Data-Intensive Ecological Research Is Catalyzed by Open Science and Team Science

KENDRA SPENCE CHERUVELIL AND PATRICIA A. SORANNO

Many problems facing society and the environment need ecologists to use increasingly larger volumes and heterogeneous types of data and approaches designed to harness such data—that is, data-intensive science. In the present article, we argue that data-intensive science will be most successful when used in combination with open science and team science. However, there are cultural barriers to adopting each of these types of science in ecology. We describe the benefits and cultural barriers that exist for each type of science and the powerful synergies realized by practicing team science and open science in conjunction with data-intensive science. Finally, we suggest that each type of science is made up of myriad practices that can be aligned along gradients from low to high level of adoption and advocate for incremental adoption of each type of science to meet the needs of the project and researchers.

Keywords: data-intensive science, open science, team science, ecology, science culture, gradient of adoption

Ecologists are increasingly being asked to answer twenty-first century research questions—questions connected to the major environmental problems facing society that are fundamentally ecological in nature, that cover broad spatial and temporal scales and that cross disciplines (Lubchenco et al. 1991). Such problems include sustaining nature's services amid an increasing human population (Palmer et al. 2005), forecasting the effects of global change on ecological systems at both fine and broad scales (Peters et al. 2014), and quantifying the contribution of ecological systems in key global cycles (Cole et al. 2007). Addressing these kinds of problems requires ecologists to do two things—synthesize diverse knowledge from a range of disciplines and perspectives (Carpenter et al. 2009) and expand the breadth of ecological knowledge and theory to a wide range of spatial and temporal scales (Palmer et al. 2005, Heffernan et al. 2014).

In fact, ecology has been moving in both of these directions for the last several decades. For example, a rich body of synthetic research has been conducted to address some of these types of problems at synthesis centers, such as the US National Center for Ecological Analysis and Synthesis, the Chinese Ecosystem Research Network, the French Centre for the Synthesis and Analysis of Biodiversity, and the US Socio-Environmental Synthesis Center (e.g., Carpenter et al. 2009, Baron et al. 2017). Second, the scale and scope of ecological research questions has been expanding through the increased development of both grassroots and top-down research networks (e.g., AmeriFlux, FluxNet, Global Lake Ecological Observatory Network, National Phenology Network, Nutrient Network, National Ecological Observatory Network). Finally, broadscale understanding of regions and continents and the integration of understanding and theory across these scales have progressed through the subdisciplines of biogeography, landscape ecology, macroecology, and macrosystems ecology (Brown 1995, Heffernan et al. 2014, Turner and Gardner 2015, Rose et al. 2017).

Much of this broadscale, networked, or synthetic research includes some form of prediction using one of three strategies: extrapolating findings from one location to another, scaling up knowledge and processes from local to regional and global extents, or forecasting knowledge from past to current and future states (figure 1). These strategies are challenging to implement and are associated with high levels of uncertainty, in part, because they often require large amounts and types of data across space and time, as well as data-intensive analytical approaches. Fortunately, ecology is currently experiencing a rapid rise in the availability of larger amounts and new types of data that expand the spatial and temporal scales of observation (Porter et al. 2012, Schimel et al. 2013, Heffernan et al. 2014). Such data include those generated from genomic sequencing, low-cost high-frequency sensors

BioScience XX: 1–10.
doi:10.1093/biosci/biy097

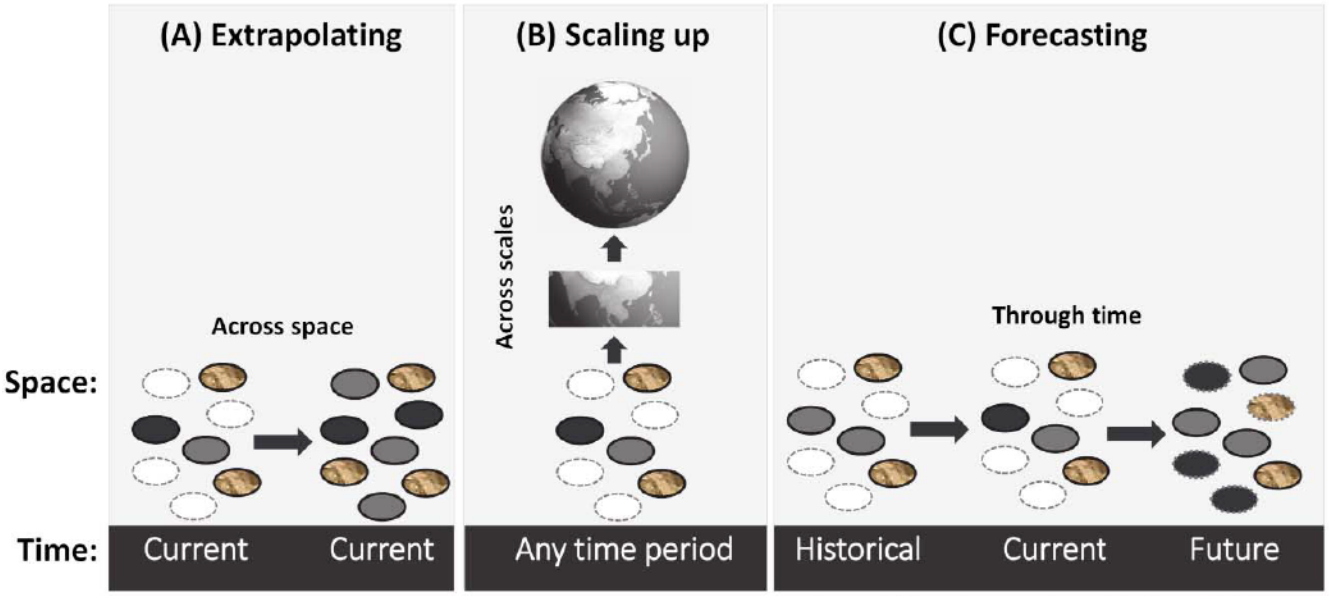

Image of globe and continent from: Globe Master 3D [CC BY 3.0 (http://creativecommons.org/licenses/by/3.0)]

Figure 1. Three strategies ecologists use to address a wide range of broadscale ecological questions often requiring data-intensive methods. Ovals represent individual ecosystems, ecosystems with similar states are the same color, and open ovals represent ecosystems with no in situ observations. These strategies are often used in synthetic analyses in which the researchers attempt to extrapolate knowledge gained from a small number of studies or ecosystems to a broader range of ecosystems (a), scale up observed estimates of important ecological processes from a number of different site-based studies to regional and global scales (b), and forecast future states of ecosystems using knowledge from current and past time periods, often in multiple sites and regions (c).

deployed across many systems, data compilation of many small studies into large integrated databases, and multispectral satellite imagery of increasingly large geographic extents, to name a few. Therefore, to answer twenty-first century ecological questions, ecology is and will become increasingly a data-intensive discipline (Peters et al. 2014, Elliott et al. 2016, Hampton et al. 2017, Farley et al. 2018).

We propose that for ecological data-intensive research to be most successful, it will be used in combination with open science and team science, both of which have already been influencing ecological practices. Although many ecologists have explicitly made compelling arguments for data-intensive, open, and team science individually or implicitly argued for some combination of these types of science (see the citations throughout this article), we believe that deliberately combining these three types of science causes synergy. For example, open science practices provide publicly accessible data, code, and methods for combining data across broad scales of time and space and across disciplines, greatly facilitating data-intensive research. Similarly, publishing open-access research articles ensures that all disciplinary scientists can access and read the content. Team-based practices provide a diverse set of people who have the knowledge and perspectives needed to combine these data across scales or disciplines using expertise from such disciplines as bioinformatics, geospatial science, statistics, or computer science (Peters et al. 2014, Soranno et al. 2015a). Finally, proponents of team science and open science have perspectives and methods that are complementary to each other and catalyze each other's practices.

Due to these (and other) examples of synergies among these three types of science, we argue that data-intensive research in ecology will be catalyzed by open science and team science. An important element of the synergies among these three types of science is that each type of science comprises a wide range of behaviors and practices that can be arranged along a gradient of adoption. We propose that as scientists move further along each gradient from low to very high levels of adoption, they maximize the synergies from open science and team science to fully accrue the benefits of data-intensive science. In this article, we (a) describe the synergies among data-intensive, open, and team science; (b) discuss cultural barriers that exist for each type of science in ecology; and (c) propose three gradients of adoption as a way to address these barriers and promote data-intensive science to answer twenty-first century ecology questions. In the subsequent sections, we summarize the evidence for these ideas from the literature drawing on past and contemporary examples of ecology's use of these types of science and provide our own experiences as examples of incremental progression along each type of science's gradient to conduct ecology research.

Definitions and benefits of data-intensive, open, and team science for ecology

Below, we define data-intensive, open, and team science and summarize the benefits of conducting each type of science individually. The exact definitions of these types of science differ within and among disciplines, and the definitions continue to evolve. However, we define data-intensive science as empirical research in which the capture, curation, and the analysis of large volumes of data is central to the scientific question (Hey et al. 2009, Critchlow and Dam 2013, Farley et al. 2018). What counts as large volumes of data is highly variable across disciplines and will change through time; however, data-intensive research in this context is strongly empirical and often relies on data from multiple sources, including those not collected by the scientist conducting the data-intensive research. Data-intensive science is distinct from data science, which develops novel methods for the analysis of data (e.g., computer science, machine learning and data mining, and statistics). Much has been written about the potential of data-intensive science for a wide range of disciplines, including ecology (e.g., Hey et al. 2009, Kelling et al. 2009, Peters et al. 2014, Hampton et al. 2017, LaDeau et al. 2017). In situ measurements of ecosystems are being compiled into databases of unprecedented broad spatial and temporal scales to ask basic ecological questions about whether results and knowledge obtained from relatively small numbers of well-studied systems can be extrapolated through time and to ecosystems across regions, continents, and the globe (Kelling et al. 2009, Hampton et al. 2013, Sharma et al. 2015, Soranno et al. 2017). In fact, empirical data-intensive approaches are helping ecologists test existing theory and understand empirical patterns that cannot be studied with finer-scale studies, ask new questions about the role of broadscale factors for driving patterns through time, and provide new insight about the scales and processes underlying patterns never before studied (O'Reilly et al. 2015, Thessen 2016, Lottig et al. 2017, Collins et al. 2018).

Open science uses a set of transparent, inclusive, and reproducible practices that result in publicly available research, data, and dissemination products (e.g., Fecher et al. 2015, Hampton et al. 2015, Lowndes et al. 2017); it also expects that all data, code and other research products are findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). Open science has the potential to benefit ecology in many ways (Hampton et al. 2015) and has been gaining popularity (Parr and Cummings 2005, Duke and Porter 2013, Lowndes et al. 2017). It stands to advance discovery, foster reproducibility, leverage investments in research, democratize science, foster a more inclusive science, and improve communication between scientists, the public, and decision-makers (Hampton et al. 2015, Soranno et al. 2015b). Such benefits are especially needed to answer integrative research questions about the patterns that exist and the processes that operate at more than one scale across space and time, which can be done by integrating smaller data sets into larger, heterogeneous ones (Peters et al. 2014, O'Reilly et al. 2015, Lottig et al. 2017, Lowndes et al. 2017, Novick et al. 2017) or through formal organizations that make data available for research on ecological change at national (US National Ecological Observatory Network) and global scales (Global Biodiversity Information Facility). Increasingly, networks and organizations are also making more than their data accessible; they are sharing lab notebooks and code with services such as GitHub, which facilitates reproducible and transparent methods.

Finally, we define team science as collaborative science that leverages the expertise of a diverse range of professionals and that incorporates practices to maximize team functioning (e.g., Stokols et al. 2008, Cheruvelil et al. 2014). Scientists have long been capitalizing on the benefits of collaborative research. For example, multiauthored publications have higher citation rates than single-authored studies (Wuchty et al. 2007), research teams that are interdisciplinary have increased creativity and productivity (Boix Mansilla et al. 2016), and those that are more diverse produce more creative and impactful outcomes (Woolley et al. 2010). In ecology, there are numerous examples of successful research collaborations, particularly in the last 20 years, including synthesis center working groups (Carpenter et al. 2009, Hampton and Parker 2011, Campbell et al. 2013, Baron et al. 2017), grassroots networks (e.g., Borer et al. 2014, Hanson et al. 2016, Novick et al. 2017), and big-science observatories (e.g., Kuhlman et al. 2016). Such efforts are producing large amounts of data, research outputs, and early career researchers with the penchant for working as part of productive teams (e.g., Read et al. 2016b, Lowndes et al. 2017). However, collaborative research is not always the same as team science, which has its roots in well-established fields such as organizational psychology and is highly informed by the recently formed, interdisciplinary science of team science, which studies the processes by which scientific teams conduct research and the circumstances that facilitate or hinder the effectiveness of collaborative research (National Research Council 2015). Therefore, team science is collaborative science that uses knowledge and practices from these two important disciplines to create and maintain productive teams. Research on science teams has shown that high-performing research teams rely on practices that maximize team function, such as clearly defining roles, responsibilities, and expectations and establishing team policies for many components of the research enterprise (e.g., authorship, data sharing; Stokols et al. 2008, Cheruvelil et al. 2014, Read et al. 2016b). As team sizes grow and team diversity increases, these practices become even more important. Therefore, data-intensive ecological research conducted by diverse teams will benefit greatly from engaging with team science behaviors and practices.

Synergies among data-intensive, open, and team science

There are many synergies among the underlying principles, strategies, and approaches of data-intensive, open, and team

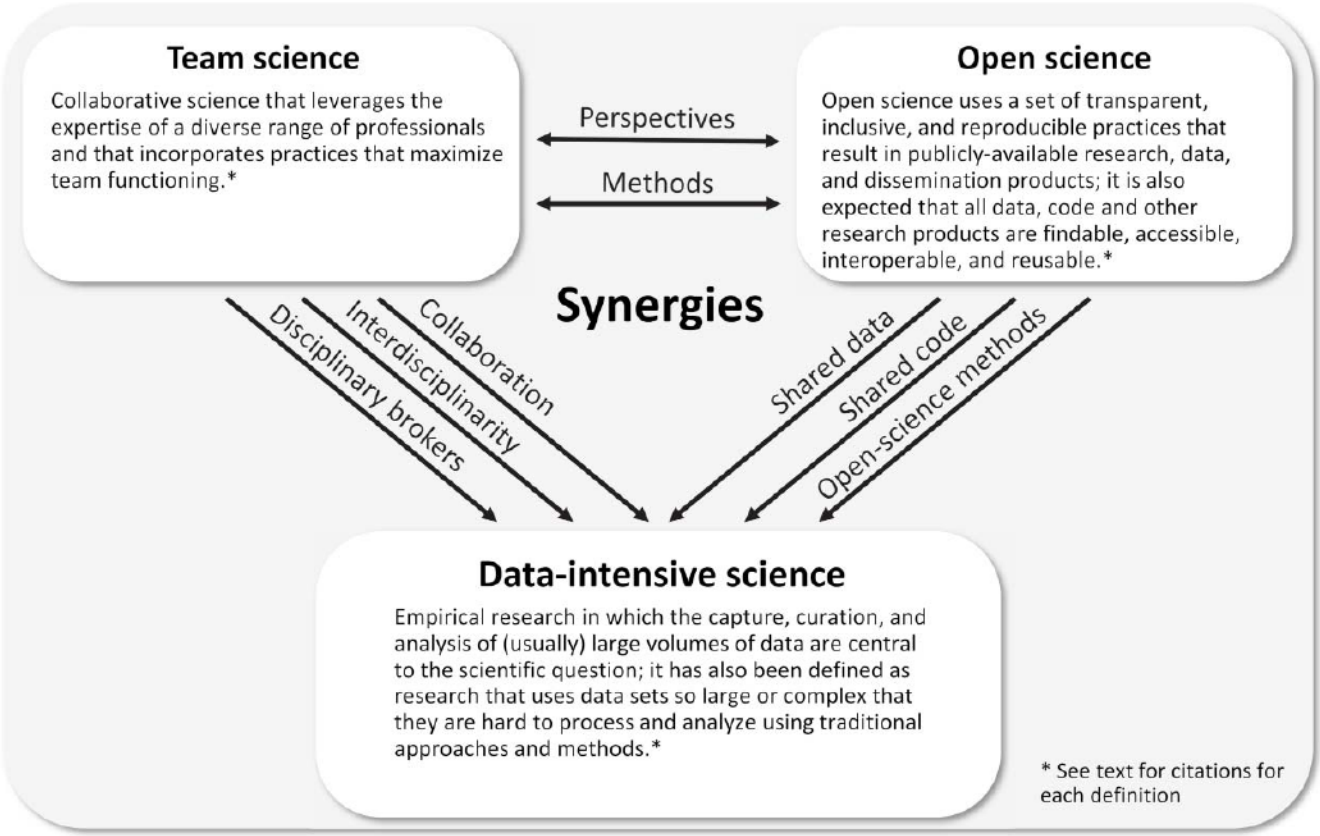

Figure 2. The synergies resulting from the combined use of data-intensive, open, and team science to answer twenty-first century ecological questions. Data-intensive science is facilitated by synergies between team science and open science (depicted by the double-headed red arrows), such as the perspectives, methods, and expertise contributed by those who practice open and team science and contribute to collaborative data-intensive research. Furthermore, team science practices facilitate data-intensive science by providing practices that ensure effective and productive collaboration, often across disciplines, and by identifying key individuals, such as disciplinary brokers who are essential for spanning disciplines to meet research objectives. Likewise, open science greatly facilitates data-intensive science by providing the tools and approaches for effectively sharing data and code and fostering the effective transfer of knowledge, data, and tools among team members and eventually all researchers.

science for those who engage with them in combination (figure 2). For example, open science and team science facilitate each other. The strategy of creating written team policies (a team science strategy) related to data sharing can increase willingness to embrace open science principles, because teams are forced to confront and reconcile differing views within the team (Cheruvelil et al. 2014). Similarly, having team members who are strong proponents of open science can help teams implement practices that facilitate team functioning. For example, using code-sharing platforms such as GitHub can promote code sharing for similar functions, which increases overall productivity and research reproducibility while also facilitating collaborations across individuals with different levels of coding expertise. Written data-sharing policies and the use of code-sharing platforms also facilitate the eventual public sharing of research products (e.g., Lowndes et al. 2017, Soranno et al. 2017).

Such synergies are highly beneficial for data-intensive ecology research, because many ecologists do not have an extensive background in creating, managing, and analyzing large and complex data sets (figure 2). For example, an open science perspective that assumes sharing of data, code, and methods, facilitates data-intensive research by developing the strategies to share early in the research project, as well as to document both the data and methods. Such practices increase overall team-level productivity by providing all team members access to the same well-documented data, code, and methods (Lowndes et al. 2017). One could imagine the inefficiencies if a single person created the database and had intimate knowledge of the data to use it but did not document that knowledge—they would have to work with each individual scientist on the team who wanted to use the data. Read and colleagues (2016a) provide an example of such promotion of data-intensive research by open science through the development of openly shared tools for data quality controls, data visualizations, and models.

Team science can also promote data-intensive ecology research by providing the strategies to ensure effective

implementation of practices and tools from other disciplines during data-intensive research. For example, Hanson and colleagues (2016) show how interdisciplinary collaborations can support data-intensive ecology research by providing essential informatics and computational expertise to create cyberinfrastructure support of network data, operations, and software tools. However, interdisciplinary research can be difficult to implement because people from different disciplines have a wide range of perspectives and values (O'Rourke et al. 2013). To overcome such challenges, team science includes such practices as inclusion of one or more members of a team who are *disciplinary brokers*—people who speak the languages of multiple disciplines (Wenger 2000, Pennington 2011). A broker greatly enhances interactions among team members of different disciplines by improving overall communication of ideas and approaches and by helping to identify important and compelling research questions for the team to address. Furthermore, team science includes training and providing opportunities to develop collaboration skills such as facilitating team discussions and conflict negotiation (Cheruvelil et al. 2014, Read et al. 2016b), which have obvious benefits for team idea generation and research productivity. Because data-intensive research teams often include more than one discipline (or perspectives within a discipline, such as an empirical and theoretical scientist), such attention to strategies that best harness these differing perspectives are especially important to ensure successful project outcomes.

Cultural barriers to data-intensive, open, and team science in ecology

Although there are many ecologists using practices from data-intensive, open, and team science, adoption of these three types of science are still the exception rather than the rule in ecology. Since technical barriers to adoption are quickly being removed (Hampton et al. 2015), it is unlikely that they are the sole cause of slow adoption; another factor slowing their adoption is a lack of cultural support. Below, we describe some of the cultural barriers to adoption of these three types of science that exist and some of the solutions that are necessary for data-intensive, open, and team science practices to flourish in ecology.

Data intensive science

Despite the potential for data-intensive science, some empirical ecologists are hesitant to add it to their research program. Much of this hesitancy stems from the argument that data-intensive approaches will result in at best, weak, and at worst, erroneous, inferences. There are three main reasons for this argument. First, critics argue that data-intensive methods are atheoretical and not hypothesis-driven, inefficient fishing expeditions or that they result in mostly spurious correlations (Elliott et al. 2016). Second, critics argue that large data sets themselves are the problem in that they are prone to characteristics that limit their usefulness including data that are noisy (because of measurement error, outliers, and missing values) and lead to weak signals, contain many correlated variables, and have complex data structures (Fan et al. 2014). Third, ecologists are used to collecting data themselves for a specific purpose; therefore, they have confidence in their data and are critical of using either other people's data or contributing their data for purposes outside of the original study design.

These major criticisms of data-intensive research argue that data-intensive approaches cannot result in strong inference when compared with well-established approaches from site-specific intensive study of single of a few ecosystems or processes. However, these criticisms arise from taking a frequentist perspective when, in fact, improper statistics can be conducted with any amount of data—large data sets are not alone in this regard (Hand 1998). Ecologists have been trained that when a sufficiently large number of statistical tests are performed with a data set, some proportion will be significant by chance alone and that the outcomes of hypothesis testing is biased if the data are not independent nor drawn from an unbiased sample. However, these concerns can be addressed using other analytical perspectives that do not rely on statistical significance, many of which have been around for several decades, including data mining, machine learning, computational approaches, Bayesian statistics, and weight of evidence modeling approaches (e.g., Burnham and Anderson 2003, Hochachka et al. 2007). In addition, dependent, unbalanced, and noisy data can be accounted for by modeling quantitative estimates of uncertainty that arise from such sources (e.g., Harwood and Stokes 2003, Petchey et al. 2015). Therefore, as more ecologists include nonfrequentist approaches as part of their research program, what constitutes "good" methods in ecology will broaden. By valuing, teaching, and rewarding the use of a variety of quantitative approaches (and their interpretation) in ecology, we can promote the inclusion of data-intensive approaches and perspectives (e.g., Peters et al. 2014, Elliott et al. 2016, LaDeau et al. 2017). Concurrently, we should incentivize researchers to build large databases and the associated tools needed to use them such as software, metadata, code, and other research products by crediting these valuable research products for career advancement (Goring et al. 2014, LaDeau et al. 2017).

Open science

Open science practices are not yet the norm in ecology. For example, the sharing of ecological data is still far from common (Wolkovich et al. 2012, Hampton et al. 2013, Roche et al. 2015), and studies have shown a large proportion of peer-reviewed ecology articles suffer from underreporting of key methods or results to make them reproducible (Parker et al. 2016). In a recent analysis of a well-established scientist-led research network, AmeriFlux, data availability from participating sites was shown to be lower than the network wanted (approximately 65%), and the proportion of sites sharing data appeared to be decreasing (Novick et al. 2017). Arguably, past (and current) scientific cultures could be characterized as highly competitive for limited

resources, including funding and data. Therefore, scientists with sole access to data, code, methods, and other tools had a competitive advantage over others. This legacy may be one contributing factor to an existing resistance to and low adoption of open science tools and practices in ecology. For example, arguments against data sharing include a reluctance to release control of the data, fear of being "scooped," fear that others will misinterpret data, fear that sharing will restrict the types of data that scientists can use, and a lack of resources or the ability to fully document the data (Wolkovich et al. 2012, Lindenmayer and Likens 2013, Fenichel and Skelly 2015, Mills et al. 2015). In fact, on the basis of such arguments, some researchers may think there are few incentives to change behavior from the past, closed, culture to a more open one.

However, when one reframes the issue to include the full range of science outcomes—answering twenty-first century ecological questions, participating in potentially high-impact research, broadening participation in science, and getting credit for research outputs in addition to publications—incentives for open-science practices become more apparent (Uriarte et al. 2007, Soranno et al. 2015b, Elliott et al. 2016, McKiernan et al. 2016). Also, scientists sharing their methods, data, and research outputs, can be identified as potential collaborators who know the data and methods well (McKiernan et al. 2016, Lowndes et al. 2017). These collaborations can increase science productivity of individual scientists, expose them to new research ideas and tools, and broaden the range of research questions an individual scientist can answer. Therefore, the apparent conflict between open and closed science resides in perception—open science practices can benefit society and science, while also benefitting individual scientists.

Team science

There remain two main cultural barriers to adoption of team science in ecology. First, the popular perception of ecology—and much of science—remains dominated by the ideals of individual achievements (Uriarte et al. 2007, Geman and Geman 2016). Many scholars were trained in the single-discipline, single-investigator approach that has dominated science disciplines to date. And, the myth of the lone genius is perpetuated by some prominent scientists falsely suggesting that great ideas rarely come from teams and calling for a return to the good old days of maverick scientists (Geman and Geman 2016). Second, most ecologists have not been trained in effective teamwork and in fact, do not include such training as a valuable component of the scientific enterprise. Because scientists rarely receive formal training in team science, they may not recognize the common power imbalances inherent in research teams (i.e., early career scientists who have low power and voice within the team) or how those power differentials may affect team practices—although early career researchers are often all too aware of such imbalances. Despite well-intentioned team members and leaders, teams without explicit procedures to ensure that low-power team members have a voice and receive deserved credit have the potential for the opposite result (e.g., Elliott et al. 2017). Luckily, the interpersonal skills required for effective and inclusive teamwork can be learned and practiced through workshops or training in practices such as facilitating discussions, critical listening, and conflict negotiation (Cheruvelil et al. 2014). The Global Lakes Ecological Observatory Network (GLEON) is a recent example of a research network that trains early career researchers in such skills, has adopted many team science approaches network-wide, and has demonstrated the benefits network science (Hipsey et al. 2015, Hanson et al. 2016, Rose et al. 2017).

To fully incorporate teams and their research into the scientific enterprise, we must stop idealizing the lone-genius mythology of science, value and incentivize team science skills such as interpersonal skills, and recognize the wide variety of ways that scientists work together to conduct research. Growing empirical evidence points to the benefits of team science and indicates that the scientists who make tomorrow's crowning achievements will be increasingly part of diverse teams with skills and perspectives that span dimensions of race, ethnicity, gender, sexuality, country of origin, career-level, and discipline, to name a few. Therefore, we need to learn for ourselves and teach the next generation of scientists effective team-based research skills. We must also value and reward team-based cultures and practices.

Gradients of adoption for data-intensive, open, and team science

One way to advance the adoption of data-intensive, open, and team science practices in ecology is to recognize that each type of science is made up of myriad practices that can be aligned along gradients from low to high level of adoption. Other researchers have argued for a similar incremental approach to incorporating open science methods into a highly collaborative research effort (Lowndes et al. 2017). However, we recommend that practices from *all three of these types of science* be incorporated incrementally along each gradient as required to meet the research needs of the project. An incremental approach to adoption that is directly linked to the needs of the project is beneficial because a new practice is more likely to be successfully adopted if there is a good justification for doing so, it is relatively easy to adopt, and there are clearly articulated benefits of doing so for both the individuals and the entire team.

For argument sake, we organized some major practices of data-intensive, open, and team science into three discrete categories of adoption, which we refer to as levels; however, we acknowledge that there are also practices between the levels. Level I is the lowest level of adoption that requires the least amount of training or new methods to implement (figure 3) and, therefore, should be the easiest to adopt. In contrast, level III requires the most training or new methods to implement. We do not argue that every ecologist or ecology research team should be at level III for all three types of

	Level of adoption I	II	III
Team science	•**Team creation** is organic or assigned by others	•**Team creation** is deliberate, leaders aware of the importance of diversity within the team	•**Team creation** is deliberate, leaders explicitly create diverse team across many dimensions
	•**Team policies and procedures** are implicit and discussed as needed	•**Team policies and procedures** are explicitly discussed	•**Team policies and procedures** are discussed, documented (written), revised, and assessed
	•**Team functioning** is implicit and discussed as needed	•**Team functioning** is explicitly discussed	•**Team functioning** is discussed, trained, practiced, and assessed
	•**Issues of power dynamics, diversity, and inclusion** are implicit and minimally acknowledged	•**Issues of power dynamics, diversity, and inclusion** are acknowledged and discussed as needed	•**Issues of power dynamics, diversity, and inclusion** are explicitly discussed and assessed
Open science	•**Data and metadata** are published in article tables or supplements	•**Data and metadata** are published in repository at article publication or project end	•**Data and metadata** are version-controlled and shared before article publication or project end
	•**Code and documentation** are on lab or individual website	•**Code and documentation** are published in repository at article publication or project end	•**Code and documentation** are version-controlled and shared before publication or project end
	•**Publications** are in closed access journals that allow self-archiving	•**Publications** are in closed access journals, but articles are made open access (paid)	•**Publications** are in open access journals
	•**Reproducibility and transparency** is based on methods in published articles	•**Reproducibility and transparency** is enhanced by metadata and documentation in repositories	•**Reproducibility and transparency** is further enhanced by accessible 'open lab notebook'
Data-intensive science	•**High volume of data** that comes from a single entity or site	•**High volume of data** that comes from multiple entities OR sites	•**High volume of data** that comes from multiple entities AND sites
	•**Type or class of high-volume data** is mostly one	•**Type or class of high-volume data** is many	•**Type or class of high-volume data** is many with complex data structures
	•**Data management or quantitative tools** are mostly existing or disciplinary	•**Data management or quantitative tools** are advanced or from other disciplines requiring new expertise	•**Data management or quantitative tools** require novel ideas from other disciplines requiring new expertise
	•**Computational constraints** are few	•**Computational constraints** are moderate	•**Computational constraints** are large and may limit analysis

Figure 3. Example behaviors and practices for data-intensive, open, and team science organized into three discrete categories of adoption (i.e., levels). The levels are arranged from left to right with the lowest level of adoption (level I) requiring the least amount of training or new methods to implement to the highest level of adoption (level III) requiring the most training or new methods to implement, and level II between these two extremes. In reality, there are three spectrums of behaviors and practices rather than three discrete categories.

science; rather, it will depend on the team, the goals of the research project, and the institutional requirements (e.g., universities, funding agencies, or publishing standards). In fact, we advocate careful consideration of the needs and goals of each project (and team) to decide where along each of the three gradients will result in the best outcomes. Many publishers require data sharing on publication, whereas just a few funding agencies are requiring the data sharing on project completion. A team that is new to open science may plan from project inception to share all data associated with project publications (level II) and develop associated policies, procedures, and templates early in the project for use by everyone in the team. Similarly, some funding agencies require a project management plan for collaborative teams as part of the proposal review process. A team that is new to team science may be prompted by this requirement to learn and practice the principles of team science to run the team effectively (moving from level I to level III).

Within this gradient of adoption framework (figure 3), we expect higher levels of adoption to result in the most positive synergistic effects of open and team science on data intensive ecology research (figure 2). However, a comparison of the outcomes of data-intensive research across levels of adoption for open science and team science warrants further research. In our own experiences described below, we took an opportunistic approach and implemented open science and team science practices as needed to conduct our data-intensive research. This incremental approach had the advantage that we could easily justify the practices to our research team rather than implementing them on the basis of principles alone. Although there may be cases in which implementing a practice on the basis of principle alone makes sense, we expect that differences in underlying values among team members could result in conflicts and resistance to such implementation (Elliott 2017), especially without explicit discussions of these underlying values (O'Rourke et al. 2013). The gradients of adoption can help provide teams with concrete practices to better implement open, team, and data-intensive science.

Our experience adopting data-intensive, open, and team science incrementally

As an example, we describe the progression of adoption that we have taken in our own research program. Although each

team will progress in its own way, we share the benefits we experienced by taking a gradual approach to incrementally incorporate more practices and behaviors along all three science gradients. We were each trained using the single-investigator mode of conducting research (albeit still collaborative) and not inculcated in open science or data-intensive research. However, during the past 20 years of working together on a range of collaborative research projects, we have increased the proportion of our research program that is data-intensive and uses team and open science. This shift happened in an unplanned way: We realized that our broadscale spatial and temporal ecology questions were impossible to answer using our past cultures and practices. We were interested in answering questions related to the extrapolation and forecasting of lake water quality at subcontinental scales in thousands of lakes in diverse regions, which were questions we could not answer alone, nor could we answer using data that only we collected. We had conducted this type of research before but at smaller scales, with fewer collaborators, with smaller data sets, and using closed science that never included publishing our data sets in public repositories.

We led an NSF-funded project of approximately 15 individuals across 6 years to build the Lake Multiscaled Geospatial and Temporal Database (LAGOS-NE) for over 50,000 lakes in the United States (Soranno et al. 2017). The project was data-intensive in that we integrated hundreds of heterogeneous data sets and types of data from thousands of systems. We estimate that our project fit within the level II of data-intensive science (figure 3) because the ecological data sets were considered large by the standards of ecology, the data complexity was high and the integration procedures could not be done manually so had to be automated. We quickly realized that conducting this kind of data-intensive research required us to include experts in database development and design and data-intensive analytical approaches (i.e., data scientists). However, we had not realized the extent to which we also needed to more fully adopt and embrace the cultures and practices of open and team science to successfully conduct this type of data-intensive research.

For example, a key practice from open science—data sharing—was an important foundation of this project because we asked approximately 70 university and government agency personnel to share their data and metadata with us and agree to make it publicly accessible at the end of our project. However, we would not be able to acknowledge all 70 of the data providers in acknowledgment sections of future manuscripts. Therefore, we decided to use the open science practice of writing a data paper that included all data providers as coauthors so that they would receive credit for their important contributions (Soranno et al. 2017). Concurrently, we made the data publicly available in a data repository (Soranno and Cheruvelil 2017), which required us to learn additional database and open science tools related to data set version control and standards and tools for metadata and database documentation (Soranno et al. 2015a). This experience also made us reflect on the broader issue of data sharing in ecology, which helped us articulate our ethical obligations for sharing our data about freshwaters, a valued natural resource (Soranno et al. 2015b). We continue to progress in our adoption of open science practices. For example, for the LAGOS-NE project, we made the data available at the end of the project or at publication time (level II, figure 3). However, our current project that is building a database for the entire US (LAGOS-US), will make each quality-controlled version of the database publicly accessible as soon as it is complete (level III, figure 3).

Team science was equally important to the success of our project that included collaborative and interdisciplinary manuscript writing. Because of the complexity of the research, we needed a team of approximately 15 individuals from a range of disciplines to conduct the work. We had to learn and implement many important principles of team science to ensure that the team of individuals with different backgrounds, needs, experiences, and power worked together to meet the project goals across the 6 years of this project. In particular, we had to learn other disciplinary cultures and create inclusive policies and practices to ensure all team members received credit and met their personal goals on the project. Because we struggled finding resources to help us in this aspect of our research project, we wrote articles describing some of these practices to help other research teams (Cheruvelil et al. 2014, Goring et al. 2014, Elliott et al. 2017, Oliver et al. 2018), and we share our team policies on our project website. Because we had more prior experience with team science before starting this project, we were able to reach a level III adoption (figure 3) within the first year or two of the project by implementing the following practices: We worked with the team to develop team policies and procedures that were written collectively and shared, we included team functioning exercises in our all-project workshops, we conducted assessment in a variety of different forms, and we discussed and considered issues related to diversity, inclusion, and power dynamics of our team members.

Many examples of research efforts are somewhere along these three gradients of adoption. Most projects (such as ours) will operate somewhere along the three gradients rather than fully adopting all practices of data-intensive, open, and team science. However, for all levels of adoption, there are practices that represent important progress to enable ecological teams to tackle challenging research questions facing the environment and society.

Conclusions

We have made the case that data-intensive ecology research can be catalyzed by using it in conjunction with open science and team science. However, because ecology is a cultural construct (as all science is), cultural barriers exist to adopting all three of these types of science. As an example of how culture influences the way ecology is conducted and evaluated, consider the debate regarding the value of

big science in ecology (Hampton et al. 2013, Soranno and Schimel 2014, Schimel and Keller 2015). Some ecologists argue that ecology is not conducive to a big science approach and use the perceived failure of the 1970s International Biosphere Program (IBP) as evidence (as was described in Aronova et al. 2010). However, when viewed through the lens of the contemporary practices of data-intensive, open, and team science, it seems plausible that many of the IBP's objectives were not met because of cultural factors related to science practice (Aronova et al. 2010). In fact, it is difficult to imagine such a broadscale, integrated, and complex undertaking succeeding without implementing team and open science practices, which were extremely rare and not valued at the time. Therefore, a 2020 IBP would look and function completely differently and would likely have vastly different outcomes. Data-intensive, open, and team-based approaches will be an essential part of the ecological toolbox for improving predictions that use extrapolation, scaling up, and forecasting, and for addressing the important research problems of the twenty-first century. Therefore, ecologists must capitalize on the cultural changes since the IBP, and continue to push the culture by making incremental steps along the gradients of adoption for each of these three types of science. As ecologists increasingly use a broader range of quantitative approaches, learn and practice interpersonal and team-functioning skills within diverse teams, and use larger accessible data sets and data sources, they will also advance the culture in ecology to value, teach, and reward these practices and perspectives.

Acknowledgments

We thank Kevin Elliott and Georgina Montgomery for 4 years of rewarding interdisciplinary scholarship that greatly informed this article (along with their friendly reviews). The manuscript also benefited from conversations or funding from the STEP team, including Isis Settles and Sheila Brassel (funding from NSF grant no. SES-1449466), Eric Tans and Devin Higgins (funding from MSU's Science Studies at State Program), Karl Smith who started KSC down the team science road, GLEON leaders Kathleen Weathers and Paul Hanson for their innovations in team and network science, the CSI Limnology team (NSF grant no. EF-1065786) and the Continental Limnology team (NSF grant no. EF-1638679), the MSU Data-intensive Limnology Lab, and funding from the USDA National Institute of Food and Agriculture, hatch project no. 176820 to PAS. Thanks to two anonymous reviewers and Karthik Ram who provided valuable comments on earlier drafts.

References cited

Aronova E, Baker KS, Oreskes N. 2010. Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. Historical Studies in the Natural Sciences 40: 183–224.

Baron JS, Specht A, Garnier E, et al. 2017. Synthesis centers as critical research infrastructure. BioScience 67: 750–759.

Boix Mansilla V, Lamont M, Sato K. 2016. Shared Cognitive–Emotional–Interactional Platforms: Markers and Conditions for Successful Interdisciplinary Collaborations. Science, Technology, & Human Values 41: 571–612.

Borer ET, Harpole WS, Adler PB, et al. 2014. Finding generality in ecology: A model for globally distributed experiments. Methods in Ecology and Evolution 5: 65–73.

Brown JH. 1995. Macroecology. University of Chicago Press.

Burnham KP, Anderson DR. 2003. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed. http://www.barnesandnoble.com/w/model-selection-and-multi-model-inference-kenneth-p-burnham/1101635043?ean=9780387953649.

Campbell LG, Mehtani S, Dozier ME, Rinehart J. 2013. Gender-heterogeneous working groups produce higher quality science. PLOS ONE 8: e79147.

Carpenter SR, Armbrust EV, Arzberger PW, et al. 2009. Accelerate synthesis in ecology and environmental sciences. BioScience 59: 699–701.

Cheruvelil KS, Soranno PA, Weathers KC, et al. 2014. Creating and maintaining high-performing collaborative research teams: The importance of diversity and interpersonal skills. Frontiers in Ecology and the Environment 12: 31–38.

Cole JJ, Prairie YT, Caraco NF, et al. 2007. Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. Ecosystems 10: 172–185.

Collins SL, Avolio ML, Gries C, et al. 2018. Temporal heterogeneity increases with spatial heterogeneity in ecological communities. Ecology 99: 858–865.

Critchlow T, Dam KK van (eds). 2013. Data-Intensive Science. Chapman and Hall/CRC.

Duke CS, Porter JH. 2013. The ethics of data sharing and reuse in biology. BioScience 63: 483–489.

Elliott KC. 2017. A Tapestry of Values: An Introduction to Values in Science. Oxford University Press.

Elliott KC, Cheruvelil KS, Montgomery GM, Soranno PA. 2016. Conceptions of good science in our data-rich world. BioScience 66: 880–889.

Elliott KC, Settles IH, Montgomery GM, et al. 2017. Honorary authorship practices in environmental science teams: Structural and cultural factors and solutions. Journal of Accounting Research 24: 80–98.

Fan J, Han F, Liu H. 2014. Challenges of big data analysis. National Science Review 1: 293–314.

Farley SF, Dawson A, Goring SJ, Williams JW. 2018. Situating ecology as a big-data science: Current advances, challenges, and solutions. BioScience 68: 563–576.

Fecher B, Friesike S, Hebing M. 2015. What drives academic data sharing? PLOS ONE 10: e0118053.

Fenichel EP, Skelly DK. 2015. Why should data be free; don't you get what you pay for? BioScience 65: 541–542.

Geman D, Geman S. 2016. Opinion: Science in the age of selfies. Proceedings of the National Academy of Sciences 113: 9384–9387.

Goring SJ, Weathers KC, Dodds WK, et al. 2014. Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. Frontiers in Ecology and the Environment 12: 39–47.

Hampton SE, Anderson SS, Bagby SC, et al. 2015. The tao of open science for ecology. Ecosphere 6: 1–13.

Hampton SE, Jones MB, Wasser LA, et al. 2017. Skills and knowledge for data-intensive environmental research. BioScience 67: 546–557.

Hampton SE, Parker JN. 2011. Collaboration and productivity in scientific synthesis. BioScience 61: 900–910.

Hampton SE, Strasser CA, Tewksbury JJ, et al. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11: 156–162.

Hand DJ. 1998. Data mining: Statistics and more? Journal of the American Statistical Association 52: 112–118.

Hanson PC, Weathers KC, Kratz TK. 2016. Networked lake science: How the Global Lake Ecological Observatory Network (GLEON) works to understand, predict, and communicate lake ecosystem response to global change. Inland Waters 6: 543–554.

Harwood J, Stokes K. 2003. Coping with uncertainty in ecological advice: Lessons from fisheries. Trends in Ecology and Evolution 18: 617–622.

Heffernan JB, Soranno PA, Angilletta MJ, et al. 2014. Macrosystems ecology: Understanding ecological patterns and processes at continental scales. Frontiers in Ecology and the Environment 12: 5–14.

Hey T, Tansley S, Tolle K (eds). 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

Hipsey MR, Hamilton DP, Hanson PC, et al. 2015. Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. Water Resources Research 51: 7023–7043.

Hochachka WM, Caruana R, Fink D, et al. 2007. Data-mining discovery of pattern and process in ecological systems. Journal of Wildlife Management 71: 2427–2437.

Kelling S, Hochachka WM, Fink D, et al. 2009. Data-intensive science: A new paradigm for biodiversity studies. BioScience 59: 613–620.

Kuhlman MR, Loescher HW, Leonard R, et al. 2016. A new engagement model to complete and operate the National Ecological Observatory Network. Bulletin of the Ecological Society of America 97: 283–287.

LaDeau SL, Han BA, Rosi-Marshall EJ, Weathers KC. 2017. The next decade of big data in ecosystem science. Ecosystems 20: 274–283.

Lindenmayer D, Likens GE. 2013. Benchmarking open access science against good science. Bulletin of the Ecological Society of America 94: 338–340.

Lottig NR, Tan P, Wagner T, et al. 2017. Macroscale patterns of synchrony identify complex relationships among spatial and temporal ecosystem drivers. Ecosphere 8: e02024.

Lowndes JSS, Best BD, Scarborough C, et al. 2017. Our path to better science in less time using open data science tools. Nature Ecology and Evolution 1: 0160.

Lubchenco J, Olson AM, Brubaker LB, et al. 1991. The Sustainable Biosphere Initiative: An ecological research agenda: A report from the Ecological Society of America. Ecology 72: 371–412.

McKiernan EC, Bourne PE, Brown CT, et al. 2016. Point of view: How open science helps researchers succeed. eLife 5: e16800.

Mills JA, Teplitsky C, Arroyo B, et al. 2015. Archiving primary data: Solutions for long-term studies. Trends in Ecology and Evolution 30: 581–589.

National Research Council. 2015. Enhancing the Effectiveness of Team Science. National Resaerch Council, The National Academies Press.

Novick KA, Biederman JA, Desai AR, et al. 2017. The AmeriFlux network: A coalition of the willing. Agricultural and Forest Meteorology 249: 444–456.

Oliver SK, Fergus CE, Skaff NK, et al. 2018. Strategies for effective collaborative manuscript development in interdisciplinary science. Ecosphere 9: e02206.

O'Reilly CM, Sharma S, Gray DK, et al. 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophysical Research Letters 42: 2015GL066235.

O'Rourke M, Crowley S, Eigenbrode SD, Wulfhorst JD (eds). 2013. Enhancing Communication and Collaboration in Interdisciplinary Research. SAGE.

Palmer MA, Bernhardt ES, Chornesky EA, et al. 2005. Ecological science and sustainability for the 21st century. Frontiers in Ecology and the Environment 3: 4–11.

Parker TH, Forstmeier W, Koricheva J, et al. 2016. Transparency in ecology and evolution: Real problems, real solutions. Trends in Ecology and Evolution 31: 711–719.

Parr CS, Cummings MP. 2005. Data sharing in ecology and evolution. Trends in Ecology and Evolution 20: 362–363.

Pennington DD. 2011. Collaborative, cross-disciplinary learning and co-emergent innovation in eScience teams. Earth Science Informatics 4: 55–68.

Petchey OL, Pontarp M, Massie TM, et al. 2015. The ecological forecast horizon and examples of its uses and determinants. Ecology Letters 18: 597–611.

Peters DPC, Haversustad KM, Cushing J, et al. 2014. Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. Ecosphere 5: 1–15.

Porter JH, Hanson PC, Lin C-C. 2012. Staying afloat in the sensor data deluge. Trends in Ecology and Evolution 27: 121–129.

Read JS, Gries C, Read EK, et al. 2016a. Generating community-built tools for data sharing and analysis in environmental networks. Inland Waters 6: 637–644.

Read EK, O'Rourke M, Hong GS, et al. 2016b. Building the team for team science. Ecosphere 7: e01291. doi:10.1002/ecs2.1291.

Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015. Public data archiving in ecology and evolution: How well are we doing? PLOS Biology 13: e1002295.

Rose KC, Graves RA, Hansen WD, et al. 2017. Historical foundations and future directions in macrosystems ecology. Ecology Letters 20: 147–157.

Schimel DS, Asner GP, Moorcroft P. 2013. Observing changing ecological diversity in the Anthropocene. Frontiers in Ecology and the Environment 11: 129–137.

Schimel D, Keller M. 2015. Big questions, big science: Meeting the challenges of global ecology. Oecologia 177: 925–934.

Sharma S, Gray DK, Read JS, et al. 2015. A global database of lake surface temperatures collected by in situ and satellite methods from 1985–2009. Scientific Data 2: 150008.

Soranno PA, Bacon LC, Beauchene M, et al. 2017. LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. lakes. GigaScience 6: 1–22.

Soranno PA, Bissell EG, Cheruvelil KS, et al. 2015a. Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. GigaScience 4: 28.

Soranno PA, Cheruvelil KS. 2017. LAGOS-NE-LIMNO v1.087.1: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. Lakes: 1925–2013. http://dx.doi.org/10.6073/pasta/b1b93ccf3354a7471b93eccca484d506.

Soranno PA, Cheruvelil KS, Elliott KC, Montgomery GM. 2015b. It's good to share: Why environmental scientists' ethics are out of date. BioScience 65: 69–73.

Soranno PA, Schimel DS. 2014. Macrosystems ecology: Big data, big ecology. Frontiers in Ecology and the Environment 12: 3–3.

Stokols D, Misra S, Moser RP, et al. 2008. The ecology of team science. American Journal of Preventive Medicine 35: S96–115.

Thessen A. 2016. Adoption of machine learning techniques in ecology and Earth science. One Ecosystem 1: e8621.

Turner MG, Gardner RH. 2015. Landscape Ecology in Theory and Practice: Pattern and Process. Springer.

Uriarte M, Ewing HA, Eviner VT, Weathers KC. 2007. Constructing a broader and more inclusive value system in science. BioScience 57: 71–78.

Wenger E. 2000. Communities of practice and social learning systems. Organization 7: 225–246.

Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: sdata201618.

Wolkovich EM, Regetz J, O'Connor MI. 2012. Advances in global change research require open science by individual researchers. Global Change Biology 18: 2102–2110.

Woolley AW, Chabris CF, Pentland A, et al. 2010. Evidence for a collective intelligence factor in the performance of human groups. Science 330: 686–688.

Wuchty S, Jones BF, Uzzi B. 2007. The increasing dominance of teams in production of knowledge. Science 316: 1036–1039.

Kendra Spence Cheruvelil is a professor in Lyman Briggs College and the Department of Fisheries and Wildlife and Patricia A. Soranno is a professor in the Department of Fisheries and Wildlife, at Michigan State University, in East Lansing. The two authors contributed equally to the conceptualization and writing of this article.