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

Istem Fer[1,*], Anthony K. Gardella[2,3], Alexey N. Shiklomanov[4], Eleanor E. Campbell[5], Elizabeth M. Cowdery[2], Martin G. De Kauwe[6,7,8], Ankur Desai[9], Matthew J. Duveneck[10], Joshua B. Fisher[11], Katherine D. Haynes[12], Forrest M. Hoffman[13,14], Miriam R. Johnston[15], Rob Kooper[16], David S. LeBauer[17], Joshua Mantooth[18], William Parton[19], Benjamin Poulter[4], Tristan Quaife[20], Ann Raiho[21], Kevin Schaefer[22], Shawn P. Serbin[23], James Simkins[24], Kevin R. Wilcox[25], Toni Viskari[1], 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]Biospheric Sciences Laboratory (618), NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA [5]Earth Systems Research Center, University of New Hampshire, Durham, NH 03824, 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]Evolution & Ecology Research Centre, University of New South Wales, Sydney, NSW 2052, Australia [9]Department of Atmospheric and Oceanic Sciences, University of Wisconsin-Madison, 1225 W Dayton St, Madison, WI 53706, USA [10]Harvard Forest, Harvard University. 324 North Main Street, Petersham, MA 01366, USA [11]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA, 91109, USA [12]Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523, USA [13]Computational Earth Sciences Group and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6301, USA [14]Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, TN 37996, USA [15]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA [16]NCSA (National Center for Supercomputing Applications), University of Illinois at Urbana Champaign, Urbana, IL, 61801-2311 USA [17]College of Agriculture and Life Sciences, University of Arizona, Tucson, AZ 85721, USA [18]The Fulton School at St. Albans, St. Albans, MO 63073, USA [19]Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA [20]UK National Centre for Earth Observation and Department of Meteorology, University of Reading, Reading, RG6 6BB, UK [21]Fish, Wildlife, and Conservation Biology Department, Colorado State University, Fort Collins, CO 80523, USA [22]National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309, USA [23]Brookhaven National Laboratory, Environmental and Climate Sciences Department, Upton, NY, 11973, USA, [24]University of Delaware, Newark, USA, [25]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 to informing models with observations have reduced our capacity to fully exploit the growing volume and variety of available data. Here, we take a critical look at the information infrastructure that connects ecosystem modeling and measurement efforts, and propose a roadmap to community cyberinfrastructure development that can reduce 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, including both modelers and empiricists. This roadmap focuses on five key opportunities for community tools: the underlying foundations of 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.

Introduction

Kindled by rapid environmental change, the scientific community is deeply invested in understanding and predicting nature's dynamics (Dietze et al. 2018; Rineau et al., 2019; Hanson and Walker, 2020). Thankfully, recent decades have seen an explosion of environmental data globally that is being delivered to us faster than ever before (LaDeau et al. 2017; Farley et al., 2018; Reichstein et al. 2019; Schimel et al., 2019). Process-based ecosystem models play a critical role in translating data into mechanistic understanding, as they provide us with the ability to synthesize and reformulate knowledge across organizational, spatial, and temporal scales, and to generate testable predictions from alternative hypotheses (Fisher et al., 2014; Medlyn et al., 2015; Hanson and Walker, 2020). Despite having more data than ever before, we have not seen comparable progress in our capacity to forecast natural systems with process-based models (Lovenduski and Bonan, 2017; Bonan and Doney, 2018; Dietze et al., 2018). For example, model projections out to the year 2100 do not agree on whether terrestrial ecosystems will be a carbon sink or source in response to climate change, and these discrepancies have not changed despite years of apparent model improvement (Friedlingstein et al., 2006, 2014; Arora et al., 2020). Perhaps this is not unexpected: adding model complexity without being informed by data does not equate to improved predictions, new processes (e.g. nutrients) may increase realism but may undo previous calibrated performance unless calibration is renewed easily. Overall, it is not a simple task to evaluate multiple model ensembles, making conclusions about forecast capacity complicated (Lovenduski and Bonan, 2017; Herger et al., 2019). A new strategy is needed to approach challenges in advancing our ecological understanding, reducing uncertainties and integrating the disparate science communities of global change biology (Bonan and Doney, 2018; Dietze et al., 2018). The goal of this paper is to better characterize the bottlenecks that have obstructed the rates at which new information has been integrated into ecosystem models, and to lay out a roadmap to overcome these bottlenecks. While many of the examples here are focused on terrestrial ecosystem models, the principles highlighted are general across different systems and processes.

A more predictive global change science needs to be based on ecosystem models that capture important processes rather than merely reproducing patterns (Medlyn et al., 2015; Lovenduski and Bonan, 2017; Bonan and Doney, 2018). Modeling efforts should be geared towards generating hypotheses that are testable against data (Hanson and Walker, 2020). Most current modeling activities, however, are more likely to be informed by high-volume high-level observational data (e.g., landscape level biogeochemical fluxes) than experimental manipulations (Wieder et al., 2019) or studies focused on low-level process details (e.g., interactions between non-structural carbohydrate reserves, drought, and mortality; Keenan et al., 2013). This is in direct contrast with the incredibly diverse range of data generated by ecology as a discipline (Hanson and Walker, 2020). Until modeling tools become more accessible, new communities of model users who can expand model-based interpretation and hypothesis testing beyond its limited scope will be curbed by informatics bottlenecks that impede wider representation.

More importantly, current approaches in confronting models with data frequently fail to actively engage the non-modeler community, who often possess a more detailed understanding of processes and study systems (Jeltsch et al. 2013; Seidl, 2017). This bottleneck not only impacts the pace and the quantity, but also the quality of modeling efforts. The division between empirical and modeling research is further exacerbated by the current "uniqueness of models"; that is, each model comes with an idiosyncratic learning curve due to the lack of standards around model interfaces and operation. To restore the balance, we need to concurrently increase modeling literacy and lower the technical barrier for modeling activities (Seidl, 2017). This barrier, overall, hinders efforts to replicate findings, extend analyses to other models and locations, and routinely confront model-based hypotheses with data (Gil et al., 2016).

We argue that a major step towards reducing these model-data bottlenecks lies in the development and support of community-wide cyberinfrastructure: a computational environment where we can effortlessly operate on data, simulate natural phenomena, perform model evaluation, and interpret results (Dietze et al., 2013; Gil et al., 2016; Eyring et al., 2019; also see Appendix A for a glossary of terms). While the general idea is not new, their application has been limited in ecology. However, there are several converging initiatives that make it timely to reinvigorate efforts (see Appendix C and D for example initiatives and their overview, and the box for "How to support and sustain community cyberinfrastructure?").

In the following sections, we present a roadmap to the key features of a community cyberinfrastructure, and discuss specific challenges and solutions for model-data activities. These activities include but are not limited to: i) obtaining and processing 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). We provide specific recommendations for the measurement community, the modeler and developer community, and the broader community throughout each section (Fig 1 and Appendix B).

FAIR Cyberinfrastructure essentials

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 (Dietze et al., 2013). Tackling this at the individual model level leads to redundant efforts across-models and inhibits economies of scale that could be gained by sharing informatics tools across communities (for examples of shared ecological informatics infrastructure please see Appendix C). Besides, the larger community of users associated with common infrastructure will foster innovation and create an incentive for developers to make better, more sophisticated algorithms that have gone through more extensive testing (Gil et al., 2016). The revolutionary success of the open source and free programming language R (R Core team, 2020) aptly exemplifies the importance of community involvement in developing and sharing standard tools for a massive reduction in redundant efforts, as well as having access to a much larger community support (Boettiger et al., 2015; Lai et al., 2019).

Here we briefly highlight the FAIR (*findable, accessible, interoperable, and reusable)* cyberinfrastructure essentials to facilitate a catalogue of model-data activities (for more details on FAIR principles for research software and data, please see Gil et al., 2016; Culina et al., 2018; Hasselbring et al., 2020 and the references therein):

- *Findability* refers to the ease with which permanent records of the key metadata about each model-data activity and computational output can be found (Hasselbring et al, 2020). Recording the full, transparent history of an analysis to enable findability is known as provenance. For large model-data workflows executing multiple models or experiments, we recommend **[R1; R for recommendation]** model developers utilize open community provenance databases, which assign unique and persistent identifiers to each model-data activity (LeBauer et al., 2013; Gil et al., 2016). Such identifiers could be used in publications, pointing readers to the full computational output and the metadata required to repeat a model run (Fer et al., 2018). **[R2]** The workflow and provenance system themselves should also be version controlled (e.g. using GitHub) to ensure a fully reproducible record (Piccolo and Frampton, 2016). **[R3]** Then, any changes to their code need to be automatically tested to ensure expected behaviour by tools for continuous integration (e.g Travis CI, travis-ci.com; Github Actions, github.com/features/actions).

- *Accessibility* in modeling goes beyond obtaining the model code. A broader technical barrier exists in terms of the abilities required to effectively deploy simulation models and perform complex analyses. **[R4]** A well-defined automated workflow that coordinates individual tasks (Fig 1) should be set up by the developers to (1) reduce barriers to entry, (2) ensure replication is possible, and (3) reduce costs of manual operation. The process of focusing on the design of this workflow, which is also known as abstraction, requires standardizing and generalizing the important tasks involved, and devising how they are related to one another. Leveraging systemized approaches (e.g. tidyverse in R, or pandas in Python) throughout the workflow design promotes consistency, creates predictable expectations and fosters knowledge transfer across projects. Abstraction further facilitates presenting the user with a **[R5]** more intuitive and accessible interface that handles everything from running ecosystem models in place to submitting complex analyses to remote high-performance computing resources under the hood.

- *Interoperability* is critical to building cyberinfrastructure that works seamlessly across many models, but this requires predictable file formats for model inputs, outputs, and data constraints used by the community. While reducing the proliferation of both data and model formats would alleviate this in the long term, in the short-term **[R6]** using standard data pipelines can remedy the redundant efforts put into building custom tools. For example, consider the common problem of managing the data streams in and out of the models with two cases where i) every developer team works independently (Fig 2, top panel), ii) a common pipeline with internal standards is used (Fig 2, bottom panel). Not only is the latter approach much more scalable, but these tools can be made more reliable and sophisticated as less code will be written and tested by more people. **[R7]** We recommend the ecological community leverage existing standard formats as the internal standards, such as the Climate and Forecast (CF) convention (Eaton et al., 2017), and the use of ontologies to provide harmonized vocabularies and semantic frameworks (e.g. Stucky et al., 2018).

- *Reusability* of community models and tools builds on interoperability but also requires **[R8]** individual tasks involved be isolated and modularized in the workflow (Fig 1). Modularity would allow (1) internal modifications to their implementation without altering the overall behavior of the system; (2) independent reuse of tools outside of specific systems; and (3) users to swap in/out alternative algorithms/tools and customize their workflow. Community cyberinfrastructure should further be available to users without having to deal with obscure system requirements and dependencies. Similar to what programming language R has achieved, more standardized installation procedures and fewer configuration steps significantly reduce user time for setup and increase adoption, reusability and reproducibility. Fortunately, modern virtualization technologies offer a number of tools that allow users to run packaged software, called containers, complete with all its dependencies (Piccolo and Frampton, 2016). **[R9]** We recommend developer communities adopt recent light-weight containerization systems (such as e.g., Docker - www.docker.com; Singularity - singularity.lbl.gov) that are easy to install, set up, upgrade, and scale up with new locations to run the models. Containerization allows existing infrastructures to be run reliably across a variety of computing resources, including cloud-based virtual services (Farley et al., 2018; Hasselbring et al., 2020).

Data ingest opportunities

Data play a critical role in modeling activities; however, due to their sheer volume and diversity, they can be difficult to locate and obtain as sifting through deluge of data manually is impractical (Waide et al., 2017; Reichstein et al., 2019). **[R10]** To make data FAIR, we recommend data producers use consistent naming structures (e.g. Assistance for Land-surface Modelling activities [ALMA] convention, also please see Appendix A for more details) and open file formats (e.g. CSV, netCDF) (Hart et al., 2016). **[R11]** Next, data should be stored in data repositories where datasets are versioned, data citations are provided, and that support **[R12]** standard, searchable metadata, and machine-readable Application Programming Interfaces (APIs) (e.g. the Oak Ridge National Laboratory Distributed Active Archive Center, Cook et al. 2016; Environmental Data Initiative, Gries et al., 2019; Open Science Framework, Sullivan et al., 2019). When those repositories are part of jointly searchable networks (e.g. DataONE - www.dataone.org), it could further allow developers to leverage one set of tools for many sources.

Admittedly, data providers may have to invest significant time and resources to follow these recommendations. These costs include; preparing descriptive metadata to prevent misuse, choosing the right repository with appropriate licensing and without isolating data from relevant disciplines, and finding means (funding and expertise) to manage data especially for small projects (Gil et al., 2016; Waide et al., 2017; Culina et al., 2018). Furthermore, other valid concerns such as data leakage and insufficient recognition are frequently raised (Bond-Lamberty et al., 2016). While these issues are not specific to the roadmap discussion here, community cyberinfrastructure tools can alleviate them to a certain extent. For example, investments in optimizing standardized protocols, terminologies and file formats for community tools during data collection and processing .will help with metadata preparation and repository selection. By getting involved with community cyberinfrastructure, small projects can gain access to larger community expertise and support. Cyberinfrastructure data ingest pipelines can automatically query licenses as chosen by the data provider (Culina et al., 2018) and streamline citations to credit researchers seamlessly. Community tools (such as Brown Dog, browndog.ncsa.illinois.edu) can access and index data collections, in particular small uncurated and/or unstructured data collections, thereby preventing data loss, increasing discovery and further securing recognition.

On the big data side, approaches for scientifically and computationally interacting with high volume, high velocity data become increasingly available (Reichstein et al., 2019). While it is important to generalize these cutting-edge tools and share with the community, modeling activities frequently involve a subset of data (e.g., a specific region or period) for which time to transfer data often exceeds the time to process it. Thus, we endorse the recent paradigm of **[R13]** cloud computing and online services (e.g. Google Earth Engine) that allow users to select, subset, transform, or perform other operations on the data without having to download and expand (see Gomes et al., 2020 for more examples). Within this set up, community cyberinfrastructure also provides a medium where a diverse array of data delivered by Internet of Things (IoT) techniques can be integrated into models in a sensible manner (Fang et al., 2014). As developers combine cloud-based cyberinfrastructure tools with cutting-edge data platforms, this would free the users from their local constraints altogether. Empowering more groups to interact with large datasets brings its own push towards progress in terms of scientific proficiency and diversity (Nagaraj et al., 2020).

Way forward in calibration

After data ingest, another persistent challenge in process-based ecosystem modeling is calibration: the process of using data to constrain model parameters (Dietze et al. 2013; van Oijen, 2017, Seidel et al., 2018). Some model parameters may be directly informed by ecological trait data (e.g., turnover rates). In this case, meta-analysis tools can pull data together from open-access, machine-readable, curated databases (LeBauer et al. 2013, 2018; Shiklomanov et al. 2020). A non-negligible portion of model parameters, however, are often not directly measurable, therefore, there is a need to estimate parameters indirectly using inverse methods that infer what parameter combinations produce model predictions compatible with observations (Hartig et al., 2012). **[R14]** When doing this, we recommend the community take the Bayesian approach to transfer the information from data to probability distributions about models and parameters (Hartig et al., 2012; LeBauer et al., 2013). Bayesian approach allows combining information from multiple sources and scales, iteratively updating our understanding as new data become available, propagating uncertainty into model predictions to inform decision making, and it is becoming more effective in dealing with complex systems with the increase of computing power and numerical methods (van Oijen, 2017).

Most off-the-shelf Bayesian tools (e.g. JAGS - mcmc-jags.sourceforge.net; STAN - mc-stan.org), however, are not designed 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 specific syntax of these software. In addition, **[R15]** these tools need to support re-reading their own outputs (posteriors) as new inputs (priors), which is critical for iterative updating of the analyses. Due to lack of available tools, models are frequently used uncalibrated (or hand-tuned) (Seidel et al., 2018). 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, hindering overall progress in model development (van Oijen, 2017). **[R16]** Using multiple data constraints can be critical to ensuring that a model is getting the right answer for the right reason (Medlyn et al., 2015). Even when a model is calibrated for one setting (*e.g.*, site or period), it does not guarantee reliable performance at another setting because there is variability and heterogeneity in natural systems. More flexible techniques, such as hierarchical Bayesian calibration, can formally quantify the scales of unexplained system variability and inform directions for model development (van Oijen, 2017), but there are even fewer available tools for their standard implementation with external models.

Within a community cyberinfrastructure, the challenge of developing advanced calibration tools only needs to be faced by statistics experts. Software alternatives for calibrating 'black-box' models are becoming increasingly available (Fer et al., 2018; Hartig et al., 2019; Huang et al., 2019). **[R17]** Community cyberinfrastructure will be most successful if hierarchical calibration tools are able to account for all kinds of ecological variability and heterogeneity (Farley et al., 2018), and if coupling to a calibration workflow is part of model development. When calibration tools are implemented in community cyberinfrastructure, they can seamlessly link multiple data constraints with multiple models. As such workflows are tracked by provenance systems, **[R18]** results from one analysis (e.g. posteriors) can readily be used by a subsequent analysis elsewhere, accelerating our ability to confront models with data. 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 **[R19]** developing novel, more advanced calibration techniques (e.g. with emulators, Fer et al., 2018; deep learning, Tao et al., 2020).

Model intercomparison and benchmarking

Comparing models to data is at the heart of hypothesis testing and model evaluation (Fisher et al., 2014; Best et al., 2015). While process-models are frequently compared to multiple datasets across their lifespan, it is remarkably rare to put an ecosystem model through all its past assessment exercises every time it is updated unless a workflow has been automated (Best et al., 2015; Collier et al., 2018). **[R20]** To verify progress, and assess the tradeoffs between model parsimony and complexity, key datasets need to be set as "benchmarks" to track and compare performance through time (Luo et al., 2012; Best et al., 2015). Benchmark data can also be used to compare across models as part of model intercomparison projects (MIPs). However, the lack of automated and shared workflows also makes traditional MIPs logistically challenging to coordinate and repeat (Fig 3, top panel). Modeling groups could face incompatibilities in their results due to differences in their model configurations (e.g. calibrated vs. uncalibrated). Furthermore, due to the cost of performing a MIP, model output requests and experimental designs are typically kept simple. For example, MIPs largely focus on single model realizations which can lead to biased or overprecise decisions about model performances.

Many of the utilities that are particularly valuable for MIPs and benchmarking are already included in embedding each individual model in the community cyberinfrastructure (Fig 3, bottom panel). The use of a cyberinfrastructure also opens up the possibility of more advanced MIP benchmark activities, such as running ensembles to propagate input uncertainty to model output uncertainty. Generating multi-model ensembles with uncertainties are also practical for studying model structural errors (Bonan and Doney, 2018) and for model averaging which could potentially reduce prediction errors (Dormann et al., 2018). **[R21]** We recommend the community move towards benchmarks that account for model and data uncertainty, and leverage this information when computing model performance scores (e.g. benchmarking that takes into account the uncertainty bounds in models and observations to calculate a score based on overlap probability).

Once a model is integrated into community cyberinfrastructure, it becomes trivial to add its alternative versions, benchmark against existing MIPs and seamlessly feedback to future model developments (Kelley et al., 2013; Collier et al., 2018; Wieder et al., 2019). For example, advancing model versions would benefit from being continually tested against the Free-Air CO_2 Experiments (FACE-MIP, De Kauwe et al., 2014; Hoffman et al., 2017) and the Arctic-Boreal Vulnerability Experiment (ABoVE, Fisher et al., 2018). Within or in addition to existing frameworks, interactive environments (e.g. Rstudio/Jupyter) would allow users to perform more extensive analyses with pre-loaded and aligned models and data. However, a number of challenges remain, including how to deal with data sets and metrics that are incomplete or inconsistent with each other (Hoffman et al., 2017; Collier et al., 2018). **[R22]** Thus, we further recommend model developers enable direct comparison to observations when possible. For example, instead of relying on modeled data products (e.g. leaf area index) whose uncertainties are harder to determine, models can be augmented to predict observations (e.g. reflected spectral radiance) as measured by the instruments. In other words, 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. Concomitantly, community cyberinfrastructure would facilitate **[R23]** interaction with a compilation of standard data sets that models need to be able to reproduce repeatedly (Anderson-Teixeira et al., 2018; Kraemer, et al. 2020; Reyer et al., 2020).

Who sets up benchmarks?

To address the bottleneck that only a small fraction of the data collected by ecologists (often with the aim of improving projections) ever makes its way into ecosystem models and scale up, data generators and disciplinary experts need also be equipped with tools for data-model comparison, not only the "modeler" minority (Seidl, 2017). Through community cyberinfrastructure, **[R24]** domain experts will more easily be able to compare multiple models to their data and set up persistent benchmarks. 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 while assessing the underlying processes and mechanisms represented in the models. As cyberinfrastructure automates tedious activities associated with a MIP, experts can focus on their analysis rather than the logistics, making modeling activities more relevant for their science.

Yet, even before the challenges of running a model or a MIP, it is nearly impossible for non-modelers to keep abreast of which models exist, their most updated version, and their respective strengths and weaknesses (Jeltsch et al. 2013; Schwalm et al., 2019). **[R25]** Therefore, 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 models. As a result, process representations that repeatedly perform below-average across multiple MIPs can be considered rejected hypotheses (Schwalm et al., 2019), which community cyberinfrastructure could track and in return inform the development of the next generation of models as advancing new hypotheses can regain focus. In time, by centralizing these comparisons into databases, community cyberinfrastructure allows new users to discover new models and to evaluate their updated process representations with minimal technical barriers while simultaneously allowing the modeling minority to focus on learning from their colleagues and improving models, rather than the status quo where the majority of their time is spent on mundane informatics issues.

Data assimilation and ecological forecasting

For ecology to respond to the pace of global change, and better inform environmental decisions, the nature of the relationship between ecological models and data must be reconsidered. While most ecological analyses tend to be non-specific and a posteriori (e.g. ANOVA models), and most ecological forecasts are long-term (e.g. 2100 projections), there is much to be learned from **[R26]** making near-term ecological forecasts that can be tested and updated as new observations become available (Fox et al., 2009; Dietze et al., 2018). Adopting an iterative forecasting approach will not only make ecology more relevant to the society, by providing information on fast, decision-relevant timescales, but will also transform basic ecological science and theory (Dietze et al. 2018), by accelerating the pace at which specific, quantitative, and falsifiable predictions are confronted with data.

Like calibration, the data assimilation methods that drive forecasting, through a formal fusion of data and modeled states (or both states and parameters), also require advanced statistical and computational expertise. Ecological models and data frequently violate the statistical assumptions embedded in assimilation algorithms developed in other disciplines (e.g. normality, homoscedasticity, independence), hence, **[R27]** many existing tools need to be reassessed and generalized by experts within community tools to appropriately meet the ecological model-data characteristics (Raiho et al., 2020). Making a forecast operational also requires **[R28]** a higher level of repeatability and efficient scheduling of cyclic workflows, where a large number of jobs are executed at regular intervals and each forecast cycle depends on previous ones (Oliver 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 pipelines provide a key economy of scale in data assimilation and forecasting and builds upon many of the features already discussed (Dietze et al., 2018): informatics tasks of gathering,processing and standardizing new data will maximize data use and diversity of contributions. Managing the execution of analytical workflows will refine analyses and make them applicable to new problems. **[R29]** By publicly archiving and reporting results community cyberinfrastructure enables comparisons of different forecasting approaches, future syntheses, and assessment of improvement over time. 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 (Oliver et al. 2019) and data assimilation (Fox et al., 2018; Raiho et al., 2020; Pinnington et al. 2020).

Box. How to support and sustain community cyberinfrastructure?

The ongoing maintenance and development of common cyberinfrastructure tools are essentially conditioned upon uptake and support by the community. This effort typically starts with building a bottom-up community (Boettiger et al., 2015) involving:

- Support widely adopted languages by the domain scientists (e.g. R and Python) so that;
    - experienced users can get off to a running start,
    - inexperienced users would be motivated to invest efforts with the co-benefit of learning a popular language,
    - larger communities of these languages can bring further support.
- Initiate strong ties with the demographic that can highly benefit from community solutions such as early career researchers.
- Establish codes of conduct for inclusion and diversity, and encourage participation regardless of experience level.
- Always adhere to open software best practices to build a reputation that can in return attract human resources and funding.

Luckily, these efforts do not need to start from scratch: the community can adopt and build upon existing systems (Appendix C). While we acknowledge that getting involved with community development requires upfront investment of time and resources of individuals, the benefits from participation are significant overall:

- Contributions to community tools perpetuate and increase their value, elevate recognition of their contributors (Lowndes et al., 2017; Dai et al. 2018).
- Community involvement provides larger support and career networks (McKiernan et al., 2016).
- In a research landscape that is ever diversifying, community cyberinfrastructure will be an active learning platform where ecologists gain advanced capability (Dietze et al., 2013).

As the community grows, successful strategies could be taken as an example, such as the WRF (The Weather Research and Forecasting Model) community (Powers et al, 2017):

- Financial and personnel burdens are spread out among the community, while the main support and steering responsibility could remain centralized.
- A help service that is responsible for user assistance is fundamental.
- Building committees in charge of coordination and direction is effective, e.g.:
    - Developers committee, to maintain code design, testing and upkeep
    - Release committee, to oversee and time major releases
    - Review committee, for scientific evaluation of major module/package contributions

Open software and data management plans are increasingly becoming an important requirement by funding agencies (Powers and Hampton, 2019) for which use of community cyberinfrastructure could be fittingly proposed. Thus, we suggest such proposals to include a budget item or person hours for the support of community tools when possible. While projects without funding should also be welcome, short-term funding opportunities for open research (McKiernan et al., 2016; Powers and Hampton, 2019) will help bottom-up community building. However, viability over the long-term requires sustainable funding structures and top-down support from funding agencies, networks, and the private sector. There are currently several appropriate venues for cyberinfrastructure projects (e.g. NSF Cyberinfrastructure for Sustained Scientific Innovation), but as communities make their cyberinfrastructure needs better known (e.g. through communication with funding agencies and uptake), we expect such opportunities to increase in number and variety. Ultimately, **[R30]** it is important that community and funding agencies support the sustainability of these tools as critical components of the collective scientific infrastructure in a similar way they do with the physical infrastructure (field stations, sensor networks, satellites) and data repositories.

Conclusions

Scientists, managers, and policy makers increasingly rely on models to understand the impact of decisions on ecological processes (Arneth et al. 2014; Bonan and Doney, 2018; Smith et al., 2019). As the barriers to entry for using the latest models and data are lowered, decisions will be made with better information, and scientific problems will be solved more quickly. Community cyberinfrastructure is the engine to bring time frames associated with model-data integration in line with the pressing needs of managers, policymakers, and society more broadly. We summarize our major recommendations for promptly meeting the dispersed and variable model-data synthesis needs of the ecological community as follows.

(1) Integrated community principles and practices

Modeling needs to be open, verifiable and credible. Three key concepts in modeling cyberinfrastructure — abstraction, automation, and provenance — open up the possibility for realistic replication, community-wide transparency, and model-based ecological analysis. Adopting common cyberinfrastructure tools that are accessible, reproducible, interoperable, scalable, and community-driven, will play a critical role in reshaping how ecologists interact with models.

(2) Reusable data and software

Data processing remains a bottleneck to model improvement. To foster effective discovery and reuse of both data and software, we recommend human- and machine-friendly community-scale approaches. Developing reusable tools based on community standards and involving the measurement community more deeply in data-model integration, are both essential for scaling up modeling efforts.

(3) More advanced calibration techniques

Testing hypotheses should be done with properly calibrated models. Inconsistencies in model comparison due to different calibration procedures will be reduced by employing shared Bayesian calibration tools that are set up to work with process-based models. Hierarchical Bayesian calibration solutions and novel algorithms, developed and generalized under community cyberinfrastructure, will help us better capture the inherent variability and heterogeneity in ecological systems.

(4) Persistent benchmarks

Model benchmarking and intercomparison are dynamic activities that need to continually inform model improvement. We recommend a more streamlined, easily repeated and modified process for benchmarking a suite of models with varying levels of process complexity and scale. Community cyberinfrastructure will allow domain experts to determine and more directly influence the most salient datasets that models need to replicate to demonstrate that they are capturing processes correctly, and then take the lead in setting up and performing these benchmarks.

(5) Near-term ecological forecasts

Automated data assimilation and forecasting pipelines are a necessity for ecology to support decision making in an increasingly non-equilibrium world that has moved outside of historical norms. Building these forecasting systems requires complex automated systems, and community cyberinfrastructure is well-positioned for putting the parts of operational forecasts together.

Process-based models, though imperfect, are our window into the future functioning of ecosystems under global change. The next generation of ecological models will need to ingest increasingly diverse and expansive 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. 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.

Data availability

Data sharing not applicable to this article as no datasets were generated or analysed during the particular study.

Code availability

Code availability not applicable to this article. However, we note for the interested reader that all example community tools mentioned in Appendix C are open source and available on online code repositories.

References

Anderson-Teixeira K.J., Wang M.M.H., McGarvey J.C., Herrmann V., Tepley A.J., Bond-Lamberty B., LeBauer D.S. (2018). ForC: a global database of forest carbon stocks and fluxes. Ecology, 99:1507-1507. doi:10.1002/ecy.2229

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

Arora V.K., Katavouta A., Williams R.G., Jones C.D., Brovkin V., et al. (2020). Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models, Biogeosciences, 17, 4173–4222, doi:10.5194/bg-17-4173-2020.

Best M.J., Abramowitz G., Johnson H.R., Pitman A.J., Balsamo G., et al. (2015). The Plumbing of Land Surface Models: Benchmarking Model Performance. Journal of Hydrometeorology, 16:1425–1442. doi:10.1175/JHM-D-14-0158.1

Boettiger C., Chamberlain S., Hart E., Ram K. (2015). Building Software, Building Community: Lessons from the rOpenSci Project. *Journal of Open Research Software*, *3*(1), e8. doi: http://doi.org/10.5334/jors.bu

Bonan G.B., Doney S.C. (2018). Climate, ecosystems, and planetary futures: The challenge to predict life in Earth system models. Science, 359. doi:10.1126/science.aam8328

Bond-Lamberty B., Smith A.P., Bailey V. (2016). Running an open experiment: transparency and reproducibility in soil and ecosystem science. Environmental Research Letters. 11:084004. doi: 10.1088/1748-9326/11/8/084004

Collier N., Hoffman F.M., Lawrence D.M., Keppel-Aleks G., Koven C.D. et al. (2018). The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. Journal of Advances in Modeling Earth Systems, 10: 2731– 2754. doi:10.1029/2018MS001354

Cook R.B., Vannan S.K.S., McMurry B.F., Wright D.M., Wei Y., Boyer A.G., Kidder J.H. (2016). Implementation of data citations and persistent identifiers at the ORNL DAAC. Ecological Informatics, 33:10-16. doi:10.1016/j.ecoinf.2016.03.003

Culina A., Baglioni M., Crowther T.W. et al. (2018). Navigating the unfolding open data landscape in ecology and evolution. Nature Ecology and Evolution, 2, 420–426. doi:10.1038/s41559-017-0458-2

Dai S.-Q., Li H., Xiong J., Ma J., Guo H.-Q., Xiao X., Zhao B. (2018). Assessing the extent and impact of online data sharing in eddy covariance flux research. Journal of Geophysical Research: Biogeosciences, 123, 129–137. doi:10.1002/2017JG004277

De Kauwe M.G., Medlyn B.E., 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 Phytologist. 203:883-899. doi:10.1111/nph.12847

Dietze M.C., LeBauer D., Kooper R. (2013). On improving the communication between models and data. Plant, Cell & Environment, **36**: 1575-1585. doi:10.1111/pce.12043

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

Dormann C.F., Calabrese J.M., Guillera-Arroita G. et al. (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecol Monogr. **88:** 485-504. doi:10.1002/ecm.1309

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

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

Fang S., Xu L.D., Zhu Y., et al. (2014). "An Integrated System for Regional Environmental Monitoring and Management Based on Internet of Things," in *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1596-1605, doi: 10.1109/TII.2014.2302638.

Farley S.S., Dawson A., Goring S.J., Williams J.W. (2018). Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*. **68:** 563–576. doi:10.1093/biosci/biy068

Fer I., Kelly R., Moorcroft P.R., Richardson A.D., Cowdery E.M., Dietze M.C. (2018). Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation, *Biogeosciences*. **15:** 5801–5830. doi:10.5194/bg-15-5801-2018

Fisher J.B., Huntzinger D.N., Schwalm C.R., Sitch S. (2014). Modeling the terrestrial biosphere. *Annual Review of Environment and Resources*. **39:** 91-123 doi:10.1146/annurev-environ-012913-093456

Fisher J.B., Hayes D.J., Schwalm C.R., et al. (2018). Missing pieces to modeling the Arctic Boreal puzzle. *Environ. Res. Lett.* **13:** 020202. doi:10.1088/1748-9326/aa9d9a

Fox A., Williams M., Richardson A.D., et al. (2009). The REFLEX Project: Comparing Different Algorithms and Implementations for the Inversion of a Terrestrial Ecosystem Model against Eddy Covariance Data. Agricultural and Forest Meteorology. **149:** 1597–1615. doi:10.1016/j.agrformet.2009.05.002.

Fox A., Hoar T.J., Anderson J.L., 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. doi: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. doi: 10.1175/JCLI3800.1

Friedlingstein P., Meinshausen M., Arora V.K., Jones C.D., Anav A., et al. (2014). Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *J. Clim.* **27**: 511–26. doi: 10.1175/JCLI-D-12-00579.1

Gil Y., David C.H., Demir I., et al. (2016), Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance, *Earth and Space Science*, 3, 388– 415, doi:10.1002/2015EA000136.

Gomes V.C., Queiroz G.R., Ferreira K.R. (2020). An Overview of Platforms for Big Earth Observation Data Management and Analysis. *Remote Sens*, *12(8)*, 1253. doi:10.3390/rs12081253

Gries C., Servilla M., O'Brien M., Vanderbilt K., Smith C., Costa D., Grossman-Clarke S. (2019). Achieving FAIRData Principles at the Environmental Data Initiative, the US-LTER Data Repository. Biodiversity Information Scienceand Standards 3: e37047. doi: 10.3897/biss.3.37047

Hanson P.J., Walker A.P. (2020). Advancing global change biology through experimental manipulations: Where have we been and where might we go? *Glob Change Biol*.; 26: 287– 299. doi:10.1111/gcb.14894

Hart E.M., 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. doi:10.1371/journal.pcbi.1005097

Hartig F, Dyke J, Hickler T, Higgins S, O’Hara R, Scheiter S, Huth A. (2012). Connecting dynamic vegetation models to data – an inverse perspective. *Journal of Biogeography.* **39**: 2240-2252. doi:10.1111/j.1365-2699.2012.02745.x

Hartig F, Minunno F, Paul S. (2019). BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics. R package version 0.1.7

Hasselbring W, Carr L, Hettrick S, Packer H, and Tiropanis T. (2020). From FAIR research data toward FAIR and open research software, *it - Information Technology*, *62*(1), 39-47. doi: doi:10.1515/itit-2019-0040

Herger N, Abramowitz G, Sherwood S. *et al.* (2019). Ensemble optimisation, multiple constraints and overconfidence: a case study with future Australian precipitation change. *Clim Dyn* **53,** 1581–1596. doi:10.1007/s00382-019-04690-8

Hoffman F.M., Koven C.D., 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 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. doi:10.5194/gmd-12-1119-2019

Huntzinger D.N., Schwalm C.R., 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. doi:10.3334/ORNLDAAC/1225.

Jeltsch F., Blaum N., Brose U., Chipperfield J.D., Clough Y. et al. (2013). How can we bring together empiricists and modellers in functional biodiversity research? Basic and Applied Ecology. 14:2, 93-101. doi:10.1016/j.baae.2013.01.001

Keenan T.F., Davidson E.A., Munger J.W., Richardson A.D. (2013). Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle. *Ecological Applications, 23(1), 273–286.* doi:10.1890/12-0747.1

Kelley D.I., Prentice I.C., Harrison S.P., Wang H., Simard M., Fisher J.B., Willis K.O. (2013). A comprehensive benchmarking system for evaluating global vegetation models. *Biogeosciences*. **10:** 3313–3340, doi:10.5194/bg-10-3313-2013

Kraemer G., Camps-Valls G., Reichstein M., Mahecha M.D. (2020). Summarizing the state of the terrestrial biosphere in few dimensions, Biogeosciences, 17, 2397–2424, doi:10.5194/bg-17-2397-2020.

LaDeau S.L., Han B.A., Rosi-Marshall E.J., et al. (2017). The Next Decade of Big Data in Ecosystem Science. *Ecosystems* **20:** 274–283. doi:10.1007/s10021-016-0075-y

Lai J., Lortie C.J., Muenchen R.A., Yang J., Ma K. (2019). Evaluating the popularity of R in ecology. Ecosphere 10(1):e02567. doi:10.1002/ecs2.2567

LeBauer D.S., Wang D., Richter K.T., Davidson C.C., Dietze M.C. (2013). Facilitating feedbacks between field measurements and ecosystem models. *Ecol. Monogr.* **83:** 133–154. doi:10.1890/12-0137.1

LeBauer D.S., Kooper R., Mulrooney P., Rohde S., Wang D., Long S.P., Dietze M.C. (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

Lovenduski N.S., Bonan G.B. (2017). Reducing uncertainty in projections of terrestrial carbon uptake. Environmental Research Letters. **12.** doi:10.1088/1748-9326/aa66b8

Lowndes J., Best B, Scarborough C. et al. (2017). Our path to better science in less time using open data science tools. *Nat Ecol Evol* **1,** 0160.doi:10.1038/s41559-017-0160

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

McKiernan E.C., Bourne P.E., Brown C.T. et al. (2016). How open science helps researchers succeed. Elife 5:e16800. doi:10.7554/eLife.16800

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

Nagaraj A., Shears E., de Vaan M. (2020). Improving data access democratizes and diversifies science. Proceedings of the National Academy of Sciences, 117 (38) 23490-23498; doi: 10.1073/pnas.2001682117

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

Piccolo S.R., Frampton M.B. (2016). Tools and techniques for computational reproducibility, *GigaScience*. **5.** doi:10.1186/s13742-016-0135-4

Pinnington E., Quaife T., Lawless A., Williams K., Arkebauer T., Scoby D. (2020). The Land Variational Ensemble Data Assimilation Framework: LAVENDAR v1.0.0, Geosci. Model Dev., 13, 55–69, doi:10.5194/gmd-13-55-2020.

Powers J.G., Klemp J.B., Skamarock W.C., et al. (2017). The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions. *Bull. Amer. Meteor. Soc.,* **98**: 1717–1737. doi:10.1175/BAMS-D-15-00308.1

Powers S.M, Hampton S.E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications* 29(1):e01822. doi:10.1002/eap.1822

R Core Team (2020). R: A language and environment for statistical computing. R version 4.0.3. Vienna, Austria: R Foundation for Statistical Computing.

Raiho A., Dietze M., Dawson A., Rollinson C.R., Tipton T., McLachlan J. (2020). Determinants of Predictability in Multi-decadal Forest Community and Carbon Dynamics. bioRxiv, doi:10.1101/2020.05.05.079871

Reichstein M., Camps-Valls G., Stevens B. *et al.* (2019). Deep learning and process understanding for data-driven Earth system science. *Nature* **566,** 195–204. doi:10.1038/s41586-019-0912-1

Reyer C.P.O., Silveyra Gonzalez R., Dolos K., Hartig F., et al. (2020). The PROFOUND Database for evaluating vegetation models and simulating climate impacts on European forests, Earth Syst. Sci. Data, 12, 1295–1320, doi:10.5194/essd-12-1295-2020.

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

Schimel D., Schneider F.D., 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 C.R., Schaefer K., Fisher J.B., et al. (2019). Divergence in land surface modeling: linking spread to structure. *Environ. Res. Commun.* **1:** 111004 doi:10.1088/2515-7620/ab4a8a

Seidel S.J., Palosuo T., Thorburn P., 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, doi: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. doi:10.1007/s10021-016-0068-x

Shiklomanov A. N., Cowdery E. M., Bahn M., Byun C., Jansen S., et al. (2020). Does the leaf economic spectrum hold within plant functional types? A Bayesian multivariate trait meta-analysis. *Ecological Applications* 30(3):02064. doi: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. doi:10.1111/gcb.14815

Steeneveld G., de Arellano J.V.-G., (2019). Teaching Atmospheric Modeling at the Graduate Level: 15 Years of Using Mesoscale Models as Educational Tools in an Active Learning Environment. *Bull. Amer. Meteor. Soc.,* **100:** 2157–2174. doi:10.1175/BAMS-D-17-0166.1

Stucky B.J., Guralnick R., Deck J., Denny E.G., Bolmgren K., 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. doi:10.3389/fpls.2018.00517

Sullivan I., DeHaven A., Mellor D. (2019). Open and reproducible research on open science framework. Current protocols, 18(1), e32. doi:10.1002/cpet.3

Tao F., Zhou Z., Huang Y., Li Q., Lu X., Ma S., Huang X., Liang Y., Hugelius G., Jiang L., Doughty R., Ren Z., Luo Y. (2020). Deep Learning Optimizes Data-Driven Representation of Soil Organic Carbon in Earth System Model Over the Conterminous United States. Frontiers in Big Data. 3:17. doi:10.3389/fdata.2020.00017

van Oijen M. (2017). Bayesian Methods for Quantifying and Reducing Uncertainty and Error in Forest Models. *Current Forestry Reports*. **3:** 269–280. doi:10.1007/s40725-017-0069-9.

Waide R.B., Brunt J.W., Servilla M.S. (2017). Demystifying the Landscape of Ecological Data Repositories in the United States, *BioScience*, 67(12): 1044–1051. doi:10.1093/biosci/bix117

White E.P., Yenni G.M., Taylor S.D., et al. (2019). Developing an automated iterative near-term forecasting system for an ecological study. *Methods Ecol Evol*. **10**: 332– 344. doi:10.1111/2041-210X.13104

Wieder W.R., Lawrence D.M., Fisher R.A., Bonan G.B., Cheng S.J., Goodale C.L., et al. (2019). Beyond static benchmarking: Using experimental manipulations to evaluate land model assumptions. *Global Biogeochemical Cycles*, 33, 1289– 1309. doi: 10.1029/2018GB006141

Acknowledgements

We are grateful to Kristina Anderson-Teixeira and Mingkai Jiang for their reviews which improved this manuscript to a great extent. We further thank Gab Abramowitz, Veronika Eyring, Michael Fienen, Andy Fox, Yuan Gao, Birgit Hassler, Xin Huang, Randall Hunt, Lifen Jiang and Jeremy White for their helpful overview on example cyberinfrastructure tools. The PEcAn project which organized the workshop where the authors of this paper came together, is supported by the NSF (ABI no. 1062547, ABI no. 1458021, ABI no. 1457897, ABI no. 1062204, 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

All authors were present in the workshop where these ideas were discussed. 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.

Competing interests

The authors declare no competing interests.

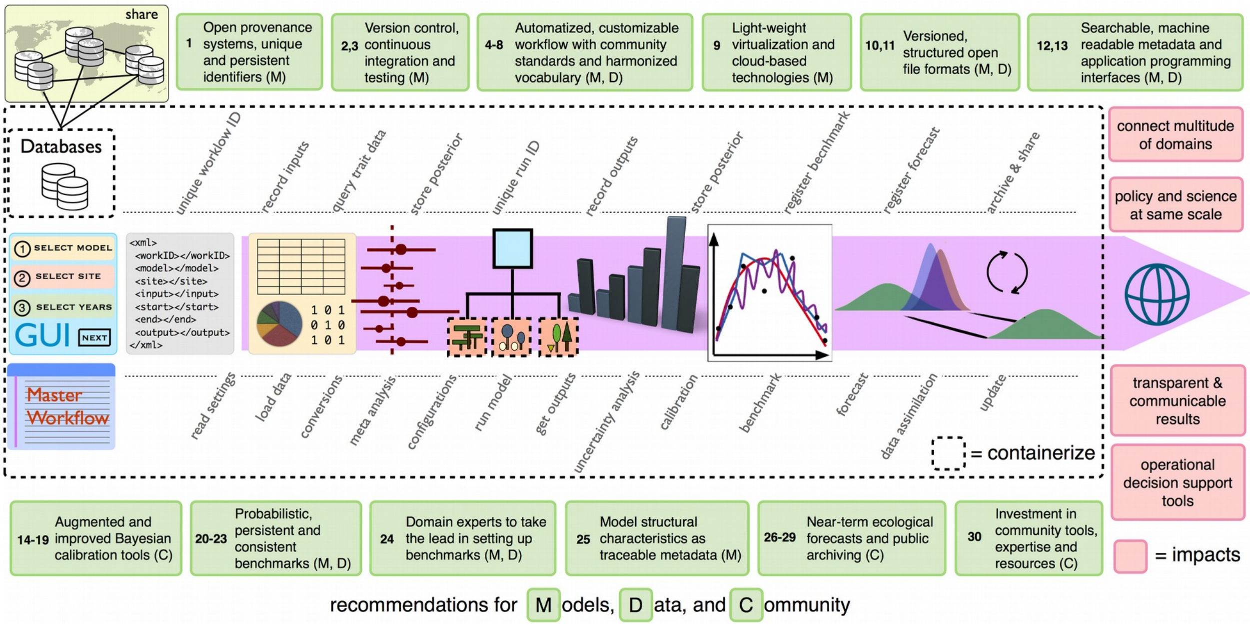

Figure 1. Schematic of a community cyberinfrastructure example and summary of recommendations (numbers in the green boxes refer to our recommendations in the main text). Users start with a high-level Graphical User Interface (GUI) to provide their setup for a modeling activity. These selections are translated into a human and machine-readable markup language and read in by the master workflow which then executes a sequence of modularized 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 in an earlier modeling activity 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 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.

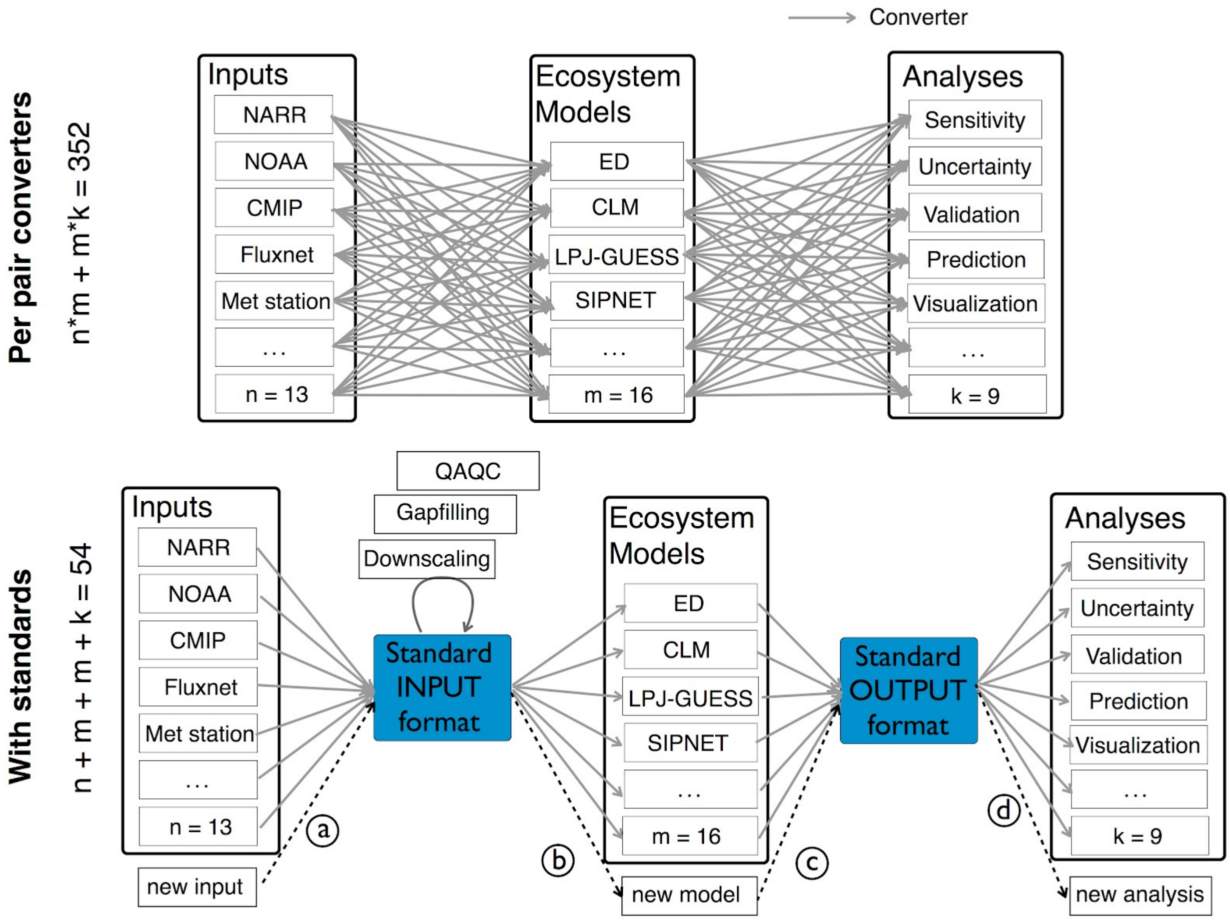

Figure 2. Reduction in redundant work when adopting common formats. There are "*n*" data types that must be linked to "*m*" simulation models and "*k*" post simulation analyses. In the top panel, the conventional approach where modeling teams work independently requires implementing *n**m* different input and *m**k* different output conversions. As data, models, and analyses are added, effort scales quadratically. On the other hand, the 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 reduces to an *m*+*n* and *m*+*k* problem, and scales linearly. 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. This scaling 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.

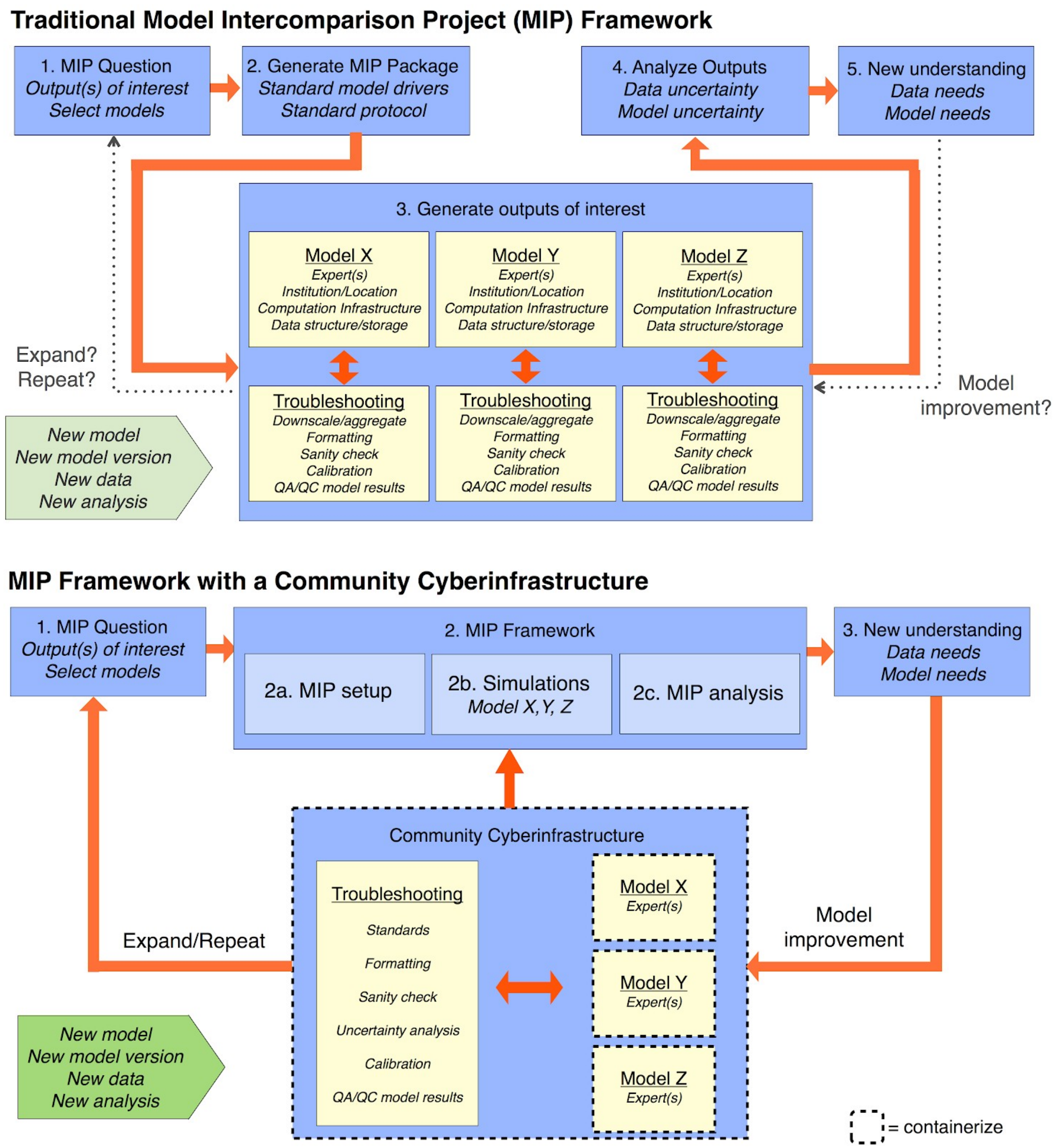

Figure 3. Traditional multi-model intercomparison project (MIP) workflow versus Community Cyberinfrastructure. Historically, each model and associated experts/infrastructure individually engage with MIPs (top). While stimulating model improvement is intended, it is not inherently nor readily available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