Jun 16th, 3:40 PM - 5:20 PM

Integrating NEON data with existing models: An example with the Community Land Model

Edmund M. Hart
National Ecological Observatory Network, thart@neoninc.org

Andrew Fox
National Ecological Observatory Network, afox@neoninc.org

Steve Berukoff
National Ecological Observatory Network, sberukoff@neoninc.org

T. J. Hoar
IMAGE, CISL, National Center for Atmospheric Research, thoar@ucar.edu

International Environmental Modelling and Software Society (iEMSs)
7th Intl Congress on Env. Modelling and Software, San Diego, California, USA
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Integrating NEON data with existing models: An example with the Community Land Model

Edmund M. Hart [a] **Andrew Fox** [a], **Steve Berukoff** [a], **T. J. Hoar** [b]

[a]*National Ecological Observatory Network, 1685 38th St, Boulder, CO 80301, USA(thart@neoninc.org, afox@neoninc.org, sberukoff@neoninc.org)*

[b]*IMAGEe, CISL, National Center for Atmospheric Research, 1850 Table Mesa Dr, Boulder, CO 80305(thoar@ucar.edu)*

Abstract:

A central challenge to environmental forecasting in hydrological and land surface modeling is how to integrate multiple data sources over a wide range of spatial scales. Furthermore how can this complex task be achieved in the most productive and reproducible way with a robust informatics architechure? At the National Ecological Observatory Network (NEON) we are collecting a variety of biophysical and biogeochemical measurements which can be used with models to perform temporal forecasting on decadal timescales. To take advantage of this data we are developing a data assimilation framework. Using this framework NEON data can be combined with the Community Land Model, which features a fully coupled carbon and nitrogen cycle (CLM-CN). Our goal is to produce optimal solutions for model states, fluxes and parameter values, with their associated uncertainties, at regional to continental scales. Here we describe our initial trials of programmatically integrating NEON data streams with the CLM using the Data Assimilation Research Testbed (DART), a community tool for ensemble data assimilation (DA). We will provide an overview of the NEON informatics architecture, the workflow we employ, and outline how our emphasis on metadata and semantic infrastructure from the NEON project will enable others to use these data within their own data assimilation frameworks.

Keywords: informatics; data integration; metadata

1 INTRODUCTION

The National Ecological Observatory Network (NEON) is a continental scale 30 year observatory that will provide standardized observations for hundreds of different environmental variables, and provide over 500 data products for users [Schimel et al., 2011]. It will have 60 sites located across the US and Puerto Rico, including Alaska and Hawai'i, with 3 in each of 20 eco-regions delineated by a multivari- ate geographic clustering algorithm (MGC) [Schimel et al., 2011]. It's intention is to provide decadal and continental scale observations to better understand how ecosystems will respond to anthropogenic forcings and feedbacks at a broad temporal and spatial scale [Keller et al., 2008]. NEON will collect this data through three major avenues: tower mounted sensors, seasonal field crews collecting obser- vational measurements, and hyperspectral and lidar data via an airborne observation platform [Schimel et al., 2011]. These varied collection activities create a heterogeneous array of data products at varying spatial and temporal resolutions. Thus, one of the great challenges of NEON is: how can we build a robust informatics infrastructure that encompasses these different data products for use by the modeling community? Furthermore, how can we ensure that this massive amount of data is useful to modellers and the broader community of data consumers? Building this infrastructure to facilitate ecological and hydrological modeling serves both internal NEON data product creation as well as providing a workflow for the broader community.

Enabling coupled ecological and hydrological modeling using these data in Earth System Models (ESMs) is central to the operations of NEON with a goal to quantitatively predict, not just retroactively explain, land surface processes [Committee on the National Ecological Observatory Network, 2003; Keller et al., 2008]. Whether predicting the future state of a catchment, or assessing the likely impact of a particular climate scenario, forecasting has great societal benefit and can help advance theory by forcing hypotheses (as theory codified in the model) to be confronted with observations. Predicting the effects of drivers that occur over decades or longer (climate change, land use and land cover change) requires information on mechanisms that act over a range of time scales, as well as the parameters that influence their behavior. The covariance between environmental drivers and responses illustrates the strength of underlying mechanisms. By examining how it varies over time and space, we can provide the basis for spatial extrapolation and temporal forecasting. NEON will provide consistent observa- tions that are comparable with many important prognostic and diagnostic ESM variables, such as soil moisture and soil temperature, leaf area and biomass and fluxes of water and carbon. To support the integration of NEON data with existing models we are developing an informatics infrastructure to work with existing ESM's. One example is the incorporation of NEON data into the Community Land Model (CLM) [Lawrence et al., 2011] via a Data Assimilation Research Testbed (DART) [Anderson et al., 2009]. Integrating NEON data into the CLM via DART will allow us to create important gridded data products such as soil moisture, evaporation and transpiration for area of the landscape where there are no direct observations. Below we will outline how the CLM and DART work, and how they can work with NEON data. Furthermore, while we are still in construction, we will detail how the developing informatics in- frastructure will help facilitate data integration in the CLM and can serve as an example for how other models can integrate NEON data products.

2 CLM AND NEON

The long-term nature of the observatory will enable iterative comparisons of predictions and observa- tions as well as analyses of factors that most strongly influence forecast errors. This experience will also enable the measurements made by NEON to be evaluated over time, ensuring they remain relevant, ef- fective and efficient in a changing environment. By observing processes at different scales, from single organisms to the continent, NEON and non-NEON data (such as remotely-sensed products including MODIS LAI and MODIS snow cover and gridded biomass estimates derived from remote sensing and forest inventory analysis) provide detailed, site-specific information and spatial measures of patterns of water storage and fluxes in nature. For example, long-term eddy covariance flux measurements made over a variety of ecosystem types, and global products derived from them, have become critical in guiding the evaluation and development of land surface models. Models can be informed by flux tower records at NEON sites that will complement site characterization data and measurements of soil properties, observations of vegetation properties and dynamics that control evapotranspiration, such as

airborne derived estimates of leaf area index, and soil moisture and groundwater dynamics. Infrastructure to help explore some key modeling questions is being developed at NEON; these efforts include employing community tools for land surface modeling and advanced data assimilation, as well as identifying data streams and establishing processing pipelines to connect these tools to NEON observations. This ability will be useful for developing and testing ESM's, overcoming the limitations of short-term or episodic data collections to describe inherently non-stationary ecological systems. Iterative forecasts through time enable much larger areas of the solution space to be explored over a diverse range of conditions, leading to an orderly forecast evaluation/update/improvement cycle.

To provide a number of required high level ecohydrological and biogeochemical data products, we are developing a data assimilation framework that couples NEONs and remotely-sensed satellite data products to the CLM. The CLM is used as the land component in the Community Earth System Model (CESM), a collaborative effort between the National Center for Atmospheric Research (NCAR), the Department of Energy (DOE) and many university researchers with the aim of predicting and understanding the couple climate system. CLM simulates terrestrial ecosystem processes including the cycling of energy, water, carbon and nitrogen and is driven by a limited set of climate variables, which may come from site observations, reanalysis or a coupled atmospheric model, while the sensitivity of ecosystem processes to climate is controlled by the initial states and parameter sets of the model. In principle, estimates of initial conditions and parameters do not require long-term, standardized observations. Within the scope of a short-term research project, initial conditions at a site can be surveyed (for example, biomass or population data), key rate constants can be measured, and a model can be developed and exercised. Examples of such research abound, but this type of forecasting is limited by a dearth of long ecological time series [Clark et al., 2003; Magurran et al., 2010].

The measurements made by NEON will gradually change over time as experience is gained through cyclic prediction-observation comparison and the analysis of factors that most strongly drive forecast errors. Our goal is to produce optimal solutions for model states, fluxes and parameter values, with their associated uncertainties, at regional to continental scales. These gridded, land surface products will be somewhat analogous to atmospheric reanalyses, whereby a fixed model and data assimilation scheme ingest many varying observations to produce a dynamically consistent estimate of hundreds of state variables with consistent spatial and temporal resolutions. This removes the requirement of the end user to understand and analyze separately all the many different observations and makes the reanalysis data sets easy to handle from a processing standpoint, although the files sizes can become large. The downside of this approach is that changing mixes of observations, and observation and model bias, can introduce spurious variability and trends into the reanalysis. Thus reanalysis reliability can vary considerably depending on location, time period and the variables considered. One way to address some of these concerns is with a robust data assimilation framework.

Data assimilation is a general term for methods that systematically combine information from observations with information from a model to achieve an understanding of the system that is more accurate than the observations or the model independently. The data assimilation approach adopted by scientists working with land surface models draws on tools developed in meteorology and applied mathematics to support numerical weather prediction. The goal of land surface data assimilation is often parameter, as well as state, estimation, particularly for describing processes related to the terrestrial water and carbon cycles. Our approach has been to use ensemble filter techniques, approximate Monte Carlo solutions to the DA problem that have grown rapidly in popularity since their first description in the 1990s [Evensen, 1994]. By using careful software engineering, it is possible to develop a state-of-the-art ensemble filtering system that is mostly independent of the geophysical model and observations being assimilated. Such a system is the DART [Anderson et al., 2009], a community facility that employs a modular programming approach for ensemble DA developed and maintained at the NCAR that provides a number of enhancements to basic filtering algorithms (Figure1). We developed a multi-instance version of CESM that more easily facilitates ensemble-based data assimilation techniques that is now released as part of the normal CESM distribution. A key component to facilitate assimilating NEON data into the CLM via DART is a strong informatics architecture.

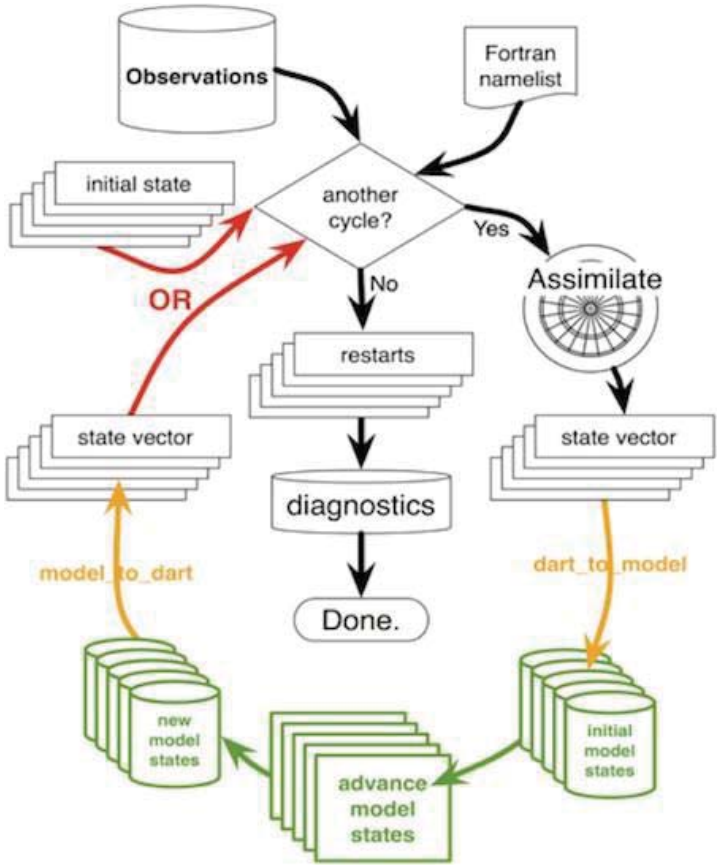

Figure 1. DART-model coupling and execution

3 INFORMATICS INFRASTRUCTURE

The CESM and DART have large and complex code bases. As scientists at NEON developing high level data products we rely upon this existing code base and the broader community's experience in using the model and ensemble filter tools. A critical activity at NEON then, in addition to contributing to development and testing efforts, is identifying which NEON observations are useful to constrain models and processing them into observation sequence files suitable for ingestion by DA schemes like DART. Observation sequences are complicated and trying to automatically accommodate a myriad of observation file formats, structure, and metadata is a difficult task. For this reason, DART has its own format for observations and a set of programs to convert observations from their original formats to DART's. The DART framework enforces a clean separation between observations and the models used for assimilation (Figure 2). The same observations can be used in any model which understands how to generate a value for the requested type of observation from the models' state-space values (i.e. the forward observation operator must exist - DART provides many for the most common state variables). In many cases, the original datasets are in a standard scientific format like netCDF, HDF, or BUFR, and library routines for those formats can be used to read in the original observation data. The DART software distribution includes Fortran subroutines and functions to help create a sequence of observations in memory, and then a call to the DART observation sequence write routine will create an entire observation file in the correct format. In many cases, a single, self-contained program can convert directly from the observation location, time, value, and error into the DART format. In other cases, especially those linking with a complicated external library (e.g. BUFR), there is a two-step process with two programs and an ASCII intermediate file.

Given the complications of ingesting external data into DART, a well thought out and robust informatics infrastructure can serve to facilitate this process. At NEON we are taking a multi-stepped approach to

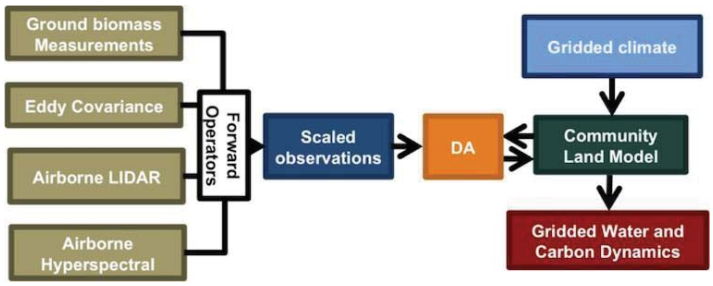

Figure 2. A data assimilation scheme for CLM, brown boxes represent data that NEON can provide.

building this infrastructure. Our first step is to adopt same basic guidelines that make using our data internally (and externally) simpler [White et al., 2013]. These include developing naming conventions, adopting file format standards across heterogeneous data sources, creating controlled vocabularies, and implementing community metadata standards. Using standard naming conventions and file formats allows for easy programmatic ingestion of NEON data into DART. If files are all structured with similar names, programmatic tools built for a single data type use case allow easy extensibility to other datum. This is especially important given the complexity moving from formats like HDF5 to the DART (Figure 1). While internal workflows are facilitated this way, external users will benefit from structured vocabularies and standardized metadata. Metadata for most measurement streams will be ISO-19115-2 [ISO-19115-2, 2009] compliant, allowing for automated ingestion of multiple files. Once we developed file format standards across all data products, we can develop tools to interface with existing community standards. One example is that much of our biophysical data will be served as HDF5 with self-described metadata for interoperability across NEON data products. However, users may want the same data product in a NetCDF format with self-describing metadata that meets Climate and Forecast (CF) conventions. To serve the goals of interoperability both across NEON data and existing data providers, another component of our infrastructure is the development of mappings to existing community standards (e.g. CF) and the tools to easily convert between formats. The development of consistent file format standards across products, providing metadata in existing standards, and developing tools for the community is the first phase of our informatics strategy. A long term goal is the development of a NEON semantic ontology's and streaming API's to integrate our data into CLM.

Semantic ontologies provide a formal way of describing relationships, and their logic between resources [Madin et al., 2007]. In the case of NEON, a formal ontology would describe the way in which data streams relate to each other, as well as to sensors, locations, and other data types. Our current plan for the development of an ontology is to create semantics that describe the NEON observation process. However, given the heterogeneity of NEON data products and community efforts at ontology development, our goal is that the internal observation ontology can then be integrated with existing community standards. Examples of existing ontologies that we potentially want to adopt for data products outside the observation process include Semantic Web for Earth and Environmental Terminology (SWEET) [Raskin and Pan, 2005] for our biophysical data, and the Biological Collections Ontology (BCO) [Walls et al., 2014] for our organismal data. A well developed NEON ontology that integrates with other ontologies would serve two primary functions in the integration into DART. First it allows for improved data discovery [Berkley et al., 2009]. Users could discover linkages between data streams that they might not otherwise have considered. For instance NEON will collect single aspirated air temperature from observational towers, but will also have micrometeorlogical stations at certain aquatic sites, with both data potentially being useful in the CLM. Secondly it allows for the improved automation of data ingest into DART and other environmental models [Villa et al., 2009]. A formal semantic ontology will provide linkages between heterogenous data sources from temperature recorded on a tower to vegetation surveys and soil types at the site. Tools can be developed that depend on these relationships, and can integrate these data sources before their inclusion into DART. Furthermore, once we have a stable API up, these semantic relationships between resources can be used to continuously stream data, and integrate it before incorporation into DART. The current plan for the development of a NEON ontology is to develop a semantic framework that internally describes

4 CONCLUSIONS

We have provided details about how the CLM and NEON data can work together to improve model predictions via the DART. The tools we develop have applicability beyond our own internal modeling needs. Many other models are also coupled to both the CLM and DART. For instance the Noah land surface model also uses DART [Ek et al., 2003], as does the Australian governments Community Atmosphere Biosphere Land Exchange (CABLE) model. The integration of NEON data into DART can serve as an example of how informatics infrastructure facilitates data assimilation with models. We have already accomplished some of the simpler parts of building out an informatics infrastructure. Our next steps are to build out the suite of open source tools for combining NEON data with DART, and develop a semantic ontology. Given that this is an ongoing process we can leverage this time during our construction phase to build out the ontology that specifically meets the needs of data assimilation. It will allow for an iterative agile development of our ontology and help operationalize the process of data assimilation between NEON and the CLM.

REFERENCES

Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano, A. (2009). The Data Assimilation Research Testbed: A Community Facility. *Bulletin of the American Meteorological Society*, 90(9):1283–1296.

Berkley, C., Bowers, S., Jones, M. B., Madin, J. S., and Schildhauer, M. (2009). Improving Data Discovery for Metadata Repositories through Semantic Search. *2009 International Conference on Complex, Intelligent and Software Intensive Systems*, pages 1152–1159.

Clark, J. S., Lewis, M., McLachlan, J. S., and HilleRisLambers, J. (2003). Estimating population spread: What can we forecast and how well? *Ecology*, 84(8):1979–1988.

Committee on the National Ecological Observatory Network, N. R. C. (2003). *Neon: Addressing the Nation's Environmental Challenges*. The National Academies Press.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Gunnmann, P., Koren, V., Gayno, G., and Tarpley, J. D. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research*, 108(D22).

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C):10143.

ISO-19115-2 (2009). Geographic information – Metadata – Part 2: Extensions for imagery and gridded data.

Keller, M., Schimel, D., Hargrove, W., and Hoffman, F. (2008). A continental strategy for the National Ecological Observatory Network. *Frontiers in Ecology . . .* , pages 282–284.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., Bonan, G. B., and Slater, A. G. (2011). Parameterization improvements and functional and structural advances in Version 4 of the Community Land Model. *Journal of Advances in Modeling Earth Systems*, 3:1–27.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279–296.

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. M., Elston, D. a., Scott, E. M., Smith, R. I., Somerfield, P. J., and Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in ecology & evolution*, 25(10):574–82.

Raskin, R. G. and Pan, M. J. (2005). Knowledge representation in the semantic web for earth and environmental terminology (sweet). *Computers & Geosciences*, 31(9):1119–1125.

Schimel, D., Keller, M., Berukoff, S., Kao, B., Loescher, H., Powell, H., Kampe, T., Moore, D., and Gram, W. (2011). The national ecological observatory network 2011 science strategy: enabling continental scale forecasting. Technical report, National Ecological Observatory Network.

Villa, F., Athanasiadis, I. N., and Rizzoli, A. E. (2009). Modelling with knowledge: A review of emerging semantic approaches to environmental modelling. *Environmental Modelling & Software*, 24(5):577–587.

Walls, R., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P., Davies, N., Endresen, D., Gandolfo, M., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., Midford, P., Morrison, N., O Tuama, E., Schildhauer, M., Smith, B., Stucky, B., Thomer, A., Wieczorek, J., Whitacre, J., and Wooley, J. (2014). Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE*, 9(3):e89606.

White, E., Baldridge, E., Brym, Z., Locey, K., McGlinn, D., and Supp, S. (2013). Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution*, 6(2):1–10.