1 Introduction

Feedbacks from the terrestrial biosphere are one of the largest sources of uncertainty in climate change projections (1, 2). Unfortunately, despite a diverse set of modeling approaches, conventional methods to reducing these uncertainties have not progressed rapidly (Figure 1). The slow pace of improvement has occurred despite an unprecedented amount and diversity of data about the terrestrial biosphere, but this data is not being fully utilized to test and improve models.

The overarching goal of our project has been to accelerate the pace of model improvement by making data and models more accessible. In our earlier Innovation proposal, we noted that a small number of important gaps separate the information we have gathered from the understanding required to improve models and inform policy and management. The Predictive Ecosystem Analyzer (PEcAn) project has thus far focused on three of these gaps: (1) no single data source provides a complete picture of the terrestrial biosphere, and therefore multiple data sources must be integrated in a sensible manner; (2) current modeling approaches only makes use of a subset of the available data; and (3) this is in large part due to a need for tools to manage the assimilation of data into models. Our Innovation award allowed us to successfully develop and test the database, workflow, and user-friendly application interface that form the PEcAn ecoinformatics toolbox. Here, we seek to build on our success to expand, deploy, and disseminate this framework.

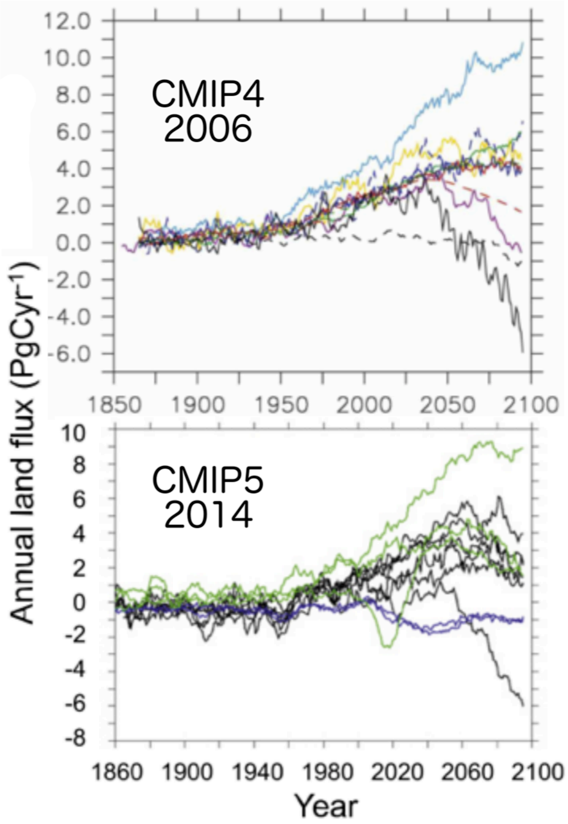


Figure 1: Climate model terrestrial carbon cycle projections show negligible improvement between IPCC assessments. Figures modified from (1, 2)

In this Development proposal we continue to focus on closing these three gaps, but have identified three additional challenges that drive many of the advances herein:

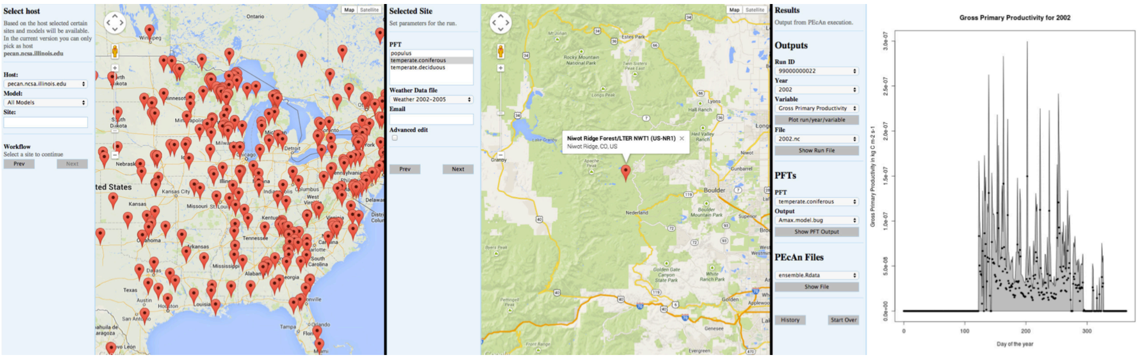
1) The need for a scalable solution to improving models, both in the cyberinfrastructure and the involvement of the research community;
2) The need to assess, track, and analyze model skill, structure, and uncertainty;
3) The need to increase the accessibility of modeling tools.

We assert that the root of the failure of models to improve (Figure 1) is that ecosystem modeling, as a discipline, is not currently scalable. Rather, modeling is mainly a cottage industry of creating and refining models where work on one model does not inform other models. If one team applies a model to a specific study system, the next model team is essentially starting from scratch because their model inputs and outputs are not interoperable, as most models use different formats. In addition, these files are rarely shared among teams, and when shared, it is primarily at the time of publication, which may lag the initial modeling effort by years. Even within the community of users for a single model, as that community grows the work done by one lab often does not benefit other users directly. **While the use of version control software for managing code is standard practice among ecosystem modeling teams, there is not an equivalent system keeping track of where a model has been run, the inputs and outputs of**

such runs, comparisons to data, and what was learned from that activity. Similarly, despite the growth in the application of model-data fusion techniques (3), there's currently no way to combine information across such studies in order to better constrain model parameters or understand how these parameters vary from site to site. **We lack a way to systematically and routinely assess, track, and analyze model skill and determine how that skill relates to model structure and parameter uncertainty**. Improving models requires a scalable cyberinfrastructure approach for coordinating information within and across modeling teams.

Furthermore, most data collected is not ever reaching these models, as most ecologists and physiologist are not modelers, though the overwhelming majority of field and lab scientists want their research to inform models (4). Most empirical observations are "long tail," uncurated data, often unpublished, and the volume and heterogeneity of these data means that the modeling community cannot keep pace with the rate of data generation. Under current conditions, even when ecologists and physiologists do collaborate with modelers, their efforts rarely do more than improve a single model for the reasons discussed earlier. To make ecosystem modeling scalable requires not only that we build the cyberinfrastructure to allow for better synthesis within and across models, but also that we build a community-based approach to confronting models with data. The larger research community needs tools that allow them to play a more active role in model calibration, assessment, and improvement. **Achieving this means that models, and the analytical tools to bring together models and data, have to be made more accessible**.

In PEcAn v1 we built an **ecoinformatics toolbox** (Figure 2). This system hides individual models beneath a common Application Program Interface (API) that uses standardized model input and output formats. This workflow allows the construction of generic, reusable tools for processing model inputs, for analyzing and visualizing model outputs, and for assimilating data into models. It also logs all of these activities and files in a database, and places a simple web-based front end on this database and a number of the PEcAn modeling tools. To date, we and our collaborators have successfully registered four models and almost thirty thousand observations into this system, and have used this toolbox to conduct novel research quantifying model uncertainty (5) and optimizing field data sampling (6).



Select host

Based on the host selected certain sites and models will be available. In the current version you can only pick as host pecan.ncsa.illinois.edu

Host: pecan.ncsa.illinois.edu

Model: All Models

Site:

Workflow

Select a site to continue

Selected Site

Set parameters for the run.

PFT: populus, temperate.coniferous, temperate.deciduous

Weather Data file: Weather 2002-2005

Email

Advanced edit

Niwot Ridge Forest/LTER NWT1 (US-NR1)

Niwot Ridge, CO, US

Results

Output from PEcAn execution.

Outputs

Run ID: 99000000022

Year: 2002

Variable: Gross Primary Productivity

File: 2002.nc

PFTs

PFT: temperate.coniferous

Output: Amax.model.bug

PEcAn Files

ensemble.Rdata

Figure 2: PEcAn web interface for site-level runs and output visualization

In PEcAn v2, we propose to build a **robust and scalable peer-to-peer network** around this toolbox. Different instances of the PEcAn database operated by different users and modeling teams will sync information across databases and share files. This will allow communication and coordination across the network of models without the need for top-down control over the modeling process, allowing dissemination of tools and data to occur organically and enhancing the long-term sustainability of the code. Users would be able to freely form subnetworks and would control what and when they share with the network. We believe such a network will enable users to spend more time doing science, less time doing redundant data management and building redundant analytical tools, and therefore **increase the pace of model improvement and our ability to forecast ecosystem services**.

2 Results from previous research: PEcAn v1

Collaborative Proposal: ABI Innovation: Model-data synthesis and forecasting across the upper Midwest: Partitioning uncertainty and environmental heterogeneity in ecosystem carbon *07/2011-06/2014 M Dietze, K McHenry NSF-ABI-1062547, $770,653 & A Desai NSF-ABI-1062204, $103,922*

Intellectual Merit

In our ABI Innovation project, we developed the PEcAn framework for ecosystem model analysis and data assimilation. PEcAn is an open-source, modular workflow that manages the flow of information into and out of ecosystem models. By wrapping models in common interface modules that use the same inputs and outputs in a common format, we have created an API that allows all tools to be developed in a general, reusable manner. This greatly reduces the amount of redundant work required by different modeling teams, allowing them to focus more on the science, and allows users to work with multiple models with less of a model-specific learning curve. To date PEcAn has been coupled with four ecosystem models, the Ecosystem Demography (ED) model (7), SipNET (8), BioCro (9), and DALEC (10), but this approach is scalable to the scores of ecosystem models currently used by the community.

In PEcAn v1, tool development focused on the processing of inputs, the visualization and analysis of outputs, and the estimation of ecosystem pools and fluxes using model-data fusion (3). Because ecosystem models aim to simulate real-world ecosystems across large spatial scales, often at a sub-hourly temporal resolution, processing model inputs can be large bottleneck. It often involves combining terabytes of data from multiple data streams in a wide array of custom binary formats, gap-filling missing observations, and scaling the inputs to the model's spatial and temporal resolution. PEcAn hides all of this under the hood, instead presenting the user with a Google-map based web interface (Figure 2). When runs complete, outputs are likewise visualized using interactive web tools and are made available for download in CSV, XML, and NetCDF formats (depending on output), using standard data formats and metadata conventions. In between, PEcAn's workflows not only handle the I/O processing, but also record the history of all operations and files in a PostgreSQL database. All previous runs can be retrieved in an interactive manner, and the full contents of the database is both human and machine accessible through a Ruby-on-Rails web interface. In addition, because many of the models run in a high-performance computing environment, PEcAn supports the remote execution of both the models and the data processing tools, including a variety of resources in NSF's XSEDE network (11).

PEcAn v1 automates a set of tools aimed at quantifying and propagating parameter uncertainty in models (4). This starts with prior probability distributions for the parameters defining each vegetation type, which are stored in the PEcAn database. Parameters are then automatically updated with a hierarchical Bayes meta-analysis of the trait data in the database any time new data is added to the database. These updated distributions are sampled to drive an ensemble of model runs, which can quantify parameter sensitivity and model prediction uncertainty. This approach has been used to better parameterize, analyze, and understand multiple ecosystem models across a wide range of biomes and to direct model-informed field research (6, 12, 13). We have also developed additional modules to deal with the uncertainties in specific data types, such a leaf-level gas exchange (14, 15), tree ring growth records (16), and plant biomass allometries (17).

Finally, one of the key goals of our ABI Innovation award was to develop state-variable data assimilation approaches to synthesize different carbon cycle observations. We are currently able to assimilate data from the MODIS satellite and eddy covariance towers (Viskari et al *in review*) and are presently working on assimilating US Forest Service vegetation inventory data (Viskari et al *in prep*).

Broader Impacts

To date, funding has supported three postdoctoral fellows and has produced eight publications, two manuscripts in review, sixteen presentations, three undergraduate training workshops, and seven training workshops at the graduate/postdoctoral level. Since the initial public release of the PEcAn system in September 2012, we have released ten software versions and the PEcAn virtual machine has been downloaded 340 times. The code repository is publicly available at http://github.com/PecanProject/pecan where it has been forked (created a separate development repository) by 37 developers who have made over 3000 commits (code revisions). In addition, project information, tutorials, videos, and an interactive demo are all available at http://pecanproject.org. As part of this grant, Dietze had developed and taught a graduate course on Ecological Forecasting and Informatics and is writing a book on this topic under contract with Princeton University Press.

PEcAn is being used successfully in a number of other research projects in order to assimilate data into ecosystem models. These projects create substantial synergy that we will be leveraging in the proposed project. These include development of the database and its web interface (Energy Bioscience Institute, **LeBauer**), the support for LiDaR and radar remote sensing (NSF EF #1318164, **Dietze** CoPI), hyperspectral remote sensing (NASA Terrestrial Ecosystems, PI Shawn **Serbin** [former postdoc on the ABI Innovation award], **Dietze** CoPI, **Desai** collaborator), LandSAT, soils, and topography (NSF EF #1241894 **Desai** and **Dietze** CoPIs), meteorological downscaling (NSF EF #1241891, **Dietze** CoPI), and a wide range of uncurated vegetation data sets (NSF DIBBs #1261582, **McHenry** PI, **Dietze** CoPI).

3 Objectives

This proposal has four specific, interconnected objectives:

1) Build Community (Sec 6)
2) Transform PEcAn into a distributed Bayesian machine learning system (Sec 5.1 and 5.3)
3) Enhance tools for multimodel evaluation, synthesis, and prediction (Sec 5.2)
4) Provide an accessible system for real-time synthesis, forecasting, and decision support (Sec 5.4)

The first objective, **Build Community**, is detailed in *Section 6.1* Plan for User Engagement, but briefly we will: (A) Solicit the needs of the community in greater detail to develop a more complete picture of system requirements, use cases, current work patterns, and bottlenecks; (B) Increase the number of models that interface with PEcAn; (C) Interface PEcAn with NSF DataNets and public repositories to allow users to connect to their own and others data; and (D) Build a sustainable 'community' approach to model-data assimilation through outreach and both face-to-face and online training.

One of the challenges in developing a community approach to characterizing and reducing the uncertainties in models is that model-data analyses done by one researcher rarely have a direct impact on other users, especially if one is working with a different model, in which case you have few options other than to repeat the whole analysis from scratch. Accelerating the improvement in models requires that we make these feedbacks more direct and faster. For the second objective, **Transform PEcAn into a distributed Bayesian machine learning system**, we intend to use a distributed architecture to allow instances of PEcAn on different machines to communicate, allowing work by different teams to contribute to the data assimilation and data mining across the whole system.

Building or leveraging tools for communication are necessary, but not sufficient, to accelerate the pace of model improvement. We also need to **enhance tools for multimodel evaluation, synthesis, and prediction.** This will begin with tools for benchmarking models against reference datasets as a way of scoring the performance of different models across difference conditions, and tracking the improvement of models over time. These tools will also allow data-model comparisons and benchmarks run on one model to be automatically replicated with other models. Next, given a distributed database cataloging runs of many models at many sites, we will build modules that leverage data mining and hierarchical Bayesian approaches to characterize and correct model structural errors, process error, and the spatio-temporal variability in model parameters. We will also extend our existing tools for uncertainty analysis to support experimental design and optimize the deployment of limited resources to further reduce uncertainties.

Finally, we aim to transform PEcAn into an **accessible system for real-time synthesis, forecasting, and decision support.** This will require that we implement additional tools for remotely sensed data and other real-time data sources, provide support for data from NSF's National Ecological Observatory Network (NEON) as it comes online (scheduled to be fully operational by 2017), and continue development of data assimilation approaches. As part of this objective we will also build upon our current web interfaces in an effort to make all of our tools more intuitive, more accessible, and deployable on-demand.

4 Example Use Case: Field assessment of carbon stocks & fluxes

To illustrate the application of the PEcAn system, consider the following use case of assessing the carbon stocks and fluxes of a forest. This is a common research and management activity that will become even more common as carbon markets expand and as traditional forester's "timber cruise" is replaced by management for multiple ecosystem services. Components to be added are identified by the section number describing these tools in the Description & Methods, while existing components or components contributed by other projects are labeled PEcAn 1.

We envision the ability to access a PEcAn server from the field using a smartphone's web browser. PEcAn would use the smartphone's geo-location information to automatically extract site-specific information, such as soils and topography, from GIS layers and to automatically downscale meteorological time-series data for the site using a combination of regional-scale gridded weather "reanalysis" product, local weather stations, and topographic corrections **(Sec 5.2)**. PEcAn would then extract data from a variety of satellite and airborne platforms, such as LandSat, MODIS, lidar, radar, and hyperspectral, and combine this with data from national forest inventories in order to estimate the current vegetation type and to estimate the current composition and structure of the forest (**PEcAn 1**). These data would also provide data time-series to constrain the ecosystem model's historical trajectory for land use, management, phenology, composition, and structure (**PEcAn 1 & Sec 5.4**).

PEcAn would next scour known observatory networks (e.g. NEON, FluxNet), data repositories (e.g. Dryad, NSF Data Nets), and its internal database for additional data constraints (**Sec 5.1**). The user could then be queried to ground-truth the PEcAn estimates based on the uncertainties in the data (e.g. the soils map for a region may be particularly coarse) and the known sensitivities of the PEcAn system based on past model runs (**Sec 5.4**).

For the identified vegetation type, PEcAn selects the latest model parameter probability distributions from across the PEcAn network (**Sec 5.3**), which themselves are constructed from the intersecting evidence of all previous users, and launches an ensemble of multiple ecosystem models, to account for uncertainties in model structure (**Sec 5.2**), with each model launching an ensemble of runs to account for uncertainties in model parameters, initial conditions, and drivers (**PEcAn 1**).

As model runs complete, PEcAn would use model-data fusion techniques to filter the ensemble based on the various data constraints previously queried (**PEcAn 1**) and update plots and output as runs complete, instead of waiting for all models to finish (**Sec 5.2**). Time series with confidence intervals of relevant ecosystem pools and fluxes of carbon, water, energy, and nutrients, would be sent to the user's screen, where they could be explored in more detail using interactive visualizations, and archived in the PEcAn system for later access and download (**PEcAn 1**). Additional ensemble runs could also be requested to explore responses to different management options and climate scenarios (**Sec 5.4**).

PEcAn's automated tools for uncertainty analysis (**PEcAn 1**) and observing system design (**Sec 5.4**) can be used to identify the dominant sources of uncertainty in the estimates and to prioritize additional measurements based on available resources. The entry of new data from the field would allow the system to refilter the ensemble output and rapidly update results without having to rerun the full ensemble (**Sec 5.4**). The results of these analyses would propagate across the PEcAn network (**Sec 5.1**), allowing parameter probability distributions to be updated (**Sec 5.3**) and the performance of alternative model structures to be evaluated (**Sec 5.2**), and thus contribute to our understanding of which models perform well under different conditions.

This use case illustrates our vision for the PEcAn network, but it is clearly not the only use case supported by the underlying automated tools for ecoinformatics and model-data fusion that we have been developing. Other use cases would include, but are not limited to: using PEcAn to evaluate past experiments or observation campaigns or to design future experiments and observations; evaluating non-trivial competing hypotheses about ecosystem function; meta-analytical synthesis leveraging mechanistic models rather than simple summary statistics; generating projections under different climate and management scenarios; and making operations, real-time forecasts of ecosystem services.

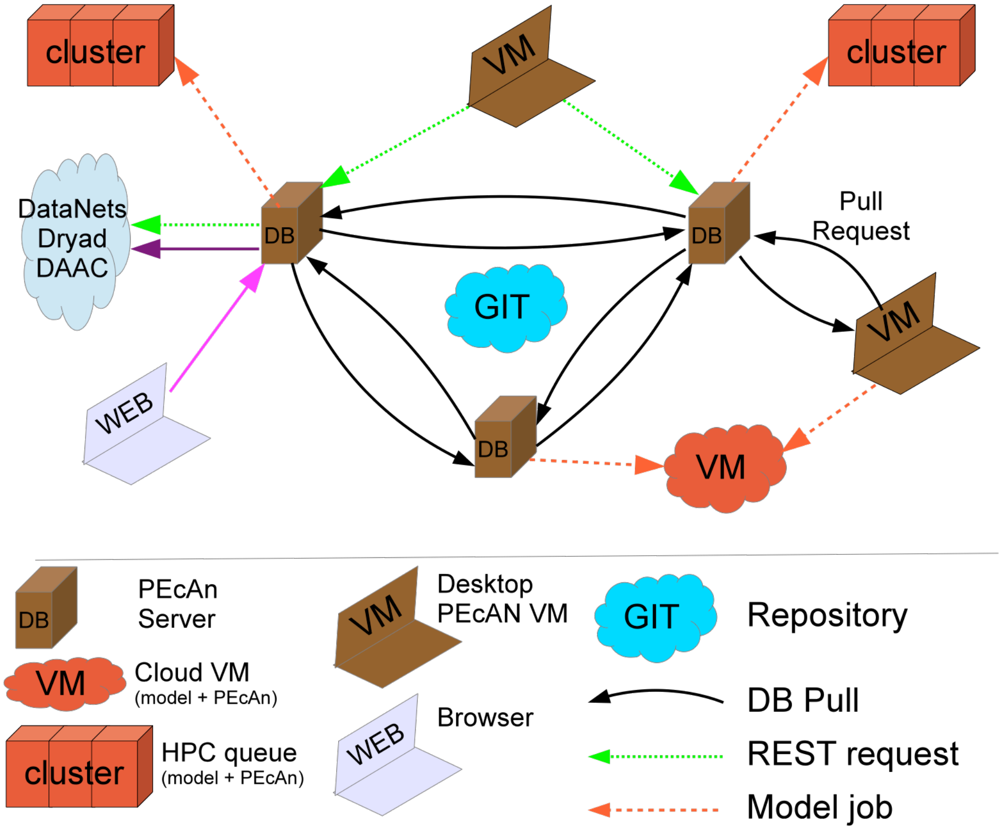


Figure 3: The PEcAn system (database + modules + web interfaces) can be installed on servers or downloaded as a fully-functional virtual machine. This system can be accessed through web-based interfaces or from within R and can run ecosystem models in place or submit these jobs to remote HPC resources (e.g. university clusters, XSEDE) or cloud VMs (e.g. Amazon, Google). PEcAn instances sync databases using a "pull" model, whereby database backups over specific row ranges on a remote machine are pulled into the local database. For persistent servers this may occur automatically on a cron job, while other PEcAn instances, such as desktop VMs, may submit pull requests asking that their databases be pulled into a server. Groups of servers or VMs may deliberately form temporary subnetworks, but under normal conditions information would rapidly percolate through the system. Information in files registered with the database (e.g. model drivers, run outputs, parameter posterior distributions) can be accessed through a REST interface, which would also make PEcAN resources available to other applications, while data sources on the web would be accessed on the fly either though automated downloads or APIs. The PEcAn API would allow whole files to be requested, common extractions (e.g. single output variable or site-level extraction of a regional model driver), or remote calculations (i.e. moving simple code to big data). This allows analyses to be automatically replicated (with the same or different models), multi-model synthesis and analysis to be distributed across the system, and the across-network updating of parameter distributions. The PEcAn codebase is maintained in a public repository (github.com/PecanProject) allowing all users access to the development code and the ability to submit code improvements and new features.

5 Description & Methods

5.1 PEcAn Network

Our first priority will be to address the need for a scalable cyber-infrastructure solution for the bottlenecks associated with working across models and research teams (Figure 3). Where PEcAn v1 was an ecoinformatics **toolbox**, PEcAn v2 will be a distributed peer-to-peer ecoinformatics **network** – a federated resource for the accessing of models, data, runs, and computation. The need for this progression in available tools has already become apparent within our team, where we encountered the need for multiple instances of the PEcAn databases at different institutions and a means for easily sharing input files and model outputs among collaborators.

We have already taken the first step toward this distributed architecture by creating a system that allows database synchronization among institutions. Specifically, we use a distributed "pull" model of synchronization whereby each server pulls the changes from other servers into their local copy of the database. These synchronizations can occur either on demand or as part of a scheduled cron job. Users have complete control over what data they post and who they pull data from. We emphasize the importance of allowing collaborators to have control over what data they share while providing them with the tools to share their data. The main improvement to database syncing in PEcAn 2 will be the **addition of a "pull request"** option, where one PEcAn node (such as a laptop running the PEcAn Virtual Machine) can ask another node (such as the BU PEcAn server) to do a pull. As with modern distributed version control, the recipient node would have the option to review new additions before accepting changes. We envision a situation where there will be a core network of high-volume, trusted nodes that sync automatically, and a broader low-volume community that contributes through pull requests.

The second key component of the PEcAn network infrastructure will be the addition of a Representational State Transfer (REST) API (18) that will allow the exchange of information outside the database. Due to the large file sizes in ecosystem modeling, the PEcAn database currently registers the metadata about these files but not their contents. This information about who has copies of which files percolates across the network through database syncs. The first stage of this API will be to enable the on-demand **sharing of whole files** between PEcAn nodes so that users can easily replicate analyses, extend those analyses to new models, or analyze outputs of model runs done on other servers. The second stage is to add the ability to **request subsets of files**, such as a single location out of a larger gridded dataset. We envision using this to request individual variables from model outputs or request model input data for a specific site. The third stage will be to **request calculations** be performed on remote PEcAn nodes. For example, someone else may have run a model at a site you are interested in, and also have field observations for that site, so instead of downloading both files locally, the requested comparison between the two could be performed on the remote server. This approach is particularly powerful when asking for comparisons for a large number sites or models, an approach often used in the MapReduce (19) paradigm where one moves computations to where the data is stored rather than moving the data to the site of computation. Finally, throughout the development of the PEcAn REST API, we will develop an R package that will provide a front end to these tools. This package will not only be used internally in PEcAn (where most of the modules are written in R) but will also make the contents of the PEcAn network accessible to the broader community without having to install the full PEcAn system.

The third component of PEcAn network infrastructure will be **tools for interacting with other data repositories**. A large number of data sources used by PEcAn are already on the web (e.g. meteorology, remote sensing, forest inventory), but are currently downloaded and inserted into the database manually. Not only will we make accessing these "big" data sources more automated but we will make it easier to search for and use smaller data sources from within PEcAn. In particular, there are a number of public data repository efforts and tools that can be interfaced with such as Dryad (http://datadryad.org/), ORNL DAAC (http://daac.ornl.gov/), DataOne (http://www.dataone.org/), SEAD (http://sead-data.net/), and other cyberinfrastructure projects (e.g. resulting from NSF's Earth Cube initiative) that either store or provide access to a wide range of observations that are directly relevant to the ecosystem models in PEcAn. Making it easier to pull data directly from repositories than to insert that same data directly into PEcAn will encourage users to deposit data in long-term repositories, creating a powerful synergism between these systems. To pull data from repositories, we will integrate tools being developed under the recently funded Brown Dog project (NSF DIBBs #1261582, PI **McHenry**, CoPI **Dietze**) to access and index data collections, in particular uncurated and/or unstructured data collections. We will also explore how to best leverage efforts such as DataONE and tools such as iRODs (http://irods.org/) as a distributed means of sharing files. One particularly important public data source is NSF's National Ecological Observatory Network (NEON, http://neoninc.org/), which is currently in build out phase and will be fully operational by 2017. NEON will provide 539 ecological data products from 60 sites across the US, generating ~10TB/yr of data for the next 30 years.

Demand for computational resources often exceeds what users typically has access to in person. PEcAn 1 already has support for remotely utilizing high performance computing (HPC), such as university clusters and XSEDE resources. As part of PEcAn 2 we will release versions of the PEcAn VM that run on the major **cloud computing** services (e.g. Amazon, Google), providing users with greater flexibility, affordability, and scalability in the types of computational resources they can leverage. Towards this end, we will enhance our current support for remote HPC execution to also support remote cloud execution, possibly within the underlying workflow system and/or leverage Swift (http://swift-lang.org/) to distribute and execute jobs. The work in Section 5.1 will be lead by **McHenry** and **Kooper** at NCSA, who have extensive experience in the technologies described above.

5.2 Enhance tools for multi-model evaluation, synthesis, and prediction

To enable synthesis across multiple models, as required by our use cases, several components of PEcAn 1 will need to be further developed in PEcAn 2. These include, 1) the enhancement of our model input workflows to work with a wider range of forcing observations, 2) incorporation of climate downscaling methodologies for coarse spatio-temporal resolution climate model output for projection, 3) improvements to model visualization and benchmarking, and 4) new modules to analyze model structural uncertainty. The outcome of these tasks will allow PEcAn to be fully realized as a tool for the evaluation of how ecosystem models project carbon cycling and vegetation dynamics in future climates. These tools will also allow us to better understand what features of these models drive differences in projections, facilitating model improvement and uncertainty reduction. These tasks will be jointly tackled as part of a research thesis by a graduate student supervised by PI **Desai**.

Currently, PEcAn 1 already has functioning tools for the **ingest of model forcing datasets**, including the North American Carbon Program site synthesis, tower-based observations at Ameriflux sites, and high-resolution reanalyzes from the NCEP North American Regional Reanalysis. In all cases, current modules can read, extract, and convert these observations to the widely used Unidata Climate Forecast (CF) NetCDF community standard, a convention used for most climate forecast models and some weather model output. Further, simple gap-filling, downscaling, and merging tools are in various stages of development. Here, we propose to 1) expand this set of drivers to include FLUXNET, NEON, and NSF LTER sites with meteorological observations, 2) assess and propagate uncertainty in our gap-filling methods, and 3) incorporate a range of existing statistical downscaling techniques (multiple linear regression, canonical correlation analysis, and Copula functions) (20, 21). These tools will have broad application beyond PEcAn as the need for processing meteorological data is common across ecology and other fields, but surprisingly there is not already a set of general, flexible tools to do this.

These same methods can be then **applied to climate model projections** available from the CMIP5 database (https://www.earthsystemgrid.org). A challenge for ecosystem models is that most existing downscaled climate model observations are: 1) not congruent with the often lengthy time period (10s-100s of years) over which ecosystem models require integration to adequately capture diurnal to successional scale dynamics; 2) with insufficient temporal resolution, usually daily or monthly, whereas models often require sub-hourly; 3) over a limited number of observation types (e.g., temperature, precipitation), whereas models require fuller accounting of the thermodynamic and radiative state of the lower atmosphere; and 4) over subsets of the region of interest. We will apply the techniques developed in the driver extraction routes and existing published scaling methods (21) to develop continuous climate projection variables that can be used by all models. In particular, de-biasing routines will be incorporated here to account for known model biases in simulation of precipitation, temperature, and radiation.

On the output side, PEcAn 1 produces plots and outputs for one model at a time, therefore we need to develop **visualizations and benchmarks for the comparison of multiple models**. These plots and benchmarks will be updated continuously as new model runs are completed locally or retrieved from the PEcAn network. Similar to the plots shown in Figure 1 & 2, interactive plots will be added allowing for overlay of a single variable (for instance, net biospheric carbon flux) over time or space with outputs from multiple models. Further, computation of the "ensemble" model (the average across all other models) will be added as an additional "virtual model" output. In PEcAn 1, we are developing the basic framework for automated computation of model benchmarks, which is scheduled for completion Fall 2014. Here, we will improve this feature into a generic benchmarking tool that provides a set of comparisons (e.g., net change in carbon stock from 1900-2000), a set of references (e.g., other models, known observational constraints), and a set of statistical evaluators (RMSE, bias, etc.). We will follow many of the practices already implemented for the ILAMB model benchmarking framework (22) and the MsTMIP model intercomparison (23, 24). Both the full set of individual benchmarks and a variety of scoring summaries will be made available in PEcAn 2. We will also develop the process to allow a new model added in PEcAn to replicate runs done by another model stored in the database, allowing for continual model intercomparison and testing of how model changes improve benchmarks.

With the above three components, multi-model synthesis and forecasting can be enabled. We plan to demonstrate this with a series of experiments at a range sites from the global FLUXNET network of 683 eddy-covariance towers, which has measured over 5000 site-years of carbon, water, and energy fluxes between the land surface and the atmosphere (25). With a considerably enlarged community of models, sites, and model-data comparisons, we will have a much greater opportunity to identify when, where, and why models are wrong or observations insufficient. To identify these errors, we will **develop new PEcAn modules to analyze model structural uncertainty**. One set of tools will focus on applying data mining techniques to the *model residual errors*, such as the wavelet approaches we previously employed in the North American Carbon Program (NACP) multi-site multi-model synthesis (26, 27). Other techniques that will be explored included cluster analyses and classification trees, general additive models, random forest models, artificial neural networks, and support vector machines. In all cases, we will make extensive use of existing R packages. The primary goal of these analyses is to identify the time scales, environmental conditions, and structural assumptions associated with high residual variance or significant model biases. We aim to provide the user with maximal flexibility in designing and conducting analyses across multiple models and scenarios. In this fashion, PEcAn will allow for rapid testing of differing model assumptions and thus lead to model improvements. In addition, we will explore whether these statistical approaches can be combined usefully with mechanistic ecosystem models when making predictions, for example by correcting biases or propagating inhomogeneous residual errors.

5.3 Transform PEcAn into a distributed Bayesian machine learning system [DESCOPED]

One of the key bottlenecks in ecosystem modeling is that what we learn from one study often does not translate into improvements elsewhere. In the previous sections we described how we were addressing this by facilitating the exchange of information among models, making replication automated, and diagnosing structural errors in models. The remaining major challenge is constraining model parameters with data. Traditionally these models had fixed parameter values and when running a model at a new site one either used these default parameters or fit them for the specific site. The result of this is that most ecosystem model perform well when tuned to a single site, but fail when applied across a wide range of independent sites (28).

In PEcAn 1 we began to address this problem by treating parameters as uncertain, with this uncertainty captured by probability distributions. Furthermore, these probability distributions are stored in the database and updated automatically when new plant trait information is added to the system. In addition, when the output of a model is compared to observations at a site, the error statistics and log

Likelihood are also stored in the database. By running the model across a range of different parameter values in our automated sensitivity analysis, we are able to generate a Likelihood surface from these records. Using Bayes rule we can then update our model's parameter distributions, as the updated distribution is proportional to the original distribution times the likelihood.

Two major advantages of this Bayesian approach are that it can be applied iteratively (i.e. we can keep updating the parameter distributions with new Likelihoods as more model-data comparisons become available) and that we get the same result regardless of the order the Likelihoods are incorporated. PEcAn 1 does not yet fully take advantage of these properties so our first task in this area is to **expand the existing parameter data assimilation** and incorporate it into PEcAn workflows and web interfaces. As part of this we will add a wide range of web-based visualizations for model-data comparisons.

Second, the modules need to be modified to **accommodate multiple rounds of updating** in a scalable manner. In PEcAn 1 parameter estimation is done using standard Bayesian Markov Chain Monte Carlo algorithms, with the one difference being that the sampling is done on an interpolated Likelihood surface rather than by running the model iteratively. This approach, called emulation, is used because the computational expense of most ecosystem models is prohibitive, both in terms of the number of runs involved and the sequential approach of traditional MCMC. By contrast, emulation is easily parallelizable because all model runs can be done simultaneously on different nodes and the Likelihood surface is then approximated with a statistical interpolator. In PEcAn 2, to allow the updating to be sequential, we will still use the Likelihood surface emulation but will switch to a particle filter (a.k.a. sequential Monte Carlo) algorithm for updating.

While the first stage of sequential updating is to perform such updates within a single system, the next stage is to **distribute this updating asynchronously across the PEcAn network**, based on the peer-to-peer interface described in 5.1. PEcAn database syncs will inform servers about new model-data comparisons that have been performed on remote machines. The database syncs will include the log Likelihood values for those comparisons, allowing the local node to reconstruct the Likelihood surface, or it could pull the full Likelihood surface from the remote servers using the API. Importantly, the local server does not need access to either the observations or the model output to perform the update, just the Likelihood surface. The net result of this is that **anytime anyone in the network compares a model to data, that additional learning then percolates across the network to all users**. As with observations and model outputs, each user has control over what information they release across the network and what information they pull from other users.

As with the structural uncertainties, we will develop additional tools to explore the variability in model parameters. Specifically, within ecological systems, there is substantial inherent stochasticity in ecological processes as well as environmental and biotic heterogeneities that are not fully accounted by current models. These sources of variability are one reason why tuning a model at different sites will give somewhat different parameter estimates. Current approaches to model-data fusion do not account for this variability, but the use of hierarchical Bayes to capture and partition these sources of ecological variability is well established and has generated substantial theoretical insight (29–31). By **incorporating a hierarchical structure** to our distributed Bayesian learning system, we will be able to ascribe the variability in ecological processes to different fixed and random effects, such as interannual and spatial variability, a key component of identifying model improvements. Overall, the research in this Section will be part of a doctoral thesis by a graduate student supervised by PI **Dietze**.

5.4 Accessible real-time synthesis, forecasting, and decision support

The final objective in our project focuses on the integration of the tools developed above and in PEcAn 1 in order to **ensure that all parts of the system are polished, interoperable, and that their use is intuitive and accessible** without being dumbed down. **All members of the team** will contribute to these goals, with **NCSA** taking the lead on integration and the **project manager** focused on accessibility. To date the interoperability of tools has been high since our project holds a weekly videoconference where we have put considerable attention into the design of the whole system and the modules before components are implemented.

In addition to overall integration, there are a number of additional tasks required to support our use cases. First, we need to integrate the existing uncertainty and power analyses in PEcAn 1 with the Bayesian learning in PEcAn 2 in order to develop **tools for observational design** (what additional data should be collected, where, and how much). These tools will then be expanded to the **interactive ground truthing** envisioned in the use case. Second, in order to support **real-time multi-model forecasting**, we need to continue to refine the state-variable data assimilation tools that were a focus of our ABI Innovation project and integrate these with the multimodel support and tools for pulling observations from public data repositories in PEcAn 2. Third, to achieve real-time multi-model **decision support** we will need to ensure that the tools for multi-model forecasting include the capacity to access commonly used scenarios (e.g. IPCC climate change projections) and also allow the user to develop and enter their own scenarios, such as scenarios of land-use change, land management (e.g. agriculture, forestry), pollutants, and disturbance. Coupling the ability of users to entertain alternative “what if” scenarios with PEcAn’s sophisticated tools for quantifying and propagating a wide range of uncertainties, both within and across models, would allow a wide range of management and policy decisions at all scales (from local to global) to be informed by the best available data. Currently, such synthesis and decision support in only available at the largest scales (e.g. IPCC climate change reports), with considerable effort, and with only a fraction of the data constraints that PEcAn 2 will provide.

6 Building Community

Building the PEcAn user community is one of our primary objectives in this project. While all members of the PEcAn team will participate in building community, we will also be hiring a dedicated **project manager** whose primary responsibilities are user engagement, dissemination, and project sustainability.

6.1 Plan for User Engagement

Coupling PEcAn to additional ecosystem models

In the process of previous modeling research activities and intercomparisons, PIs Dietze and Desai have published with 42 other ecosystem modeling teams beyond the four teams already using PEcAn (ED2, SIPNET, BioGro, DALEC). These models range in complexity from simple models that can be written in a single page of code to sophisticated Earth System models with hundreds of thousands of lines of code that are used in the IPCC climate change assessments. As we seek to reach out to the larger modeling community, we have identified 11 collaborators with whom we will work closely in the first two years in order to couple PEcAn to 16 additional well-established, publicly available, highly-used ecosystem models, many of which form the base land surface models of major climate models (Table 1). While the project collaborators are not directly supported (see letters of collaboration), travel funds have been budgeted to allow the core team to spend time on-site at each institution to train collaborators intensively in PEcAn and to work hand-in-hand on coupling each model to PEcAn.

Table 1. Ecosystem modeling collaborators (see accompany letters)

Name	Location	Models
Ian Baker	Colorado State	SiB4
Rosie Fisher	NCAR	CLM, CLM-ED
Ryan Kelly	Boston Univ	TEM
Belinda Medlyn	Macquarie Univ	G’DAY, Maestra, Maespa
Bill Parton	Colorado State	DayCENT
Ben Poulter	Montana State	Orchidee, LPJ
Afshin Pourmokhtarian	Boston Univ	PNET-BGC
Tristan Quaife	Univ. of Reading	SDGVM, JULES
Kevin Schaefer	NSIDC	SiB-CASA
Elena Shevliakova	Princeton	LM3, LM3-PPA

Jonathan Thompson	Harvard Forest	LANDIS-II

Training materials & workshops

The training materials developed for these model coupling exchanges will be implemented online and will build upon previous training materials used to date. The experience derived from working with these groups will allow us to refine and further develop these materials, and will include written material, images, and short videos. We will host these materials on our project website and will investigate other online course platforms for disseminating this material more broadly, for example as a set of undergraduate or graduate level labs. As we will discuss in *Section 6.2* Dissemination Plan, in addition to making these materials available online, we will use them in the following face-to-face training activities: (1) stand-alone PEcAn training workshops, (2) PEcAn workshops at the annual user group meetings for specific model communities, (3) PEcAn tutorials as part of other short courses, and (4) PEcAn labs within semester-long undergraduate and graduate courses. In these activities, we will also survey the experiences of users in order to gain additional feedback about the usefulness and accessibility of these tools and identify bottlenecks. Finally, tutorials on how to submit bug reports and new code (pull requests) on GitHub are already part of our documentation and will be included in online videos and all face-to-face training.

Kick-off Community Meeting

At the start of the project, we will host a meeting in Boston to engage the research community in the design of PEcAn. This first meeting will primarily target the community of model developers and model users. A key goal of this meeting is the identification and prioritization of current modeling and model-data synthesis bottlenecks and redundancies across groups. In addition, we will solicit existing tools and models that could be either coupled to PEcAn or developed into PEcAn modules. The primary deliverable from this meeting will be a manuscript on the bottlenecks facing the current generation of models and the roadmap for reducing them.

Mid-project Community Meeting

Two years into the project, we will host a second meeting in Boston which will engage both the modeling and field research user communities. The goals of this meeting are: (1) to solicit additional input from outside the model development community about their modeling and synthesis needs and bottlenecks; (2) to showcase work done to date; (3) solicit feedback from both model developers and non-modelers about the successes and failures of the PEcAn system; (4) identify any additional data sources or models that need to be coupled with PEcAn or any modules that need to be developed; and (5) engage the broader community in brainstorming additional applications and use cases. We are hosting this meeting mid-term so that many of the new tools will be operational, and the core group of new models fully coupled, but still leave adequate time for needed additions and subtractions to the development plan. The primary deliverable from this meeting will be a manuscript presenting new and innovative use cases for how ecosystem models could be used in the near future to meet research and management needs.

6.2 Dissemination Plan

PEcAn is available as a public code-base (NSCA open source license, http://opensource.org/licenses/NCSA) on GitHub (http://github.com/PecanProject) and downloadable database images are refreshed hourly. In addition, stable releases are tagged in the repository and are made available as a fully-functioning Virtual Machine that has the compiled code for PEcAn and the supported ecosystem models, the database, and the webserver to support all interfaces, with all dependencies installed. We have been working toward the goal of more frequent version releases, with stable updates being pushed on a monthly basis. New release announcements and other high-level updates are posted on the project webpage, developer's listserv, and the project's Twitter account, @PEcAnProject. Task deadlines are outlined in *Section 7* Workplan.

We operate our GitHub repository under the model where we accept pull requests from the general user community and provide documentation for how to submit code. In the proposed project, we will also allow user additions to the database, which will similarly follow a pull model. Model inputs and outputs are currently available through the PEcAn web interface and here we will be adding the ability to access these through the REST API (see section 5.1) and an R package that utilizes the API. All system contents from the core team will be flagged public, while external contributors will have the ability to control whether their data and model runs are made public, restricted to specific users, or kept private.

Throughout the project we are strongly committed to making our tools accessible to the broader biological community and beyond. We will continue to invest considerable time and effort into designing intuitive and easy-to-use web-based interfaces for our tools and formally soliciting feedback on their usability in post-workshop surveys. Our tools are extensively documented on our GitHub wiki, which allows easy, collaborative updating by all users, and all the R packages that make up PEcAn are internally documented using ROxygen to build R package documentation. In the past we have developed written tutorials and we have an hour-long video lecture about the project on YouTube that has had 145 views. Moving forward we will create a more extensive set of short video tutorials and written instructions explaining different tasks and modules.

We will continue to teach PEcAn as part of other workshops (Flux summer course, PalEON summer course, ED2 users workshop, Harvard Forest REU) and courses (Boston University GE375 Environmental Modeling & GE585 Ecological Forecasting) and will branch out to provide both stand-alone PEcAn workshops and training that targets existing user communities for different models. PEcAn workshops will be offered in conjunction with national conferences most frequented by our user community (Ecological Society of America, American Geophysical Union, North American Carbon Program, LTER All Scientists Meeting, International Association for Landscape Ecology). For the user community workshops, we will take advantage of the fact that the larger modeling communities (e.g. CLM, LANDIS) frequently have annual meetings for their user communities.

6.3 Sustainability Plan

We anticipate that the sustainability of the PEcAn project will occur at two scales: (A) within the core team and (B) across the PEcAn network. Within the current core team, PEcAn will continue to advance through the inclusion of PEcAn in new research projects. As noted in Section 2 (Results from previous research: PEcAn v1), PEcAn is being used in seven other research grants, each of which has contributed specific modules. The core team is dedicated to the long term support of PEcAn, as these tools have already become essential to our day-to-day research, and thus will continue to include this type of support in future research grants. Furthermore, we will develop a project governance document, which will lay out the protocols for public development and decision making in a manner that establishes PEcAn as a true community resource.

The second scale of sustainability will occur organically across the PEcAn network. By building a distributed network, the network gains resilience. For example, PEcAn's reliability does not depend upon any individual server because the database is mirrored across all the servers in the network. We aim to build a critical mass of users who find PEcAn indispensible, and thus will likewise contribute code and include further PEcAn development in future grant proposals. The ecosystem modeling community is already skilled in programming and software development, and PEcAn provides tools that would be used on a regular basis for everyday modeling activities, so we feel that it is realistic to anticipate that this community would contribute to the further maintenance and development of these tools. We also anticipate other future funding resources beyond this award to support development.

7 Workplan

The Work Breakdown Structure/Timeline (Table 2) summarizes the major tasks laid out in *Section 5 (*Description & Methods) and *Section 6 (*Building Community) and the schedule for their completion. In particular, automating the pulling of data from repositories will be done on an as-needed basis to support the other objectives in 2015 and 2016 leading up to the core effort starting in 2017 with support for NEON data. Likewise, efforts to couple additional models to PEcAn will continue after the core effort in the first two years, while the development of training materials and workshops will occur on a continuous basis throughout the project.

Table 2: Work Breakdown Structure & Timeline

			2015			2016				2017				2018				19
	Objective	Task	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
NCSA	PEcAn Network	API File request	■	■							W	W						
		DB pull request		■	■						▨	▨						
		API File subset				■	■				▨	▨						
		API Remote Op						■	■	■	▨	▨						
		API R package	■	■	■	■	■	■	■	■	▨	▨						
		Pull remote data	▨	▨	▨	▨	▨	▨	▨	■ NEON	■	■	■	■	■	■	W	W
		Cloud execution															■	■
Wisc	Multi-model Synthesis	Enhance inputs	■	■	■				W	W								
		Climate downscaling				■	■	■	▨	▨								
		Benchmark & Viz							■	■	■	■	W	W				
		Structure data mining										■	■	■	■	■	W	W
BU	Distributed Bayesian Learning	Refactor & Viz	■	■	■													
		Local updating			■	■	■	■	W	W								
		Distributed updating						■	■	■	■	W	W					
		Hierarchical									■	■	■	■	■	■	W	W
All	Accessible Synthesis, Forecast, & Decision Support	Accessible tools	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
		Observational design							■	■	■	■			W	W		
		Ground Truth Queries									■	■	■	■	▨	▨		
		Real Time Forecast											■	■	■	■	W	W
		Decision Support													■	■	■	■ W
Manager	Build Community	Meetings	■	W	W						■	W	W					
		Couple models	■	■	■	■	■	■	■	■	W	W	▨	▨	▨	▨	▨	▨
		Training materials	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
		Workshops	▨	■	■ C	▨	▨	■	▨	▨	▨	■	■ C	▨	▨	■	▨	▨

*The core time for each task is indicated in black, while grey areas indicate areas that will receive continued, secondary attention. Areas marked with a W indicate time devoted to the preparation and submission of at least fourteen peer reviewed manuscripts Areas marked with C indicate the teaching of GE585 Ecological Forecasting at BU, a course developed as part of the Broader Impacts.

Our team supported here includes 3 PIs spanning the biological, atmospheric, and computer sciences, two graduate students, one project manager, one undergraduate, 11 ecosystem model collaborators and 10 existing collaborators, postdocs, and graduate students supported on other projects. In terms of allocation of effort to tasks, NCSA (PI **McHenry**, Senior Programmer **Kooper**) will lead the development of the cyberinfrastructure to support the functioning of PEcAn as an integrated peer-to-peer network and will also coordinate code integration across all objectives. They will be assisted by collaborator **LeBauer**, who currently maintains the database front-end, and a summer undergraduate technician. At Wisconsin, PI **Desai** and **graduate student** will lead the development of support for multi-model analysis, benchmarking, and visualization, as well as tools for the diagnosing structural uncertainties across models. At Boston University, PI **Dietze** and a **graduate student** will lead the enhancement of the current parameter data assimilation approach into a distributed Bayesian learning system. In addition, Dietze and the **project manager** will lead community building efforts, such as community meetings, training materials, videos, and workshops. They will also couple PEcAn to additional ecosystem models with the assistance of the 11 **ecosystem modeling collaborators**. The project manager will furthermore coordinate efforts to ensure that all PEcAn tools are made as intuitively accessible as possible. Efforts to make PEcAn into an on-demand real-time system for synthesis, forecasting, and decision support requires a close integration across the tools developed in all the other objectives, and thus the workload will be shared by all three institutions. All aspects of the project will be discussed at weekly teleconferences and managed using online task scheduling and issue tracking software, as we have been doing with PEcAn 1.

8 Broader Impacts

The PEcAn network will strengthen the scientific foundation for climate and environmental science, as well as climate and environmental impact policy, by improving the reproducibility and provenance tracking of ecosystem data-model synthesis and model projections. PEcAn will provide a set of open-source tools that will make ecological modeling, synthesis, forecasting, and decision support accessible to the broader research community and beyond. It will establish a resilient ecoinformatics network that will greatly enhance the capacity for peer-to-peer modeling, data sharing, and providing easier access to high performance computing resources. PEcAn will engage the broader community of modelers and non-modelers, increase trading of ideas, structures, parameters, and make models more robust and less subject to criticism of cherry-picking results. Many of the tools developed are extensible beyond ecosystem modeling and field biology communities to weather forecasting, climate science, biodiversity and other models.

As we describe in our use case, PEcAn aims to make ecosystem modeling and ecological forecasting part of the everyday tools of not just the research community, but also land managers and on-the-ground operational personnel. Indeed, one of the models that we are coupling with, LANDIS-2, is already a required part of US Forest Service management plans. Furthermore, the models in PEcAn are not limited to just US forests, but include every biome worldwide, including major agricultural systems, and the PEcAn system already makes it straightforward to add new vegetation types or agricultural crops. As such, PEcAn has the potential to find broad application in precision agriculture, forest management, rangeland management, and conservation planning. Because some of the models even include urban ecosystems, PEcAn could also be used by city or state planners aiming to balance the management of multiple ecosystem services and anticipate responses to future climate scenarios.

As discussed in *Section 6* Building Community, we are putting considerable effort into making these tools public. This includes making the code and databases open source and modifiable, hosting community meetings to solicit input, working with modeling teams to couple more models to PEcAn, and providing extensive documentation, training materials, and online videos. In addition, we will run PEcAn training workshops at national conferences, at the annual meetings of modeling teams, and as part of ongoing summer courses such as the Niwot Ridge Flux Course. We will also be developing training materials and workshops specifically for the policy and management communities. In addition, PEcAn will continue to be used in the classroom as part of Dietze's Environmental Modeling (GE375, undergrad) and Ecological Forecasting (GE585, graduate) courses, as well as the Harvard Forest REU program.

Dietze's graduate class in Ecological Forecasting is a direct result of the Broader Impacts of our ABI Innovation award, and will continue to be taught Fall 2015, 2017, and odd years thereafter. In addition, Dietze is under contract with Princeton University Press to produce a book on the same topic, to be submitted Jan 2016. This book will be the first on the topic and provides a novel synthesis of the theories, concepts, and past research in ecological forecasting. The book is very complementary to PEcAn itself, but also discusses forecasting in many domains other than just the carbon cycle, and thus provides an important bridge to generalizing our work to other disciplines, such as environmental epidemiology, endangered species, invasive species, fisheries, agriculture, and forestry.

Finally, over the course of developing PEcAn 2 we will educate two graduate students, one undergraduate per summer, and one post-bachelor's project manager, focusing on cross-disciplinary training and recruiting of under-represented students in our field. We also plan to publish over a dozen peer-reviewed manuscripts, and present our research at national and international meetings.