Ecosystems (2017) 20: 274–283
DOI: 10.1007/s10021-016-0075-y

ECOSYSTEMS

20th Anniversary Paper

The Next Decade of Big Data in Ecosystem Science

S. L. LaDeau,* B. A. Han, E. J. Rosi-Marshall, and K. C. Weathers

Cary Institute of Ecosystem Studies, Millbrook, New York 12545, USA

Abstract

Ecosystem scientists will increasingly be called on to inform forecasts and define uncertainty about how changing planet conditions affect human well-being. We should be prepared to leverage the best tools available, including big data. Use of the term 'big data' implies an approach that includes capacity to aggregate, search, cross-reference, and mine large volumes of data to generate new understanding that can inform decision-making about emergent properties of complex systems. Although big-data approaches are not a panacea, there are large-scale environmental questions for which big data are well suited, even necessary. Ecosystems are complex biophysical systems that are not easily defined by any one data type, location, or time. Understanding complex ecosystem properties is data intensive along axes of volume (size of data), velocity (frequency of data), and variety (diversity of data types). Ecosystem scientists have employed impressive technology for generating high-frequency, large-volume data streams. Yet important challenges remain in both theoretical and infrastructural development to support visualization and analysis of large and diverse data. The way forward includes greater support for network science approaches, and for development of big-data infrastructure that includes capacity for visualization and analysis of integrated data products. Likewise, a new paradigm of cross-disciplinary training and professional evaluation is needed to increase the human capital to fully exploit big-data analytics in a way that is sustainable and adaptable to emerging disciplinary needs.

Key words: network science; eco-analytics; forecast; scale; data mining; prediction.

Introduction

Big data touches increasingly personal aspects of each of our lives, from health to shopping and entertainment preferences. Ecosystem scientists curate very little of the types of social media, consumer-based, and medical data that have motivated much of the technological and analytical development in informatics (Chang and others 2014; Tirunillai and Tellis 2014; Han and others 2015; Hoegh and others 2015; Culotta and Cutler 2016; Flechet and others 2016). This is reflected in the fact that more than 70% of the published articles from the past ten years in Web of Science that refer to *big data* are from computer science, engineering, telecommunication, and business research fields. Yet there is a growing core of ecosystem work that is persistently expanding the scope for how our field defines, handles, and exploits big-data products and approaches. This is evidenced by publications from data-driven networks like GLEON (Global Lake Observatory Network, http://gleon.

Received 29 June 2016; accepted 13 September 2016;
published online 21 November 2016
Author contributions All authors contributed to manuscript preparation.

*Corresponding author; e-mail: ladeaus@caryinstitute.org

org), FLUXNET (http://fluxnet.ornl.gov), and by ventures like the U.S. National Science Foundation's $400 million support for a National Ecological Observatory Network (NEON) (Wilson and others 2002; Baldocchi 2003; Michener and others 2012; Weathers and others 2013; McDowell 2015; Hanson and others 2016). Likewise, eco-informatics platforms like DataONE (Data Observation Network for Earth, www.dataone.org) have substantially advanced 'discoverability' of these big-data products (Michener and others 2012). Despite this progress, we identify a growing need for theoretical development and infrastructure support to better manage and integrate cross-scale data (for example, at different organizational levels). These advances must address key challenges that include real-time prediction of ecosystem properties and the need for probabilistic forecasts to better understand how ecosystem function might be altered under global change scenarios.

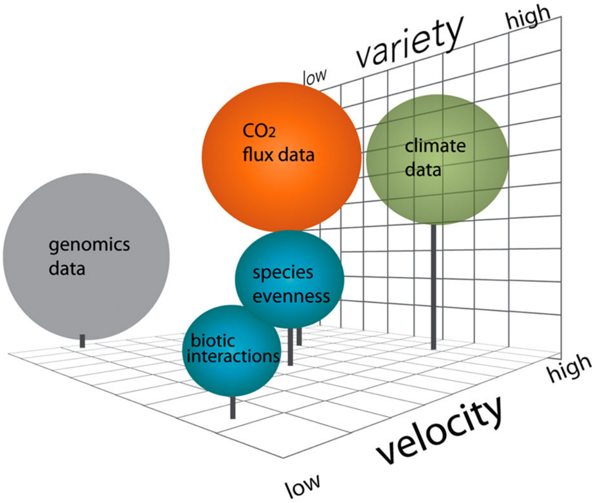

Figure 1. The next decade will demand advances driven by integration across multiple scales and sources of data to address critical questions about how ecosystems function and change. Each of the bubbles, scaled by relative Volume, represents an individual data product that by itself might be considered 'big' by axes of Volume, Velocity, or Variety. Height on this figure is arbitrary for purposes of visualization. Measures of CO_2 flux from a single tower can generate great volume due to the high velocity of measurement records. The volume is further increased if multiple tower sites are considered, although variety remains low because the data are all the same type of measurement. Likewise, even a single organism can generate high-volume genetic sequence data. Community ecology data, including biodiversity metrics and biotic interactions, are often 'smaller' data, although measuring biodiversity can require multiple types of data (high variety). Ecosystem function is a complex process that should be informed by many data types from multiple scales.

For much of the past decade, big data has been used to indicate data volumes over a terabyte—a storage capacity that was only made available on personal computers in 2007. However, volume is not the only dimension of value for big data (Figure 1), which also encompasses data variety, velocity, and, in some literature, veracity (Kwon and others 2014; Lovelace and others 2016). Variety, or structural heterogeneity, generally refers to data that integrate tabular (numeric) data with images, text, or other unstructured sources. We extend this to also encompass data from different scales and/or organizational (for example, taxonomic) levels. Velocity captures the speed at which data are generated, with the expectation that data collected at high velocities can inform real-time (or near-time) analytics. Veracity acknowledges the importance of identifying data reliability and might best be evaluated via ground-truthing and cross-validation across different data types (Lovelace and others 2016). Specific definitions vary in both time and disciplinary space—as what is 'big' gets redefined with each technological and analytical advance that is adopted. However, volume is a key dimension across fields and for all involved, big data is notable for being *bigger* than the standard data that are collected and analyzed with conventional methods. The broadest definitions incorporate a hope and expectation that big data ultimately represents an evolving capacity to search, aggregate, and cross-reference large datasets to inform decision-making and to generate new understanding about emergent properties of complex systems (Evans and others 2013; Tinati and others 2014). In this sense, big data is not exclusively generated to address a specific hypothesis but is inclusive of opportunity for addressing many yet-to-be-defined hypotheses and for generating predictive inference.

Below we provide a brief synthesis of how understanding and use of big data has developed in ecosystem research over the past ten years. We also highlight current technological and cultural challenges limiting our field from exploiting big-data approaches. Finally, we describe a vision for the next decade of big data in ecosystem science that will redefine the scope of what is currently considered *big*, as our capacity for managing and analyzing large datasets continues to grow. We anticipate that the field will further develop and support network and team science that can effectively integrate diverse data sources and generate cross-scale inference with clear and quantitative definitions of uncertainty. Further, we expect that

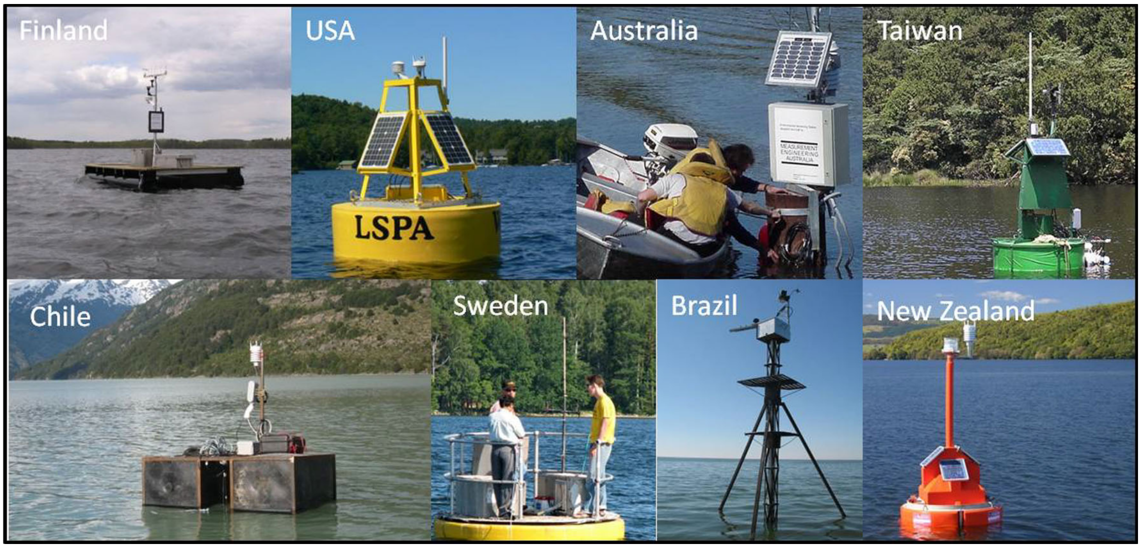

Figure 2. The global lakes ecological observatory network (GLEON) integrates buoy sensor data from more than 80 lakes in 51 countries across 6 continents.

ecosystem scientists will increasingly employ big-data approaches to understand how a growing human population and global climate change influence ecosystem function and stability. The next decade will certainly see growing demand for forecasts driven by big data that are aimed to guide policy and management needs, and ecosystem scientists will increasingly need to evaluate how research can concurrently support both theoretical understanding and forecasting needs.

Understanding Ecosystems with Big Data

Ecologists and ecosystem scientists in particular have a long history with big-data concepts. A number of recent publications illustrate the challenges and opportunities attributed to big data (Michener and Jones 2012; Hampton and others 2013; Raffaelli and others 2014; Han and others 2015; Weathers and others 2016); each describes persistent challenges, including the standardization of metadata, units, and protocols, data 'discoverability,' and ease of visualization and analysis within or between connected cyberinfrastructure and analysis platforms (for example, DataONE and statistical software). While ecosystem scientists have engaged in and leveraged big-data approaches, ecosystem science is still largely a discipline that explains pattern and process rather than one adept at predicting them (Dietze and others 2013; Niu and others 2014). This is in part because there is still so much about pattern and process that requires explaining.

Ecosystem scientists strive to understand biophysical processes, as well as the complex applications and implications of those processes (for example, ecosystem function, stability, and services). Big data could substantially advance understanding and forecasting capacity for ecosystem science over the next decade. Although data quality and keen researcher insights will remain critical in any analysis, engaging big-data technologies and philosophies can provide robust approaches and tools to make inroads towards answering some complex questions in ecosystem science, such as how ecosystem function is affected by changing climate across local to regional scales, how stability of ecosystem processes is supported by biodiversity, and others. Furthermore, the clear connections between big data and network science approaches denote a potential shift towards a future where early career faculty are expected both to develop collaborations and are rewarded for doing them well.

Individual research in ecosystem science often generates smaller data but, even so, requires integration of information across space, time, and types of data to address fundamental questions about how earth's biophysical systems function (Figure 1). Assembling and curating measurements and metadata from across spatially distinct sites, such as the GLEON model (Box 1), is a network science approach—where individual and/or collaborative groups of researchers build a database that, as a whole, can generate inference and predictions more effectively than any single data component could by itself. Network science engages multiple investigators in data collection, validation, curation, and, in some cases, synthesis processes. This requires collaborative, open data sharing and documentation (Hampton and Parker 2011; Stokstad 2011; Wallis and others 2013; Hampton and others 2015; Laney and others 2015). Examples include the following: networks that prescribe and support experimental, processing and/or data entry protocols (that is, FLUXNET, NEON); those that are more loosely organized around use of shared technology and protocols (GLEON); or a network of independent research programs that shares intent and a common data repository (for example, NSF's LTER program, Box 2). These networks provide big data, although the current infrastructure and analytics support often fall short of facilitating big-data analytics (see Boxes 1 and 2).

Box 1. The global lake ecological observatory network (GLEON)

Big data is at the core of the grassroots, Global Ecological Observatory Network's (GLEON, www.gleon.org) science, outreach, and educational activities (Weathers and others 2013; Hanson and others 2016). GLEON is really three networks: a network of lakes (80), a network of people (over 550), and a network of data (approximately 1 million data records, est.; Weathers and others 2013; Hanson and others in press). GLEON's first mission, when it was established in 2005, was to build and grow a scalable (meaning expandable to other places), persistent network of lake ecological observatories. Early activities included not only encouraging the installation of buoys around the world (Figure 2), but building a centralized database and the cyberinfrastructure to support sharing and discovery of GLEON's complex, high- and low-frequency data on lake and reservoir ecosystems. However, GLEON abandoned the creation of its own data infrastructure management program approximately five years into its existence; it was too expensive and not feasible as part of the research project. The data were too complex and heterogeneous, the task too big, and the database and infrastructure were ultimately unsustainable by a grassroots effort. GLEON then began to explore partnerships with groups focused on develop the tools to receive harmonized sensor data, as well as other data and metadata from scientific and citizen science efforts from around the world. The GLEON network is a member node in the DataONE platform, with 17 unique data streams currently maintained and 266 data downloads since 2014. Streamlining protocols for visualization and summary analyses of buoy data from multiple sites in real-time remains a goal, although no easy solution appears on the immediate horizon.

In 2013, GLEON's mission was reformulated to better capture its scientific foci: GLEON conducts innovative science by sharing and interpreting high-resolution sensor data to understand, predict, and communicate the role and response of lakes in a changing global environment. Some of the new insights and products that have resulted from GLEON analyses of big data include the variable response (in time and space) of lakes to extreme events (Jennings and others 2012; Klug and others 2012), new insights into basic ecological principles such as the role of temperature dependence as a driver of lake respiration (Yvon-Durocher and others 2012), and lake temperature responses to global climate change. Further, GLEON's data have been used to create open-source ecosystem models (Hamilton and others 2015). Although papers, theses, and products attributed to GLEON are now close to 200, and data records from GLEON sites are likely to number over 200 million, scientific discovery and data visualization could be *much* faster and more diverse if streaming data from GLEON lake ecosystems could be accessed, harmonized, and shared around the world. GLEON has built a collaborative culture that is quintessentially poised to use big data to understand lake and reservoir ecosystems (Hanson and others 2016); the lack of cyberinfrastructure to handle high-frequency data streams from diverse sources, metadata standards, and controlled vocabulary in addition to the (political and regulatory) challenge of sharing some data across institutions and cultures remains a primary limitation to scientific discovery within GLEON, and elsewhere.

Box 2. Long-term ecological research and sharing big data across space and time

The United States' federally funded long-term ecological research (LTER) sites were first established in 1980, and over 20 sites collect long-term site-based data. Each site must collect long-term data in 5 core areas (primary production, population studies, movement of organic matter, movement of inorganic matter, and disturbance patterns). As a result, LTER data at any one site have relatively high variety (multiple data types), and although developed to address site-specific questions, each site generates similar types of data. The LTER network was an early adopter of data availability, and support for network information managers to develop a common system-motivated development of the PASTA (Provenance Aware Synthesis Tracking Architecture) infrastructure to promote sharing and synthesis of data from across sites. PASTA is a data portal that was first implemented in 2013. Data entered into PASTA must meet criteria for metadata quality to ensure consistency among datasets in the system. In addition, the LTER network has developed a data use policy (http://www.lternet.edu/policies/data-access) and has DOI for all datasets to allow for appropriate citation of LTER data. The LTER program is also a member node of the DataONE platform, with 235,767 data streams from across the 27 U.S.-based sites and 959 data downloads as of August 2016

Generating and curating data to conform to a particular network platform can mean relinquishing investigator control to a wider community. This could be a daunting prospect for many scientists, especially early in their career. The academic tenure process, for example, continues to undervalue an individual's contribution to publications with many authors, or to network science more generally (Uriarte and others 2007; Goring and others 2014). Likewise, the growth of a big-data network that integrates across multiple sources of data takes time and requires sustained support—beyond the traditional 3- to 5-year funding cycle (Ruegg and others 2014). The development of the global FLUXNET program is funded by multiple, international agencies and is one example of ecosystem data networks that have advanced capability for modeling global carbon cycles, developing new mechanistic hypotheses, and exploring predictive capacity (Baldocchi 2003; Xiao and others 2014; Rao and others 2015). The level of infrastructure support that has made FLUXNET successful, for example, has only recently gained prominence in the mission of funding agencies. A majority of funding has historically prioritized the generation of new data, and the technological infrastructure and curation required for synthesis and analytics remains a significant limitation for ecosystem scientists, especially those who do not focus on terrestrial carbon or climate research.

Detailed experimental and site-specific studies can deepen our understanding of how ecosystem processes work, but greater theoretical and informatics developments are still needed to guide data integration to effectively predict ecosystem function. Cross-scale inference in space or time remains a challenge for developing predictive capacity in ecosystem science (Soranno and others 2014; Mouquet and others 2015; Petchey and others 2015; Price and Schmitz 2016). Recent ecological examples highlighting the utility of big data for answering questions at relevant spatial scales include a study that examined species distributions and habitat boundaries through high-resolution, large-scale, long-term cloud cover data using remote sensing (Wilson and Jetz 2016). Combining this data product with existing knowledge about the role of cloud cover in key life history characteristics of animal species (for example, growth, reproduction, and behavior) offers a way to combine high-frequency, high-volume data of high veracity to better estimate habitat transitions and species distributions across a large spatial extent. Although these and other recent studies in terrestrial ecosystems demonstrate how big-data products across organizational levels can help scale up inference in space (Xiao and others 2014), developing the technological and analytical infrastructure for integrating diverse data sources remains a primary challenge for ecosystem science. Investing in these developments will pay off in the near term through the immediate use of currently available big data to generate new hypotheses, and will pay dividends in the future as macroecological insights arising from cross-scale analyses contribute to the development of ecosystem theory.

Big-Data Approaches for Ecosystem Science

Big-data approaches open up unprecedented opportunities to generate new knowledge in ecosystem science. Ecologists are trained to develop hypotheses through observation of the natural world. However, formal training is often directed more toward placing our observations (data) within the context of literature rather than in conducting synthesis or meta-analysis, and even less toward eco-informatics approaches needed to create and use big data (Michener and Jones 2012; Touchon and McCoy 2016). The computational advances offered by machine learning (Peters and others 2014) and data mining (Hochachka and others 2007) tools, for example, enable analyses to subsume ecological heterogeneity and common data caveats rather than 'controlling' them. Thus, big data offers a quantitative departure from the constraints of a reductionist approach in ecosystem science, whose questions are often at much larger scales than other subdisciplines in ecology. Successfully exploiting big-data approaches will require scientific exchange between controlled experimental design and big-data products. Broadening the scope of ecological questions can mean giving up some of the control that we have been trained to consider as a gold standard of empirical research, although individual data still derive from many well-controlled studies. In exchange, using big-data approaches can reveal important contours to the ecosystem that will remain invisible to us as long as our scope of view remains fixed on questions that are conducive to manipulation (statistical or otherwise). Shifting our perspective offers us an opportunity to identify unforeseen answers to old questions, and new hypotheses that might be tested in empirically tractable, controlled settings (Stephens and others 2016).

Data collation A predominant task in any big-data approach is collating existing data. Most ecological datasets do not adhere to common standards or methods of access (Parsons and others 2011). Following pre-processing, validation, and quality control, data (and/or their metadata) can be archived as standalone synthetic products through a number of online repositories (for example, Ecological Archives (http://esapubs.org/archive/), Knowledge Network for Biocomplexity (KNB, https://knb.ecoinformatics.org/), and others). Although these online data products are becoming more commonplace in ecological research, the majority of the workflow is still largely developed on an individual basis, in part because the data products are individually unique (Michener and others 2012; Michener 2015). Finding and using existing datasets can be a time-consuming and error-prone process (Roche and others 2015). Developing a more standardized and streamlined process for data science in ecology is increasingly necessary, and the DataONE platform has greatly advanced this goal (Michener and others 2012). The DataONE platform (www.dataone.org) represents an important advance in big-data cyberinfrastructure and practice in ecosystem science (Michener and others 2012; Michener 2015). This platform is an organizational nexus for discovering data from across member nodes and a resource for training in data sharing, ethics, and informatics. There are currently 36 member nodes, including the U.S. and European LTER programs, NEON, GLEON, and USGS data repositories. The site has recorded 348,243 data downloads (of the 404,097 datasets uploaded) as of September 2016. Software such as the EcoData Retriever have been developed to automate the tasks of finding, downloading, and reformatting ecological data files, streamlining the process to get from data discovery to analysis (Morris and White 2013). The popularization of data and software carpentry workshops (http://software-carpentry.org//index.html, http://www.datacarpentry.org/) also points to greater movement in this direction. Still, the efficacy of any current platform to search, acquire, and support visualization or summary analyses across the U.S. government-funded datasets, for example, has yet to be proven. This is at least partly because productivity that results from platform use is difficult to track. Both funding agencies and individual journals are increasingly requiring that data are made available using online resources, which is a marked improvement, although much work remains if we are to achieve the ultimate goal of maximizing future utility of these archived data: a recent study showed that 56% of archived data in ecology and evolution are incomplete, and the archiving methods for 64% of datasets effectively prevent their reuse (Roche and others 2015). Rooting out unreliable values and aligning scale and unit across data types can be difficult to automate. Critical next steps will be to invest in improving the accessibility and therefore the utility of these data products.

Data Analysis Although many big-data products already drive ecosystem models of global carbon and nutrient cycles, for example, (Melillo and others 2002; McCarthy and others 2010; Xiao and others 2014), they are perhaps less appreciated for their value in the development of theory and driving the discovery of new hypotheses. There are two general analytical approaches to drawing inference from big data. If there is sufficient understanding to define a model, statistical regression-based methods are used to estimate parameters and evaluate hypotheses. Hierarchical (often Bayesian) statistics are adept at integrating diverse data sources to inform understanding of more complex ecosystem processes (Niu and others 2014; Hartig and others 2012).

However, when a priori understanding is insufficient to describe a process model, data mining approaches can reveal robust, multivariate patterns, identify important drivers, and generate predictions from the data themselves (Hochachka and others 2007). In ecology, exploratory approaches have historically meant making keen observations of our environment to identify ecological relationships and generate testable hypotheses. Data visualization and mining approaches have also been used as a step towards confirmatory (statistical inference) approaches. More formally, data mining refers to examining large, multivariate datasets to identify robust data patterns (for example, clusters of data points, anomalous signatures in data, or dependencies among variables) that are worth following up through focal hypothesis testing (confirmatory analysis). The use of computational algorithms, including ensemble regression and classification trees (Breiman 2001; Elith and others 2008) or association rule mining (Faust and Raes 2012), circumvents our limited ability to assign interactions and statistical distributions *a priori* to (possibly very) large numbers of variables. Such approaches are increasingly needed to draw insights across systems that are highly complex, or where replication is intractable, but this approach requires some reorientation in our thinking about data analysis (Breiman 2001).

Looking forward: the next 10 years

In an epoch of global change, determining our condition relative to the ''safe operating space'' of Earth's planetary boundaries is of paramount importance (Rockstrom and others 2009; Steffen and others 2015). To do this, ecosystem scientists will be called upon to generate predictions and to identify solutions (Evans and others 2013). Indeed, a current focus for ecosystem scientists is to quantitatively predict how complex drivers interact to influence ecosystem change (Weathers and others 2016). Beyond the generation of new knowledge and insight, big-data products should help guide theoretical development for understanding how ecosystems function across spatial scales and how stable these functions are across time. Likewise, there is also still much to be learned about what ecosystem properties are actually predictable and at what scale (Evans and others 2013; Mouquet and others 2015; Petchey and others 2015).

There are multiple sources of uncertainty in big-data products and in the models that use them (Schaefer and others 2012; Raczka and others 2013; Xiao and others 2014), and ecosystem scientists will continue to define and understand uncertainty in data, models, and prediction. Veracity is often listed as one of the V's defining big data (Lovelace and others 2016). Although veracity is critical for defining the boundaries of how to use and interpret big data, our focus on veracity goes beyond the traditional definition of measurement error and detection limits. Ecosystem management decisions that are based on big-data analytics will depend critically on data reliability across different spatial and temporal scales. Eddy covariance measures, for example, from a single tower have veracity defined by sensor technology, gap filling accuracy, and understanding of plant physiology and soil properties within the tower footprint (Falge and others 2001; Moffat and others 2007). However, the same data are less reliable for interpreting or predicting carbon dynamics at scales beyond that tower's footprint. Concretely defining uncertainty in big data and incorporating this explicitly into big-data analytics is a current focus of ecosystem research that will ultimately determine the forecast horizon (Dietze and others 2013; Petchey and others 2015).

Ecosystem science as a discipline must forecast how changes at local to planetary scales affect human well-being. Scientific research has generated an incredible amount of data (Boose and others 2007; Dietze and others 2013). Although mechanistic studies will remain the hallmark for hypothesis testing, network science approaches are increasingly important for deepening our understanding of variability, stochasticity, and uncertainty in large-scale ecosystem processes. The utility of data networks has been recognized through infrastructure funding by the U.S. government for NEON (McDowell 2015) and FLUXNET networks. Even with these and other large investments in infrastructure, there remains a significant lag in usage of this infrastructure (for example, DataONE) by most ecosystem scientists to access and curate big data. This usage lag is representative of a growing need for greater technological capacity and workforce training to facilitate the visualization and analysis of diverse data. For example, technological infrastructure must find and bring together existing disjointed datasets, and there needs to be support to maintain and update these platforms. Additionally, the transfer of big data from a repository to a separate analytical platform is a non-trivial technical hurdle that impedes wider capacity for data exploration. Developing tools that can be executed within the data discoverability platform is critical for increasing community use (Michener and others 2012). Funding sustained infrastructure to make data discoverable and usable will increase the returns on the investments already made in ecosystem science through the near-term production of ecological insights (both through testing of outstanding hypotheses and the generation of new hypotheses, as suggested through analysis of synthetic data), and by building the predictive capacity of ecosystem science.

Even with the best infrastructure, motivating a greater focus on big-data approaches will still require a shift in the way academia values scientific products that arise from big data in ecosystem ecology; for example, by generating metrics that value the contribution of data products in addition to metrics designed to measure the impact of peer-reviewed publications, as well as recognizing that analyses that generate novel hypotheses from big data are as valuable as analyses designed to test a focused hypothesis—in fact, these two approaches should go hand in hand. Valuing the collaborative efforts and skill set needed to contribute to and analyze big-data products as potentially equivalent to the design of novel, data-generating experiments to test a particular hypothesis will require a mental shift that may be unfamiliar to many and even uncomfortable for some (Breiman 2001). However, the scientific benefits and opportunities are clear. Scientific innovation has consistently been demonstrated to scale with the size and diversity of research teams (Wuchty and others 2007; Jones and others 2008; Cheruvelil and others 2014; Read and others 2016). Those who have access to big data (integrating across Vs in Figure 1) and can synthesize those data to generate new hypotheses and models are well positioned to derive inference at scales necessary to understand ecosystem function, as well as to generate forecasts that can inform management and promote stability in a changing global environment.

ACKNOWLEDGMENTS

The authors wish to thank Drs. Gary Lovett as well as Edward Rastetter and two anonymous reviewers for their thoughtful comments on versions of this manuscript. This manuscript is a product of the Cary Institute of Ecosystem Studies. Open Access was also supported by NSF-MSB grant (EF-1137327) to KCW and NSF-LTER grant (DEB 1027188) to ERM and SLL.

OPEN ACCESS

REFERENCES

Baldocchi DD. 2003. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. Glob Chang Biol 9:479–92.

Boose ER, Ellisona AM, Osterweil LJ, Clarke LA, Podorozhny R, Hadley JL, Wise A, Foster DR. 2007. Ensuring reliable datasets for environmental models and forecasts. Ecol Inform 2:237–47.

Breiman L. 2001. Statistical modeling: the two cultures. Stat Sci 16:199–215.

Chang RM, Kauffman RJ, Kwon Y. 2014. Understanding the paradigm shift to computational social science in the presence of big data. Decis Support Syst 63:67–80.

Cheruvelil KS, Soranno PA, Weathers KC, Hanson PC, Goring SJ, Filstrup CT, Read EK. 2014. Creating and maintaining high-performing collaborative research teams: the importance of diversity and interpersonal skills. Front Ecol Environ 12:31–8.

Culotta A, Cutler J. 2016. Mining brand perceptions from twitter social networks. Mark Sci 35:343–62.

Dietze MC, Lebauer DS, Kooper R. 2013. On improving the communication between models and data. Plant, Cell Environ 36:1575–85.

Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. J Anim Ecol 77:802–13.

Evans MR, Bithell M, Cornell SJ, Dall SR, Diaz S, Emmott S, Ernande B, Grimm V, Hodgson DJ, Lewis SL, Mace GM, Morecroft M, Moustakas A, Murphy E, Newbold T, Norris KJ, Petchey O, Smith M, Travis JM, Benton TG. 2013. Predictive systems ecology. Proc Biol Sci 280:20131452.

Falge E, Baldocchi D, Olson R, Anthoni P, Aubinet M, Bernhofer C, Burba G, Ceulemans R, Clement R, Dolman H, Granier A, Gross P, Grunwald T, Hollinger D, Jensen NO, Katul G, Keronen P, Kowalski A, Lai CT, Law BE, Meyers T, Moncrieff H, Moors E, Munger JW, Pilegaard K, Rannik U, Rebmann C, Suyker A, Tenhunen J, Tu K, Verma S, Vesala T, Wilson K, Wofsy S. 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. Agric For Meteorol 107:43–69.

Faust K, Raes J. 2012. Microbial interactions: from networks to models. Nat Rev Microbiol 10:538–50.

Flechet M, Grandas FG, Meyfroidt G. 2016. Informatics in neurocritical care: new ideas for Big Data. Curr Opin Crit Care 22:87–93.

Goring SJ, Weathers KC, Dodds WK, Soranno PA, Sweet LC, Cheruvelil KS, Kominoski JS, Ruegg J, Thorn AM, Utz RM. 2014. Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. Front Ecol Environ 12:39–47.

Hamilton D, Carey C, Arvola L, Arzberger P, Brewer C, Cole J, Gaiser E, Hanson P, Ibelings B, Jennings E, Kratz T, Lin F-P, McBride C, de Motta Marques D, Muraoka K, Nishri A, Qin B, Read J, Rose K, Ryder E, Weathers K, Zhu G, Trolle D, Brookes J. 2015. A global lake ecological observatory network (GLEON) for synthesising high-frequency sensor data for validation of deterministic ecological models. Inland Waters 5:49–56.

Hampton SE, Anderson SS, Bagby SC, Gries C, Han X, Hart EM, Jones MB, Lenhardt WC, Macdonald A, Michener WK, Mudge J, Pourmokhtarian A, Schildhauer MP, Schildhauer MP, Woo KH, Zimmerman N. 2015. The Tao of open science for ecology. Ecosphere 6(7):1–13.

Hampton SE, Parker JN. 2011. Collaboration and productivity in scientific synthesis. Bioscience 61:900–10.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future of ecology. Front Ecol Environ 11:156–62.

Han Z, Bennis M, Wang D, Kwon T, Cui SG. 2015. Special issue on big data networking-challenges and applications. J Commun Netw 17:545–8.

Hanson PC, Weathers KC, Kratz TK. 2016. Networked lake science: how the Global Lake Ecological Observatory (GLEON) works to understand, predict, and communicate lake ecosystem response to global change. Inland Waters 6:543–54. doi:10.5268/IW-6.4.904.

Hartig F, Dyke J, Hickler T, Higgins SI, O'Hara RB, Scheiter S, Huth A. 2012. Connecting dynamic vegetation models to data—an inverse perspective. J Biogeogr 39:2240–52.

Hochachka WM, Caruana R, Fink D, Munson A, Riedewald M, Sorokina D, Kelling S. 2007. Data-mining discovery of pattern and process in ecological systems. J Wildl Manag 71:2427–37.

Hoegh A, Leman S, Saraf P, Ramakrishnan N. 2015. Bayesian model fusion for forecasting civil unrest. Technometrics 57:332–40.

Jennings E, Jones SE, Arvola L, Staehr PA, Gaiser E, Jones ID, Weathers KC, Weyhenmeyer GA, Chiu CY, De Eyto E. 2012. Effects of weather-related episodic events in lakes: an analysis based on high-frequency data. Freshw Biol 57:589–601.

Jones BF, Wuchty S, Uzzi B. 2008. Multi-university research teams: shifting impact, geography, and stratification in science. Science 322:1259–62.

Klug JL, Richardson DC, Ewing HA, Hargreaves BR, Samal NR, Vachon D, Pierson DC, Lindsey AM, O'Donnell DM, Effler SW, Weathers KC. 2012. Ecosystem effects of a tropical cyclone on a network of lakes in Northeastern North America. Environ Sci Technol 46:11693–701.

Kwon O, Lee N, Shin B. 2014. Data quality management, data usage experience and acquisition intention of big data analytics. Int J Inf Manag 34:387–94.

Laney CM, Pennington DD, Tweedie CE. 2015. Filling the gaps: sensor network use and data-sharing practices in ecological research. Front Ecol Environ 13:363–8.

Lovelace R, Birkin M, Cross P, Clarke M. 2016. From big noise to big data: toward the verification of large data sets for understanding regional retail flows. Geogr Anal 48:59–81.

McCarthy HR, Oren R, Johnsen KH, Gallet-Budynek A, Pritchard SG, Cook CW, LaDeau SL, Jackson RB, Finzi AC. 2010. Re-assessment of plant carbon dynamics at the Duke free-air CO_2 enrichment site: interactions of atmospheric CO_2 with nitrogen and water availability over stand development. New Phytol 185:514–28.

McDowell WH. 2015. NEON and STREON: opportunities and challenges for the aquatic sciences. Freshw Sci 34:386–91.

Melillo JM, Steudler PA, Aber JD, Newkirk K, Lux H, Bowles FP, Catricala C, Magill A, Ahrens T, Morrisseau S. 2002. Soil warming and carbon-cycle feedbacks to the climate system. Science 298:2173–6.

Michener WK. 2015. Ecological data sharing. Ecol Inf 29:33–44.

Michener WK, Allard S, Budden A, Cook RB, Douglass K, Frame M, Kelling S, Koskela R, Tenopir C, Vieglais DA. 2012. Participatory design of dataONE-enabling cyberinfrastructure for the biological and environmental sciences. Ecol Inform 11:5–15.

Michener WK, Jones MB. 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol Evol 27:85–93.

Moffat AM, Papale D, Reichstein M, Hollinger DY, Richardson AD, Barr AG, Beckstein C, Braswell BH, Churkina G, Desai AR, Falge E, Gove JH, Heimann M, Hui DF, Jarvis AJ, Kattge J, Noormets A, Stauch VJ. 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agric For Meteorol 147:209–32.

Morris BD, White EP. 2013. The ecoData retriever: improving access to existing ecological data. PLoS ONE 8:e65848.

Mouquet N, Lagadeuc Y, Devictor V, Doyen L, Duputie A, Eveillard D, Faure D, Garnier E, Gimenez O, Huneman P, Jabot F, Jarne P, Joly D, Julliard R, Kefi S, Kergoat GJ, Lavorel S, Le Gall L, Meslin L, Morand S, Morin X, Morlon H, Pinay G, Pradel R, Schurr FM, Thuiller W, Loreau M. 2015. REVIEW: predictive ecology in a changing world. J Appl Ecol 52:1293–310.

Niu SL, Luo YQ, Dietze MC, Keenan TF, Shi Z, Li JW, Chapin FS. 2014. The role of data assimilation in predictive ecology. Ecosphere 5:65.

Parsons MA, Godoy O, LeDrew E, de Bruin TF, Danis B, Tomlinson S, Carlson D. 2011. A conceptual framework for managing very diverse data for complex, interdisciplinary science. J Inf Sci 37:555–69.

Petchey OL, Pontarp M, Massie TM, Kefi S, Ozgul A, Weilenmann M, Palamara GM, Altermatt F, Matthews B, Levine JM, Childs DZ, McGill BJ, Schaepman ME, Schmid B, Spaak P, Beckerman AP, Pennekamp F, Pearse IS. 2015. The ecological forecast horizon, and examples of its uses and determinants. Ecol Lett 18:597–611.

Peters D, Havstad P, Cushing KM, Cushing J, Tweedie C, Fuentes O, Villanueva-Rosales N. 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5(6):1–15.

Price SA, Schmitz L. 2016. A promising future for integrative biodiversity research: an increased role of scale-dependency and functional biology. Philos Trans R Soc Lond B 371:20150228.

Raczka BM, Davis KJ, Huntzinger D, Neilson RP, Poulter B, Richardson AD, Xiao JF, Baker I, Ciais P, Keenan TF, Law B, Post WM, Ricciuto D, Schaefer K, Tian HQ, Tomelleri E, Verbeeck H, Viovy N. 2013. Evaluation of continental carbon cycle simulations with North American flux tower observations. Ecol Monogr 83:531–56.

Raffaelli D, Bullock JM, Cinderby S, Durance I, Emmett B, Harris J, Hicks K, Oliver TH, Patersonk D, White PCL. 2014. Big Data and ecosystem research programmes. In: Woodward G, Dumbrell AJ, Baird DJ, Hajibabaei M, Eds. Advances in ecological research, big data in ecology, Vol. 51p 41–77.

Rao P, Kwon J, Lee SJ, Subramaniam LV. 2015. Advanced big data management and analytics for ubiquitous sensors. Int J Distrib Sens Netw, p 174894.

Read EK, O'Rourke M, Hong GS, Hanson PC, Winslow LA, Crowley S, Brewer CA, Weathers KC. 2016. Building the team for team science. Ecosphere 7:e01291.

Roche DG, Kruuk LE, Lanfear R, Binning SA. 2015. Public data archiving in ecology and evolution: how well are we doing? PLoS Biol 13:e1002295.

Rockstrom J, Steffen W, Noone K, Persson A, Chapin FS, Lambin E, Lenton TM, Scheffer M, Folke C, Schellnhuber HJ, Nykvist B, de Wit CA, Hughes T, van der Leeuw S, Rodhe H, Sorlin S, Snyder PK, Costanza R, Svedin U, Falkenmark M, Karlberg L, Corell RW, Fabry VJ, Hansen J, Walker B, Liverman D, Richardson K, Crutzen P, Foley J. 2009. Planetary boundaries: exploring the safe operating space for humanity. Ecol Soc 14:32.

Ruegg J, Gries C, Bond-Lamberty B, Bowen GJ, Felzer BS, McIntyre NE, Soranno PA, Vanderbilt KL, Weathers KC. 2014. Completing the data life cycle: using information management in macrosystems ecology research. Front Ecol Environ 12:24–30.

Schaefer K, Schwalm CR, Williams C, Arain MA, Barr A, Chen JM, Davis KJ, Dimitrov D, Hilton TW, Hollinger DY, Humphreys E, Poulter B, Raczka BM, Richardson AD, Sahoo A, Thornton P, Vargas R, Verbeeck H, Anderson R, Baker I, Black TA, Bolstad P, Chen JQ, Curtis PS, Desai AR, Dietze M, Dragoni D, Gough C, Grant RF, Gu LH, Jain A, Kucharik C, Law B, Liu SG, Lokipitiya E, Margolis HA, Matamala R, McCaughey JH, Monson R, Munger JW, Oechel W, Peng CH, Price DT, Ricciuto D, Riley WJ, Roulet N, Tian HQ, Tonitto C, Torn M, Weng ES, Zhou XL. 2012. A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis. J of Geophys Res Biogeosci 117:G03010.

Soranno PA, Cheruvelil KS, Bissell EG, Bremigan MT, Downing JA, Fergus CE, Filstrup CT, Henry EN, Lottig NR, Stanley EH, Stow CA, Tan PN, Wagner T, Webster KE. 2014. Cross-scale interactions: quantifying multiscaled cause-effect relationships in macrosystems. Front Ecol Environ 12:65–73.

Steffen W, Richardson K, Rockstrom J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de Vries W, de Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers B, Sorlin S. 2015. Sustainability. Planetary boundaries: guiding human development on a changing planet. Science 347:1259855.

Stephens PR, Altizer S, Smith KF, Alonso Aguirre A, Brown JH, Budischak SA, Byers JE, Dallas TA, Jonathan Davies T, Drake JM, Ezenwa VO, Farrell MJ, Gittleman JL, Han BA, Huang S, Hutchinson RA, Johnson P, Nunn CL, Onstad D, Park A, Vazquez-Prokopec GM, Schmidt JP, Poulin R. 2016. The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. Ecol Lett 19:1159–71.

Stokstad E. 2011. Network science open-source ecology takes root across the world. Science 334:308–9.

Tinati R, Halford S, Carr L, Pope C. 2014. Big data: methodological challenges and approaches for sociological analysis. Soc J Br Sociol Assoc 48:663–81.

Tirunillai S, Tellis GJ. 2014. Mining marketing meaning from online chatter: strategic brand analysis of big data using latent dirichlet allocation. J Mark Res 51:463–79.

Touchon JC, McCoy MW. 2016. The mismatch between current statistical practice and doctoral training in ecology. Ecosphere 7:e01394.

Uriarte M, Ewing HA, Eviner VT, Weathers KC. 2007. Constructing a broader and more inclusive value system in science. Bioscience 57:71–8.

Wallis J, Rolando CE, Borgman CL. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS ONE 8(7):e67332.

Weathers K, Hanson P, Arzberger P, Brentrup J, Brookes J, Carey C, Gaiser E, Hamilton D, Hong G, Ibelings B, Istvánovics V, Jennings E, Kim B, Kratz T, Lin F-P, Muraoka K, O'Reilly C, Piccolo M, Rose K, Ryder E, Zhu G. 2013. The global lake ecological observatory network (GLEON): the evolution of grassroots network science. Limnol Oceanogr Bu 22:71–3.

Weathers KC, Groffman PM, VanDolah E, Bernhardt E, Grimm NB, McMahon KA, Schimel J, Paolisso M, Baer S, Brauman K, Hinckley E. 2016. Frontiers in ecosystem ecology from a community perspective: the future is boundless and bright. Ecosystems.

Wilson AM, Jetz W. 2016. Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. PLoS Biol 14:e1002415.

Wilson K, Goldstein A, Falge E, Aubinet M, Baldocchi D, Berbigier P, Bernhofer C, Ceulemans R, Dolman H, Field C, Grelle A, Ibrom A, Law BE, Kowalski A, Meyers T, Moncrieff J, Monson R, Oechel W, Tenhunen J, Valentini R, Verma S. 2002. Energy balance closure at FLUXNET sites. Agric For Meteorol 113:223–43.

Wuchty S, Jones BF, Uzzi B. 2007. The increasing dominance of teams in production of knowledge. Science 316:1036–9.

Xiao JF, Ollinger SV, Frolking S, Hurtt GC, Hollinger DY, Davis KJ, Pan YD, Zhang XY, Deng F, Chen JQ, Baldocchi DD, Law BE, Arain MA, Desai AR, Richardson AD, Sun G, Amiro B, Margolis H, Gu LH, Scott RL, Blanken PD, Suyker AE. 2014. Data-driven diagnostics of terrestrial carbon dynamics over North America. Agric For Meteorol 197:142–57.

Yvon-Durocher G, Caffrey JM, Cescatti A, Dossena M, del Giorgio P, Gasol JM, Montoya JM, Pumpanen J, Staehr PA, Trimmer M, Woodward G, Allen AP. 2012. Reconciling the temperature dependence of respiration across timescales and ecosystem types. Nature 487:472–6.