Background

Ecology needs data-intensive approaches to predict the responses of biodiversity and ecosystem function to global change, address invasive species, and prioritize areas for conservation. However, the challenges common in many areas of data science are often magnified in ecology. These include highly heterogeneous data across systems and studies, an unprecedented volume and velocity of data that is now being collected using sensor networks and remote sensing, and the lack of ecologists with the necessary training to deal with this magnitude of data. Despite a clear need, the discipline has failed to embrace data-intensive approaches (see e.g., Paine 2010, Lindenmayer & Likens 2011, 2013). Ecology needs leaders in data science to: 1) demonstrate the value of data-intensive approaches in ecology; 2) mine, assemble, and model existing data; 3) integrate existing data with the large amounts of data now being generated by automated data collection and citizen science; and 4) train ecologists in the tools and approaches necessary for data-intensive science.

Major Accomplishments

I entered graduate school with a field ecology background, but became frustrated with the scope of inference that could be accomplished using typical ecological approaches. It was possible to understand a single system well, but these results could not be generalized across the globe and across the diversity of life. I joined one of the only labs doing data-intensive ecology and have spent my career using large data compilations to understand ecological systems at continental to global scales, developing tools to make it faster and easier to conduct this type of research, and training the next generation of data-intensive ecologists.

Research: My research uses large data compilations to understand the processes driving ecosystems, test ecological models, and make predictions regarding ecological systems. We have developed novel approaches to understanding patterns of biodiversity based on dividing species into resident and transient species using the structure of their population time-series (White & Hurlbert 2010, Coyle et al. 2013). We have led the way in strong general tests of ecological theory by creating the largest combined datasets ever used in community ecology and using them to evaluate and compare ecological theories (White et al. 2006, Thibault et al. 2011, White et al. 2012, Locey & White 2013, McGlinn et al. 2013, Xiao et al. 2013). Finally, we have built on these ecological theories by using them to make predictions for biodiversity, rarity, and patterns related to ecosystem level fluxes (White et al. 2012, Xiao et al. 2013). We are beginning to combine this with new work using machine learning and remotely sensed data in multi-level models to forecast the future state of ecological systems and make predictions for unsurveyed locations.

Software Development: Conducting this type of research requires the time-consuming assembly of numerous heterogeneous datasets. Typically every scientist writes custom code to handle the data or assembles them by hand. This wastes time and effort organizing data that could be spent doing science. My group built the EcoData Retriever to address this problem. It automates the discovery, downloading, cleaning, and reformatting of ecological data. This simplifies the process of finding, acquiring, and using data and therefore increases the use of the diversity of ecological datasets. Altmetrics show that the Retriever is both highly recommended and highly cited (stars and forks on GitHub in the 91-99%tile), the associated paper is in the 98th percentile of online impact for scientific papers, and the installers were downloaded over 230 times in the last two weeks.

Training & Mentoring: I actively train and mentor the next generation of data-intensive biologists. Using an NSF CAREER Award I have developed a suite of university courses that teach core aspects of data science to biologists (programming, databases, statistics, visualization). Course material is available online and has been viewed over 125,000 times by users in over 150 countries. Students and postdocs in my lab with backgrounds including field biology, mathematics, statistics, and computer science receive in-depth training in data-intensive approaches to ecology. I am also actively involved in Software Carpentry: developing material, teaching ecology focused workshops (6 in the last 2 years), and serving on the advisory board. Finally, I have taken a leadership role in making biology more open, reproducible, and data-intensive by: 1) openly sharing reproducible code, data, and grant proposals; 2) writing papers on data management and sharing, best practices in computational science, and the need for preprints in biology; and 3) explaining the importance of these approaches on my blog and through my active presence on Twitter.

Future Research Directions

My future research directions involve working in three key areas that will create a foundation for data-intensive ecology: 1) tying data and theory together more tightly by expanding the use of data-driven modeling in ecology (e.g., machine learning, hierarchical modeling); 2) leveraging existing data to test, improve, and make predictions using process based models; and 3) developing tools and personnel to help scientists handle the challenges of ecological data.

My data-driven modeling efforts have so far focused on entropy maximization models, which predict the state of a system based on a set of empirical constraints. We will build on this foundation by expanding more generally into machine learning and hierarchical modeling to make predictions for ecological patterns and processes across scales. We will develop a suite of master models to address core questions about biodiversity, population dynamics, and ecosystem processes, including biosphere influences on carbon fluxes and global change. Specifically, we will develop models for the distribution of individuals of species across the globe, the traits of those individuals, and the dynamics of both counts and traits over time. These models will use large compilations of climate, land use, and ecological data as predictors, including data on the abundances of other species and their potential interactions both within and across taxonomic groups (e.g., birds compete with other birds for food and require plants for food and nesting). These models will be trained and tested using compilations of ecological data from across ecosystems and taxonomic groups, including an expanded version of our current compilation (which currently includes data on the distribution of ~50 million individual organisms), text mining data on interactions among species, and growing compilations of individual level trait data (~2.5 million trait values).

I will leverage the large amounts of data currently being generated in ecology to provide improved testing of process based ecological theories. We will conduct strong tests and comparisons of population, community, and ecosystem level models by evaluating their performance across ecosystems and taxonomic groups, testing all of their predictions simultaneously using independent data, and directly comparing predictions of different models for the same patterns. Specifically, we will focus on models relating population dynamics to environmental drivers, unified community theories making predictions for large numbers of ecological patterns, and ecosystem models that relate properties of the organisms at a site to the system level fluxes. The information from these comparisons will be used to identify the most promising models for making broad scale ecological predictions, to identify aspects of these models that require improvement, and to change or constrain these models to yield improved predictions for ecological systems.

To help ecologists handle the challenges of data-intensive science I will continue to expand my efforts to make data-intensive ecology easier, more robust, and more reproducible. I plan to add provenance tracking and automated reproducibilty features to the EcoData Retriever that will allow data processing steps to be documented and easily reproduced. I will develop new tools to make it easier to combine ecological, environmental, taxonomic, and other datasets in reproducible ways to allow ecologists to quickly make use of the broad array of data that relates to ecological systems. To address the lack of training for scientists in data-intensive approaches I will participate in the development of a new set of Data Carpentry material that builds on Software Carpentry's knowledge and successes to train the next generation of scientists. In combination, this research, tool building, and training will help establish a data-intensive era for ecology.

1. Fundamental Question and Five-Year Impact

1.1 Impact

Despite the critical need to forecast how nature will respond to global change, ecology is still primarily a descriptive science focused on understanding, rather than predicting, nature. Little research generates meaningful predictions and, when predictions are attempted, they tend to focus on a single ecosystem. This lack of prediction is slowing progress in ecology (McGill 2012a), because the field lacks meaningful benchmarks for comparing and improving models and fails to develop models that can be applied to understudied ecosystems and species. Despite having recognized the importance of forecasting for over a decade (Clark et al. 2001), ecology has failed to embrace a predictive approach to science.

Now is the perfect time to lead ecology from a descriptive to a predictive science. We now possess the abundance of data necessary to tackle prediction, the methods to develop predictive models from these data (e.g. machine learning, hierarchical modeling), and the mathematical and computational technology to develop and assess models from across the data-driven to theory-driven spectrum (Luo et al. 2011). I will develop and test predictive models for ecology across different levels of organization, ecosystems, and the diversity of life. I will build tools to make data-intensive science easier and train the next generation of scientists in data-intensive approaches. All of this work will be conducted in an open, collaborative, and reproducible manner. This will broaden its impact by increasing knowledge of the approaches and building a community to further the development of the tools. In combination, **the impact of my work will be to make ecology a more predictive, data-intensive, and open science that is capable of addressing the major ecological and environmental challenges of our time.**

1.2 Fundamental Question

I will focus on one of the fundamental questions in ecology: **what will nature look like in the future?** This question is fundamental to knowing how well we understand ecological systems because, if we cannot predict how they will change, we do not understand how they operate. As such, answering this fundamental question requires understanding what governs the structure, dynamics, and fluxes of ecological systems. This question is also fundamental to the application of ecology in management and policy decisions. To protect at-risk species, conserve biodiversity, and maintain ecosystem services, we need to be able to predict how nature responds to stresses such as climate and land-use change, and potential interventions such as the creation of reserves and the removal of invasive species (Clark et al. 2001, Evans et al. 2013).

I will make forecasts for three major areas of ecology: 1) the abundance of individual species; 2) the structure of communities (groups of species in the same region); and 3) the function of entire ecosystems. Populations, communities, and ecosystems are three of the central levels of organization in ecology. They are all amenable to data-intensive approaches due to large amounts of existing data (from individual studies, citizen science projects, and coordinated government sampling) and to a massive influx of new data from ecological observatory networks. Focusing on these dimensions of ecology will allow us to predict how species distributions, biodiversity, rarity, ecosystem fluxes, and ecosystem services will respond to anthropogenic pressures including shifts in climate, changes in land-use, and invasive species.

I will seek to capture how these different aspects of ecology change through time and vary across ecosystems and the diversity of life. I will develop predictive models using both data-driven and theory-driven approaches, using large compilations of ecological data to evaluate and improve the models. I will use far more comprehensive and sophisticated data compilations than have been brought to bear on these questions before, including data on hundreds of millions of species' occurrences, high temporal resolution climate and land-use data, and newly available text-mining and compilation-based data on species traits and interactions.

1.3 Measuring Progress

The core measure of progress is how well my research group can predict independent data and how effectively we can forecast the future state of ecological systems. We will measure the progress of our research at three levels based on our ability to: 1) predict the state of ecological systems in different locations; 2) forecast and hindcast within existing time-series when training models only on data from the beginning or end of the time-series; and 3) forecast the future state of ecological systems. Each year we will publish predictions for the state of ecological systems one to ten years into the future and evaluate those predictions every year as new data is collected.

Beyond my own research progress, a fuller measure of my impact will be whether ecology as a field focuses more on prediction and data-intensive approaches. I hope to create an environment that fosters forecasting by explicitly publishing forecasts and evaluating their accuracy. This will help create a culture focused on producing better forecasts, like that in disciplines with successful forecasting, such as weather and climate (McGill 2012b; see e.g. Kalnay 2003). To accelerate this transition I will conduct this research in a fully open manner, using open notebooks, public code repositories, and social media as outreach for the ideas and to encourage collaboration. All code will be open source, all training material will be open access, and all papers will be open access and posted as preprints prior to submission.

I will measure the influence of these efforts, and further the outreach, by running a series of "forecasting challenges" (similar to Kaggle competitions but with a focus on ecological forecasting). Impact on the field will be measured using the level of participation as an indication of how interesting and popular these approaches are among ecologists, and using the collective performance of the competitors to measure our success as a field in using data-intensive approaches to predict the future state of ecological systems.

2. Advancing Data Science Methodologies and Human Capital

We have too much data and too many important problems to be addressed by the small number of individuals with the requisite skills to work with large amounts of heterogeneous data. Realizing the potential of data-intensive approaches requires us to both bring the data to the researchers by developing improved tools for the acquisition, assembly, and analysis of data, and bring the researchers to the data by providing training in computational, statistical and other data science methodologies. Over the last five years I have been actively building these bridges between researchers and data as part of an NSF CAREER award and I plan to significantly expand these efforts by: 1) developing methodologies for working with the variety dimension of big data by building software that automates the acquisition and assembly of heterogeneous data sources; 2) developing approaches for modeling complex data and making them available in easy-to-use software; and 3) training scientists in data science skills. To maximize the impact of these efforts, all tools and training material will be developed in public GitHub repositories using open source and open access licenses. I will use this openness to actively encourage collaboration from both scientists and members of the technology community. In combination, these efforts will allow more scientists to engage in data-intensive approaches, and will let them spend more time focusing on doing science and less time wrestling with data.

2.1 Methodologies for Automatically Combining Heterogeneous Datasets

Combining heterogeneous data from disparate sources and formats is a core challenge in many areas of data science, and one that is particularly prevalent in my research. Typically this involves individual researchers developing custom scripts to download, cleanup, and restructure individual datasets, followed by even more custom scripts for combining datasets. This is error prone, time consuming, and does not allow scientists to benefit from each other's knowledge and effort. We can do better. By building tools to automatically handle the data side of data science we can remove impediments to data-intensive approaches and allow scientists to focus on doing science.

My lab developed a platform for acquiring, cleaning, and restructuring heterogeneous data sources in reproducible ways, and is building a community who are adding and improving datasets (Morris & White 2013). We are expanding the platform to non-ecological data and plan to expand its provenance and reproducibility functionality. The next step is to tackle the challenge of combining datasets. I will lead the development of a general tool for automatically combining multiple heterogeneous datasets in reproducible ways. This tool will build on our successes in solving the individual dataset problem, using generalized routines to automate the handling of standard tasks involved in assembling datasets while leveraging human collaboration to develop the metadata describing how to combine datasets. This tool will interface with efforts for acquiring and streaming data, such as dat, rOpenSci, rOpenGov, NEON and our Data Retriever, to allow data from all three dimensions of big data to be easily combined to answer fundamental scientific questions.

2.2 Methods for Complex Data

Most data science methodologies assume that while data may be large and heterogeneous the data themselves are relatively simple: responses are linear, there is a single response variable, and data points are independent and identically distributed. However, many data-intensive questions involve data that violate all of these assumptions. For example, my research requires simultaneously predicting the interrelated abundances of hundreds of species that respond to climate in non-linear ways (Harris 2014), with context-dependent interactions among species (Poisot et al. 2014), where standard cross-validation fails due to strong spatial correlations in both features and outcomes (Bahn & McGill 2012). These challenges apply to many areas of data science. They require complex approaches that are capable of simultaneously handling non-linear responses and predicting high-dimensional joint distributions as outcomes (e.g., stochastic neural networks, Markov random fields), and methods for handling complexities such as spatial autocorrelation, irregularly sampled time-series, and missing data. We will build on existing methods (Le Rest et al. 2014) to provide general solutions to cross-validation in spatially and temporally autocorrelated contexts, build general implementations of our approaches for forecasting the distributions of species and ecosystem services, and extend methods for dealing with missing and irregularly sampled data. The solutions we develop will be broadly useful to any field that deals with complex data. Our core focus will be developing both tools and training to allow scientists across disciplines to take advantage of these approaches.

2.3 Building Human Capital

Tools can help bring data to scientists, but they cannot overcome the lack of individuals with the skills to conduct data-intensive research. To build human capital, we need to train scientists at all levels in the tools and approaches for tackling data-intensive problems. Just like open source software projects, training initiatives benefit from collaboration and community. This is why I focus my training efforts as a core member of the Software Carpentry team. While data science skills overlap with software development skills, major aspects of data science approaches are not covered in the current Software Carpentry curriculum. I am part of a core group that is in the early stages of developing a Data Carpentry curriculum that focuses on the tools and approaches of data science. I would use support from this award to help build both beginner and advanced curricula, to teach this material in workshops and university courses, and to develop approaches to engaging scientists in collaborative open-source communities. This will help produce a new generation of data-intensive scientists with the ability to work collaboratively to address fundamental questions using the variety, volume, and velocity of data that are now available.

Ethan P. White

Department of Biology and the Ecology Center, Utah State University, Logan, UT 84322
http://ethanwhite.org, ethan.white@usu.edu, 435-760-1909

Professional Preparation

2005	PhD Biology (with distinction), University of New Mexico
1998	BA Biology (*magna cum laude*), Colorado College

Appointments

2012-	Associate Professor, Dept. of Biology and Ecology Center, Utah State University
2012-	Senior Scientist, Sevilleta Long-Term Ecological Research Station
2007-2012	Assistant Professor, Dept. of Biology and Ecology Center, Utah State University
2005-2007	NSF Postdoctoral Fellow in Biological Informatics, Univ. of AZ & U.C. Merced

Awards and Fellowships

NSF CAREER 'Young Investigators' Award 2010-2015
NSF Postdoctoral Fellowship in Biological Informatics 2005-2007
NSF Graduate Research Fellowship 2000-2005

Five Most Relevant Publications and Products

Morris, B.D. and E.P. White. 2013. The EcoData Retriever: improving access to existing ecological data. PLOS ONE 8:e65848. http://doi.org/doi:10.1371/journal.pone.0065848. Website. GitHub.

Locey, K.J. and E.P. White. 2013. How species richness and total abundance constrain the distribution of abundance. Ecology Letters. 16:1177-1185. http://doi.org/10.1111/ele.12154

White, E.P., E. Baldridge, Z.T. Brym, K.J. Locey, D.J. McGlinn, S.R. Supp. 2013. Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution 6(2):1-10. http://doi.org/10.4033/iee.2013.6b.6.f

White, E.P., K.M. Thibault, and X. Xiao. 2012. Characterizing species-abundance distributions across taxa and ecosystems using a simple maximum entropy model. Ecology 93:1772-1778. http://doi.org/10.1890/11-2177.1

White, E.P., S.K.M. Ernest, P.B. Adler, A.H. Hurlbert, and S.K. Lyons. 2010. Integrating spatial and temporal approaches to understanding species richness. Philosophical Transactions of the Royal Society B 365:3633-3643. http://doi.org/10.1098/rstb.2010.0280

Five Other Publications and Products

Wilson, G., D.A. Aruliah, C.T. Brown, N.P. Chue Hong, M. Davis, R.T. Guy, S.H.D. Haddock, K. Huff, I. Mitchell, M. Plumbley, B. Waugh, E.P. White, and P. Wilson. 2014. Best practices for scientific computing. PLOS Biology. 12:e1001745. http://doi.org/10.1371/journal.pbio.1001745

Thibault, K.M., E.P. White, A.H. Hurlbert, and S.K.M. Ernest. 2011. Multimodality in the individual size distribution of bird communities. Global Ecology and Biogeography 20:145-153. http://doi.org/10.1111/j.1466-8238.2010.00576.x

Xiao, X., White, E.P., M.B. Hooten, and S.L. Durham. 2011. On the use of log-transformation vs. nonlinear regression for analyzing biological power-laws. Ecology 92:1887-1894. http://doi.org/10.1890/11-0538.1

White, E.P. and A.H. Hurlbert. 2010. The combined influence of the local environment and regional enrichment on bird species richness. American Naturalist 172:E35-E43. http://doi.org/10.1086/649578

Price, C.A., K. Ogle, E.P. White, and J.S. Weitz. 2009. Evaluating scaling models in biology using hierarchical Bayesian approaches. Ecology Letters 12:641-651. http://doi.org/10.1111/j.1461-0248.2009.01316.x

Synergistic Activities

Ecological Data Wiki: Founder and developer of a wiki based website that allows ecologists to collaborate on the discovery and use of available ecological datasets (http://ecologicaldata.org). The site has been viewed nearly 75,000 times by users in over 140 countries.

EcoData Retriever: Leader of an open source software project that downloads, cleans up, restructures, and installs ecological datasets so that scientists can focus on doing science (http://ecodataretriever.org). Code is available on GitHub. Altmetrics show that the Retriever is highly recommended and highly cited. The installers for the last release have been downloaded over 500 times.

Computational Biology Course & Website Development: Developed a suite of university courses on computational methods for biologists. Course material is available online to facilitate broader learning (http://programmingforbiologists.org) and has been viewed over 140,000 times by users in over 170 countries.

Software Carpentry: Advisory board member, instructor, and maintainer of intermediate material for a non-profit that provides training in computational best practices to scientists (http://software-carpentry.org). Examples of my contributions to the instructional material include an introduction to Python for non-Python programmers and material on modularization and documentation. I have organized and taught 8 workshops in the last 2 years, and members of my research group been involved in more than 10 additional workshops.

Impactstory Board of Directors: Member of the board of directors for Impactstory, a non-profit whose goal is to improve the practice of science by providing tools to quantify the broad impact of diverse research products including software, datasets, presentations and publications.

Collaborators and Other Affiliations

Collaborators: P. Adler (Utah State), D.A. Aruliah (U. Ontario), C.T. Brown (Michigan St.), J. Coyle (U. North Carolina), M. Davis (Data Pad), S. Durham (Utah State), B. Enquist (U. Arizona), M. Giffin (Utah State), J. Gittleman (U. Georgia), J. Goheen (U. Wyoming), J. Green (U. Oregon), R.T. Guy(U. Toronto), S.H.D. Haddock (Monterey Bay Aquarium), N.P. Chue Hong (Software Sustainability Institute), M. Hooten (Colorado State), K. Huff (UC Berkeley), A. Hurlbert (U. North Carolina), N. Isaac (Centre for Ecology and Hydrology, UK), S. Lyons (Smithsonian Institute), I. Mitchell (U. British Columbia), H. Morlon (Ecole Polytechnique), B. Morris (U. North Carolina), K. Ogle (Arizona State), M. Plumbley (Queen Mary Univ.), C. Price (U. Western Australia), R. Sibly (U. Reading, UK), F. Smith (U. New Mexico), J. Stegen (Pacific Northwest National Lab), S. Supp (Utah State), B. Waugh (University College London), J. Weitz (Georgia Tech), G. Wilson (Mozilla), P. Wilson (U. Wisconsin).

Advisors: J.H. Brown (U. New Mexico; PhD); J.L. Green (U. of Oregon; postdoc), B.J. Enquist (U. of Arizona; postdoc)

Advisees: E. Baldridge (Utah State; PhD), K.J. Locey (Indiana Univ; PhD), D.J. McGlinn (Utah State; postdoc), K. Riemer (Utah State; PhD), K. M. Thibault (National Ecological Observatory Network; Postdoc), X. Xiao (Utah State; PhD)