Requirements Driven Community Modeling Framework for the Earth and Environmental Sciences

Steve Berukoff, David Schimel, and Brian Wee
National Ecological Observatory Network (NEON), Inc.[1]

Introduction

The biosphere, Earth's living component, is one of our planet's most complex and fascinating systems and is also the source of many vital services to humanity. The biosphere influences, and is influenced by, physical, chemical and geological processes and is arguably the least well-understood of these systems. The biosphere and the physical Earth System interact strongly over diurnal and seasonal cycles, but also have critical interactions over decades and centuries. Currently, the ability to observe or reconstruct long-term, coupled behavior between living and non-living components of the Earth System is limited in both spatial and temporal scales and must be improved to support long-term ecological forecasting. Additionally, while humanity now affects nearly the entire biosphere, our understanding of how the biosphere operates in these landscapes is limited because most studies focus on organisms in pristine or minimally altered ecosystems. Thus, developing a better understanding of the physiology, distribution, and evolution of organisms in human-dominated landscapes is essential.

The National Ecological Observatory Network (NEON) will provide large amounts of information on a huge number of ecosystem attributes by deploying approximately 15,000 sensors of roughly 200 distinct types, making biological measurements on about 2000 plots distributed over 62 sites, and gathering airborne hyperspectral and LiDAR data, collecting in sum about a petabyte of information each year. The results of this data collection will be about 500 distinct primary scientific data types and 120 types of derived ecological parameters. Samples collected by the network will yield 175,000 chemical, taxonomic, isotopic and genomic analyses per year, and a similar number of samples will be stored annually for future research.

NEON will collect basic data through a combination of instrumentation and human observers. Basic calibrated data and data that have been temporally or spatially rectified will be processed using state-of-the-art algorithms and models to produce high-level, synthetic data products that both specialist and non-specialist scientists can use to rapidly and effectively to address ecological problems, as well as facilitate educational curriculum development and discovery. The algorithms to produce NEON data products, and the full provenance describing the flow from instrument to mapped output must be captured in the information system to allow replication of results, and provide for a natural introduction for the user community to extend the utility and impact of NEON data.

The identification of key algorithms, the sharing and reuse of those algorithms, their documentation, and then the ability to record entire complex workflows will be central to

[1] *The National Ecological Observatory Network is a project sponsored by the National Science Foundation and managed under cooperative agreement by NEON, Inc. This material is based upon work supported by the National Science Foundation under the following grants: EF-1029808, EF-1138160, EF-1150319 and DBI-0752017. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.*

EarthCube and NSF's Major Research Equipment and Facilities Construction (MREFC) projects like NEON, OOI, and others, as will the ability of the community to locate and use data products from complex algorithms and models as well as the models themselves. This aspect of community modeling, and its ecosystem of surrounding infrastructure, is described in more detail below.

Requirements Driven Infrastructure

Facilitation of community efforts, like the modeling discussed here, is optimized if recognition and integration of those requirements occurs at the earliest stages of the project, as it is within this framework that the ability to engage in community modeling activities is fostered, and constrained. This environment includes several aspects crucial to such efforts' success.

The "Era of Observations" and its impact
In the past fifteen years, the ecological sciences has embarked on a renaissance in the methodologies of observation, measurement, and simulation, and the recognition that local, site-based studies of ecological processes no longer fully suffice in addressing salient issues driving the field. NSF Director Suresh has remarked that science is entering an "Era of Observations" in light of the advent of large-scale, intensive, and extensive tools that observe processes ranging from nano-scale interactions to the heart-beat of an entire continental ecosystem. Such capabilities imply the need to store, process, and curate increasing amounts of data for research, education, resource management, and policy purposes. The need for heightened spatiotemporal resolution and scope in resulting data has outpaced efforts to secure data interoperability, develop communities of practice, and train new generations to develop and utilize tools that address scientific challenges. There is also a growing call for data from Federally funded research to be made easily discoverable and accessible. In what is foreseen as a trajectory towards long-term fiscal austerity, collaboration and the leveraging of existing infrastructure are the cornerstones of an architecture to accelerate innovative science. We posit that NSF's large-scale infrastructure investments have a (perhaps understated) mission to act as hubs for these activities, within their current scope and with an eye toward extension of it.

Toward an ecology of soft cyberinfrastructure
A central piece to this puzzle is the development of an ecology of a "soft" cyberinfrastructure, built and maintained as a service. Part of the impetus for this is driven by the fact that there are relatively few in the community who have interest or expertise in the engineering aspects of building software: tools for analysis, visualization, workflow management, and modeling, developing sustainable and widely-accepted definitions of data and metadata standards, "middleware" tools for interoperability, etc. However, while large facilities can develop such infrastructure for their own purposes, they must also engender community support and input if their efforts are to succeed.

Moreover, while a funding agency may contribute funds toward building a central hub of such activity, there are limitations to such a model. An alternative model that merits mention borrows on the success of Linux kernel development: while there exists substantial investment by corporate interests, a great deal of the "community" upon with Linux is built depends vitally on the distributed, constrained anarchy that encourages new ideas and development yet streamlines such effort within a larger scope. It is here that large facilities can best serve as a “hub” – by providing structure (via working groups, “hackathons”, user workshops and tutorials, software & model repositories, etc.) that encourages contributions and community engagement. Such an “ecology” is sustainable and largely sheltered from the vagaries of funding nadirs, and solicits substantial feedback and interaction while keeping costs manageable. Ultimately, encouragement of this nascent environment is a feedback loop, since the satisfaction of large project requirements - software deliverables, community algorithm and data acceptance milestones, interoperability, etc. - is streamlined and accelerated, benefiting all.

Infrastructure in a box

While the primary output of large facilities is often considered its data, the infrastructure that creates the data suites is easily as valuable. However, this is only true if both are standardized across the breadth of facility activities, from the installation of sensors and establishment of field protocols to the storage of data streams and the detailed management of subsystem linkages. One primary upshot of this “superstructure” is that all of the facility’s deliverables can be considered products – science designs, metadata schema, data formats, and, algorithms. That the algorithms are “products” like their data brethren is essential to the establishment of community modeling efforts, since while facilities like EarthCube and NEON are nominally tasked with providing data, the “nexus responsibility” of the facility provides a home for the community to further develop its acumen. Since an essential component of facility success is its working groups, and as a hub for community activity the facility is a general resource, this is a natural model for ensuring that funding agencies’ monies are spent responsibly, transparently, and with maximal utility.

Community Modeling as a Service

We have argued above for the responsibility of large facilities to act as a nexus for community activities by providing the infrastructure that supports its advancement. This is partly due to the need to model complex natural systems, wherein single analysis tools mirror that complexity to such an extent that the tools become “facilities” in and of themselves. The resulting integration of a project’s science needs with the tool’s “facility” thus places a requirement on both to interact, interdepend, and coexist, which can only occur through a community modeling approach. We discuss one current effort underway at NEON as a case study.

Assimilating land surface models and network data

Land surface models are the components of global circulation models (GCMs) or earth system models (ESMs) that represent the role of vegetation, soils and terrestrial water bodies in the earth system. They represent the surface water and energy balance and, importantly, the environmental responses and time evolution of the land carbon sources and sinks, such as vegetation. The effects of climatic forcings and responses on such carbon pools can be empirically scaled to national or continental scales, and many key processes involved agree with one another to a certain extent. However, overfitting of extant global data has manifested through divergence in model comparisons, and thus even well-validated models can retain large uncertainties in the aggregate view, highlighting the dependence of uncertainty on the form and parameterization of equations used, and the constraints on the input parameter set.

Data assimilation is a general term for methods that systematically combine observation and model information to achieve an understanding of the system that is more accurate than either independently. Data assimilation techniques basically involve making multiple comparisons between observations and model predictions to estimate the most likely values for parameters or state variables. Applications of data assimilation in carbon cycle research include (a) gap filling of missing measurements, (b) estimation of process model parameters or other non-observed model-derived quantities, or (c) forecasting of future model states. Data assimilation has been widespread in the geosciences (particularly hydrology), atmospheric sciences, and, more recently, ecology, especially with regard to estimates of carbon pools. The goal of ecological carbon data assimilation is often parameter, rather than state, estimation, because parameter estimates in carbon cycle models give insight into process-level responses to environmental variation, *e.g.*, the temperature sensitivity of respiration or the photosynthetic response to humidity. In addition, carbon modeling typically probes coupled systems with very different time constants (minutes to decades or longer) that must be considered simultaneously, and on a background of "slow" geophysical processes that typically appear only as initial or boundary conditions. While reanalysis of atmospheric and oceanic data can reveal the role of slowly varying (typically oceanic) processes, diagnosis of the role of slower processes is a key challenge in carbon modeling.

At NEON, the combination of land-surface modelling and data assimilation sits at the heart of an effort to build several foundational data products, providing information on carbon and water balance across the continent, which, in turn, informs data products that address NEON's Grand Challenges in climate change, land use change, and invasive species. The land surface is modelled at high resolution using the Community Land Model (CLM), which, as its name implies, is a successful model with international software and scientific support. For NEON's purposes, several aspects of the CLM need modification, leading to an outstanding collaboration between NEON staff, NCAR staff, and the community at large, composed of faculty, postdocs, and graduate students from around the globe. Meanwhile, a sequential, ensemble Kalman filtering (EnKF) technique provides some of the glue that enables estimation of parameter and model uncertainty, a key NEON output. The EnKF linkage is being performed in collaboration with researchers from NCAR and Oak Ridge National Lab, among others, with input and support again from the community of modelling and statistics researchers.

We would like to highlight the endpoint of this development path and why it is unique. When complete, the model-data assimilation framework under construction will be made widely available, and, via programmatic interface to NEON data, researchers will be able to freely and openly use the algorithms in concert with NEON data to perform and extend the analyses. NEON will not "own" the resulting algorithms, but will support the modelling community in the ecological sciences: all of the tools (and relevant documentation) will be freely and openly accessible, and as the science community identifies needs for alterations, improvements, or additions, NEON will facilitate development and respond to the user community through organized sessions, workshops, and even helpdesk requests made through its web portal. Further, NEON plans to curate algorithms, code, and "auxiliary" information like input files (initial conditions, namelists, etc.) enabling fully transparent provenance of the processing pipelines.

As another contribution to the community, NEON has also been involved in forums about data citation standards, such as those being advanced by the National Snow and Ice Data Center (NSIDC) via participation in the Earth Science Information Partners (ESIP) and the USGS Community for Data Integration (CDI). We, together with others, have also been promoting the idea of using data publications in performance evaluations such as those affecting tenure and pay-scale. The topic of data citation has been frequently brought up in the National Science Board’s deliberation on data policies, and while there remain substantial financial sustainability and cultural barriers, we believe that the path towards resolving these issues will see the increased sharing of data and the repurposing of well documented data for purposes not foreseen by the original publishers of those data.

Paths Forward

In the ecological community, the community model concept needs to be more heavily promoted and championed. It is therefore incumbent on large facilities like NEON to market it as an essential feature of our burgeoning “large data” science era. There are, however, a number of ‘sociological’ challenges before the scientific community at-large. For instance,

- Merging modern developments in computing, such as the use of public or private cloud resources in support of interoperability, or of business intelligence (BI) tools that enable deep data mining and data discovery, will contribute greatly to advancing our sciences, but not unless we, as scientists, think out of the box and engage industry partners;
- Past validating individual codes, the benchmarking of complex, nonlinear algorithms must be *de novo*, commonplace practice. Cross-validation engenders community engagement and support while ensuring we are doing the best science possible, and enables our funding agencies to demonstrate the benefits of Federally funded research, and how these may be translated for societal benefits;
- Recognizing that there is a distinction between research-grade, open-source code and community modeling efforts. The former is the *de facto* mode of research for most scientists, while the latter supports science while requiring a social context and rules of engagement that must be contemplated and cultured.

Large facilities projects like EarthCube, NEON, and OOI have an excellent opportunity to steward not only data but their respective scientific communities, from their sociological intrarelationships to their metadata-based schema interoperability. As we move forward in time, our success will be reflected in a concomitant understanding of our scientific processes, providing a rich legacy upon which to build.

The authors would like to acknowledge Andy Fox (NEON Postdoctoral Researcher) and representatives from NEON, Inc.’s Member Institutions (http://www.neoninc.org/about/members) who discussed the requirements for such a community driven infrastructure during a breakout session at the NEON, Inc. Fourth Annual Membership Meeting in Boulder, CO (September 16, 2011).