, *Geosci. Model Dev.* **11:** 3159–3185, https://doi.org/10.5194/gmd-11-3159-2018

Walls RL, Deck J, Guralnick R, et al. 2014. Semantics in support of biodiversity knowledge discovery: an introduction to the Biological Collections Ontology and related ontologies. *PLoS One* **9:** e89606. doi: 10.1371/journal.pone.0089606

Warszawski L, Frieler K, Huber V, Piontek F, Serdeczny O, and Schewe J. 2014. The Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP): Project framework. *Proceedings of the National Academy of Sciences*. **111:** 3228-3232. doi: 10.1073/pnas.1312330110

White, EP, Yenni, GM, Taylor, SD, et al. 2019. Developing an automated iterative near-term forecasting system for an ecological study. *Methods Ecol Evol*. **10**: 332– 344. https://doi.org/10.1111/2041-210X.13104

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

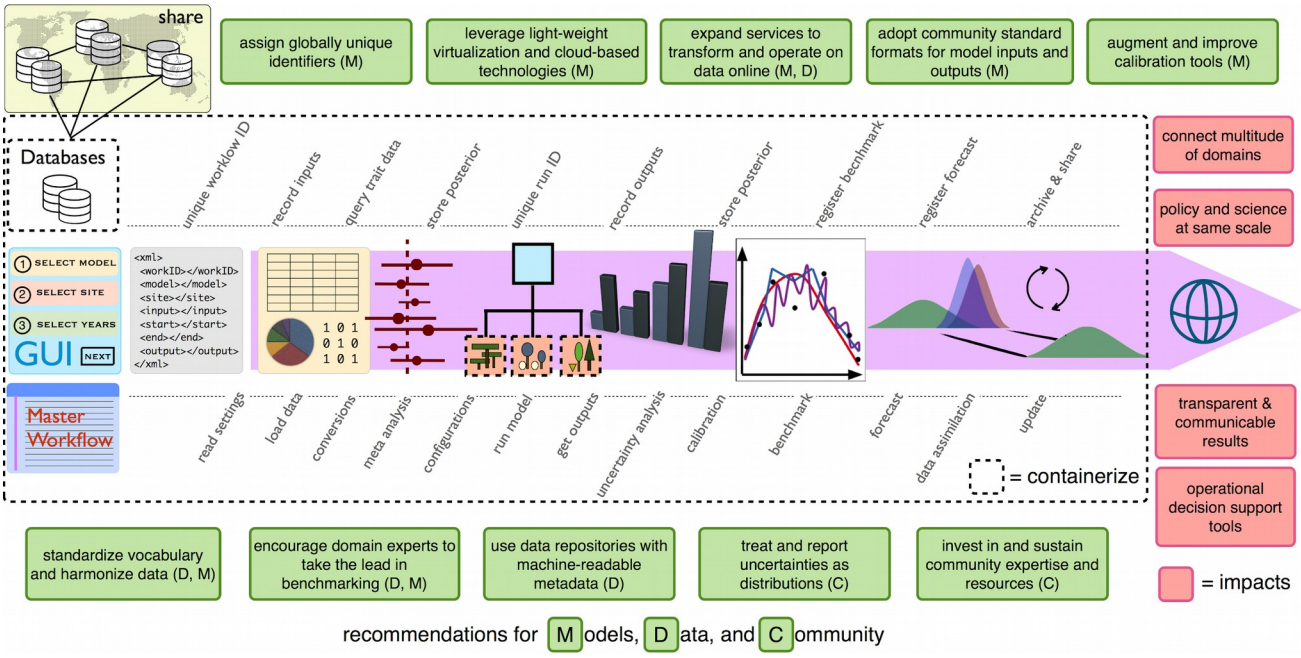

Figure 1. Schematic of a community cyberinfrastructure example and summary of recommendations. Users start with a high-level Graphical User Interface (GUI) to provide their setup for a modeling activity. These selections are translated into a human and machine readable markup language, and read in by the master workflow which then executes a sequence of tasks. At this stage, a unique identifier is assigned to the workflow to be executed. This ID, which points to the full workflow output and access to the metadata required to repeat it, can be shared among collaborators and published in papers. Next, the selections of the user are queried with the database, and actions are decided depending on whether requested items are already processed and ready to use or need to be retrieved and processed. Then, each module performs a well-defined task in the specified order. Crucial information for provenance of the whole workflow is recorded in the database during associated steps. Key outputs from analyses, such as meta analysis and calibration posteriors, are stored in a way that enables their exchange and re-use between different workflows. An important feature of this cyberinfrastructure is that both its parts and itself as a whole are virtualized (containerized) to add an additional layer of abstraction and automation, and to ensure interoperability.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

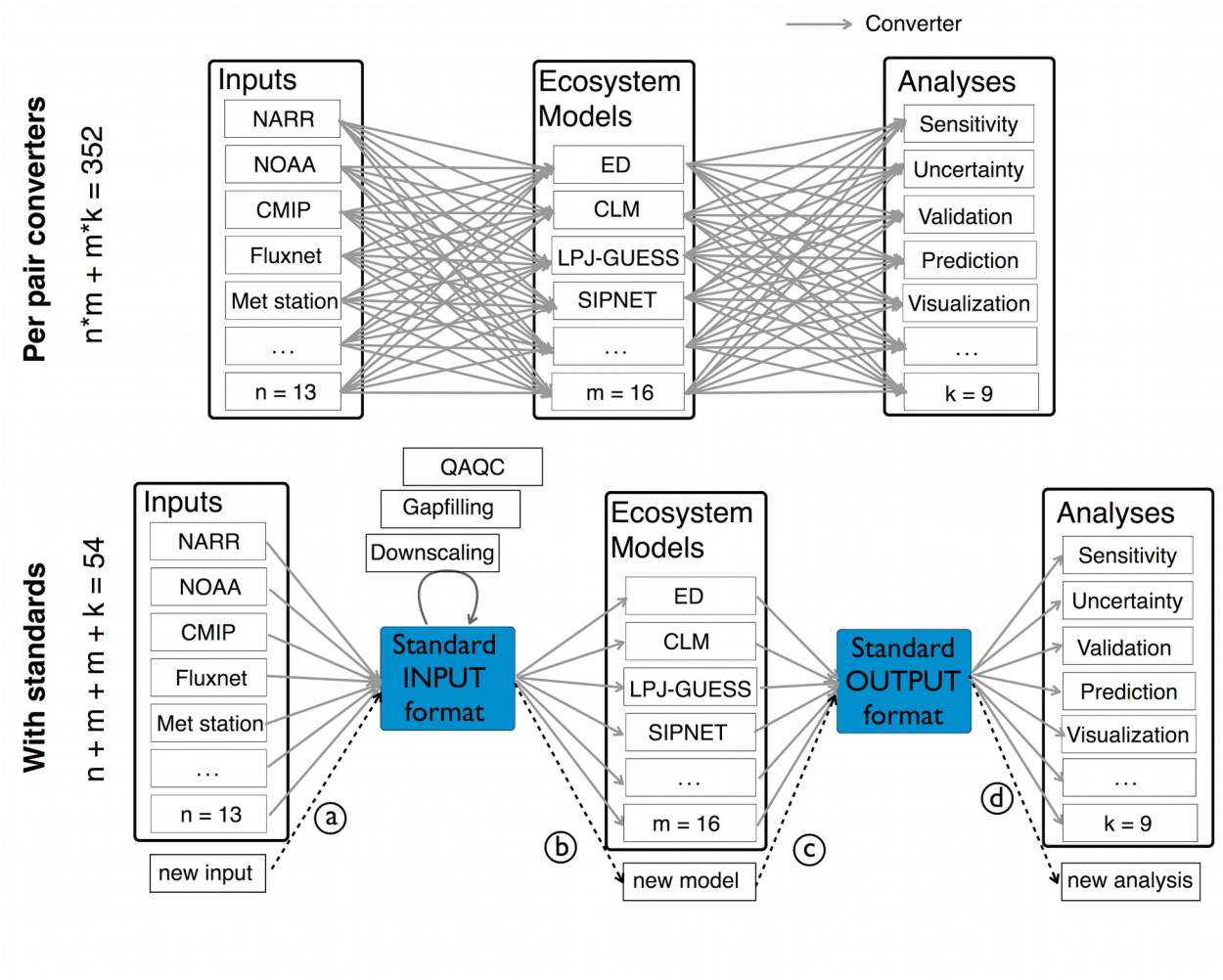

Figure 2. Reduction in redundant work when adopting common formats. There are "*n*" data types that must be linked to "*m*" simulation models and "*k*" post simulation analyses. For example, Predictive Ecosystem Analyzer (PEcAn) toolbox has 13 meteorological drivers, 16 models, 9 analyses coupled to its workflow. In the top panel, the conventional approach where modeling teams work independently requires implementing *n*m* different input and *m*k* different output conversions. As data, models, and analyses are added, effort scales quadratically. On the other hand, bottom panel shows that by working as a community, and adopting common formats and shared analytical tools, the number of converters necessary to link models, data, and analyses scales linearly. PEcAn follows the latter approach, adopts and extends the MsTMIP and CF standards as the common input and output formats. When a new input source or a new analysis is added to the system, it can immediately get access to *m* models by writing only one converter, (a) and (d) respectively. Likewise, when a new model is added, it can get access to *n* inputs and *k* analyses by writing one converter for each, (b) and (c) respectively. Furthermore, not only is there a major economy of scale in terms of marginal costs (1 for each data set or analytical tool added, 2 per model), but these tools can be made more reliable and sophisticated as less code will be written and tested by more people.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 January 2020 doi:10.20944/preprints202001.0176.v1

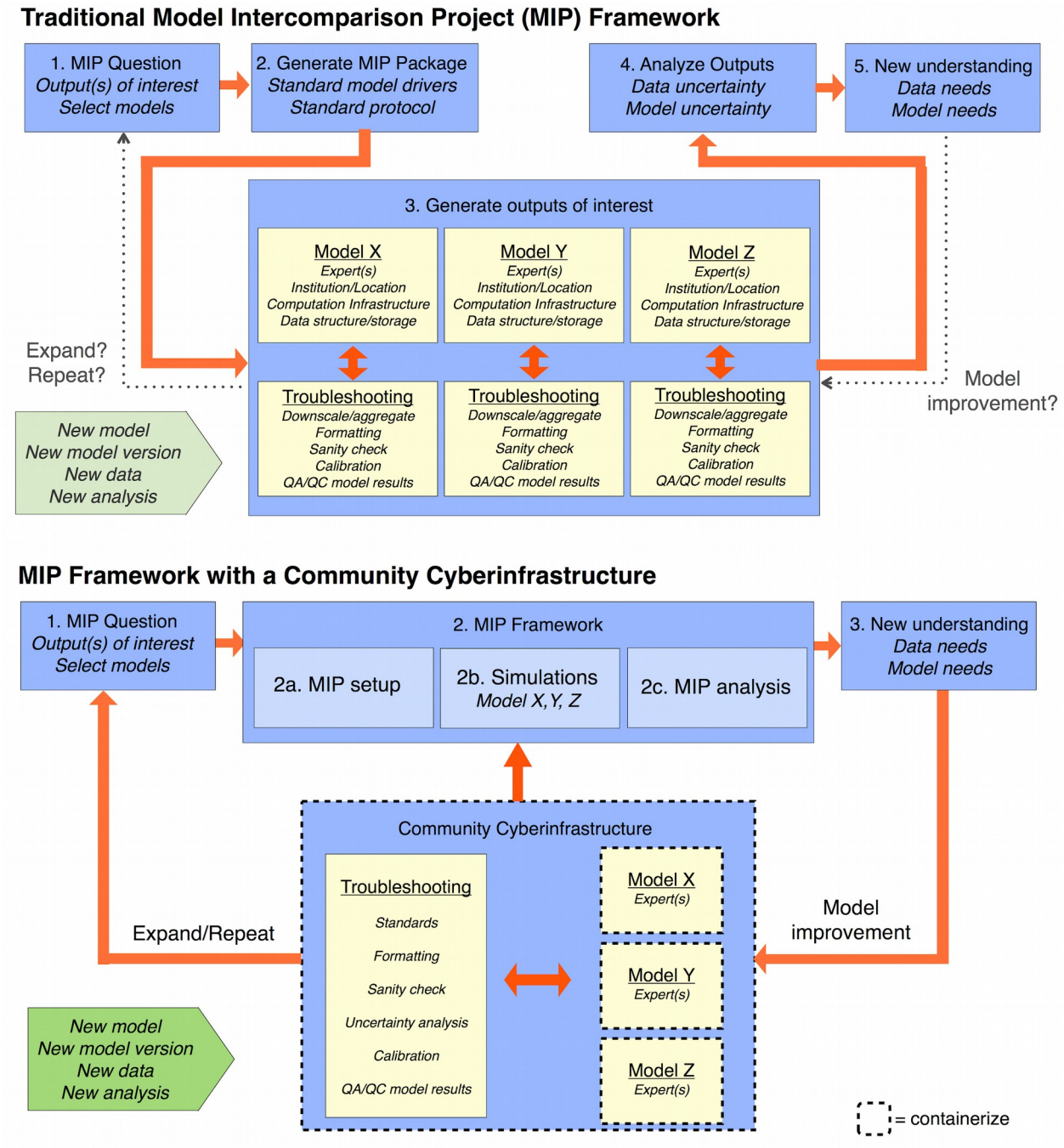

Figure 3. Traditional multi-model intercomparison project (MIP) workflow versus Community Cyberinfrastructure. Historically, each model and associated experts/infrastructure individually engage with MIPs. While stimulating model improvement is intended, it is not inherently nor readily available in traditional MIPs. In a Community Cyberinfrastructure, by contrast, both standardization of inputs and outputs and troubleshooting are included in the process of embedding each individual model in the system. MIP analyses are a use case, rather than the single purpose of the workflow, leverage ongoing Community Cyberinfrastructure development for a more streamlined and easily replicated/modified process. MIP conclusions relevant for model or cyberinfrastructure development can be fed directly back into this framework.

Acknowledgements

The PEcAn project is supported by the NSF (ABI no. 1062547, ABI no. 1458021, DIBBS no.1261582), NASA Terrestrial Ecosystems, the Energy Biosciences Institute, and an Amazon AWS education grant. We would also like to thank Boston University for providing the venue for the workshop that inspired this article. IF and TV acknowledge funding from the Strategic Research Council at the Academy of Finland (decision 327214), the Academy of Finland (decision 297350) and Business Finland (decision 6905/31/2018) to the Finnish Meteorological Institute. TQ is funded by the UK NERC National Centre for Earth Observation. JBF contributed to this work from the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. California Institute of Technology. Government sponsorship acknowledged. JBF was supported in part by NASA programs: CARBON and CMS. Copyright 2020. All rights reserved. SPS was partially supported by NASA CMS (grant #80NSSC17K0711), and through the DOE Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation Science Focus Area (RUBISCO SFA), which is sponsored by the Earth & Environmental Systems Modeling (EESM) Program in the Climate and Environmental Sciences Division (CESD), and the Next-Generation Ecosystem Experiments (NGEE-Arctic and NGEE-Tropics) supported by the Office of Biological and Environmental Research in the Department of Energy, Office of Science, as well as through the United States Department of Energy contract No. DE-SC0012704 to Brookhaven National Laboratory. MDK acknowledges funding from the Australian Research Council (ARC) Centre of Excellence for Climate Extremes (CE170100023), the ARC Discovery Grant (DP190101823) and support from the NSW Research Attraction and Acceleration Program. FMH was partially supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725. Additional support was provided by the Data Program, by the Reducing Uncertainties in Biogeochemical Interactions through Synthesis and Computation Science Focus Area (RUBISCO SFA) in the Earth & Environmental Systems Modeling (EESM) Program, and by the Next-Generation Ecosystem Experiments (NGEE-Arctic and NGEE-Tropics) Projects in the Terrestrial Ecosystem Science (TES) Program. The Data, EESM, and TES Programs are part of the Climate and Environmental Sciences Division (CESD) of the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science.

Author contributions

IF and AKG lead the writing with extensive feedback from MCD and with contributions from all authors. All authors have read and approved the manuscript.