Provoking a Cultural Shift in Data Quality

SARAH E. MCCORD, NICHOLAS P. WEBB, JUSTIN W. VAN ZEE, SARAH H. BURNETT, ERICA M. CHRISTENSEN, ERICHA M. COURTRIGHT, CHRISTINE M. LANEY, CLAIRE LUNCH, CONNIE MAXWELL, JASON W. KARL, AMALIA SLAUGHTER, NELSON G. STAUFFER, AND CRAIG TWEEDIE

Ecological studies require quality data to describe the nature of ecological processes and to advance understanding of ecosystem change. Increasing access to big data has magnified both the burden and the complexity of ensuring quality data. The costs of errors in ecology include low use of data, increased time spent cleaning data, and poor reproducibility that can result in a misunderstanding of ecosystem processes and dynamics, all of which can erode the efficacy of and trust in ecological research. Although conceptual and technological advances have improved ecological data access and management, a cultural shift is needed to embed data quality as a cultural practice. We present a comprehensive data quality framework to evoke this cultural shift. The data quality framework flexibly supports different collaboration models, supports all types of ecological data, and can be used to describe data quality within both short- and long-term ecological studies.

Keywords: data quality, quality assurance, quality control, big data, data, ecoinformatics

In the past two decades, ecology has begun a transformation toward open science (Hampton et al. 2013). Remote-sensing platforms, *in situ* sensor networks, monitoring networks, and community science initiatives have all contributed to an explosion in the kinds, amounts, and frequency of environmental data that are publicly available (Farley et al. 2018). This surge in ecological data is led by collaborative efforts such as the National Ecological Observatory Network (NEON), the US Long Term Ecological Research Network (LTER), the US Bureau of Land Management's Assessment Inventory and Monitoring strategy (BLM AIM), and the US National Phenology Network. The availability of new data streams via monitoring networks, data repositories, and aggregators (e.g., DataOne, Global Biodiversity Information Facility, FLUXNET), provide opportunities to understand ecosystem processes in new ways (Poisot et al. 2016, White et al. 2019). Data availability and new ecosystem research approaches are also facilitating an increase in transdisciplinary, interagency, and remote collaborations (e.g., Webb et al. 2016) and new subdisciplines such as macroecosystem ecology and ecological forecasting are developing rapidly (Poisot et al. 2016, Dietze et al. 2018). Advances in data integration and modeling in collaboration with community scientists and land managers provide new opportunities to synthesize, predict, test, and revise our understanding of ecosystems across spatial and temporal scales (Campbell et al. 2016, Dietze et al. 2018, Peters et al. 2018, Carter et al. 2020). Specific advances include integrating community science phenology observations into models seeking to understand vegetation responses to climate change (Taylor et al. 2019) and broadscale standardized rangeland monitoring programs that inform land management decisions at local and national scales (Toevs et al. 2011). However, these advances bring new challenges for ecological studies and data-driven decision-making.

Improving and developing new analysis techniques is not possible without quality data, which in turn can improve ecological models (e.g., Webb et al. 2016) and forecasts (e.g., Taylor et al. 2019, White et al. 2019). Addressing data quality extends beyond improving data management to the broader ways in which ecologists interact with data. Concerns of reproducibility and replicability are heightened as data complexity increases and ecologists are using new kinds of data (Bond-Lamberty et al. 2016, Powers and Hampton 2019). Whereas high quality data sets are celebrated jewels within the ecological community, erroneous data sets become increasingly problematic as errors propagate across scales, users, and applications (Foster et al. 2012). For example, Van Niel and Austin (2007) found errors in digital elevation models propagated in vegetation habitat models that undermined model accuracy for predicting rainforest tree cover. Typically, approaches to managing data quality are developed in small-team settings that rely heavily on interpersonal trust and tools such as lab notebooks. However, because there is not a tradition of developing data quality

BioScience 71: 647–657.
doi:10.1093/biosci/biab020 Advance Access publication 31 March 2021

approaches in a consistent way, data quality practices developed in small research team settings do not scale well to large data repositories, networked monitoring, and large collaborative research efforts (Farley et al. 2018). Similarly, data quality approaches that are successful for large, networked data collection efforts (e.g., NEON, LTER, BLM AIM) rely on dedicated data management staff who may not be available in small research teams (Laney et al. 2015). Breakdowns in data quality management can have dire consequences for the rigor of inferences drawn from data analyses, our understanding of ecosystems, and the predictive power of models and their uncertainty (Beck et al. 2014). Such breakdowns can also increase the risk of ill-conceived data-driven management decisions. For instance, Vauhkonen (2020) found that tree-level inventories derived from airborne methods underdetect small trees and, therefore, underpredict harvest profits, resulting in misleading future profit expectations for managers. Similarly, Brunialti and colleagues (2012) demonstrated limited comparability of lichen diversity estimates because of variability in protocol interpretation, data collector skill sets, and training procedures, which resulted in a restricted ability to monitor changes in lichen biodiversity in response to ecological drivers that would inform management. As the diversity and volume of data and ecological analyses increases, ecology needs to adopt both cultural and technological frameworks to improving and ensuring data quality throughout the data lifecycle.

Fortunately, there are a plethora of technical solutions available to improve data quality, made possible by advances in hardware and software that have increased both data storage capacity and processing speeds (Goda and Kitsuregawa 2012). Electronic data capture, which reduces data transcription and management errors, is now standard for both sensor systems and observational programs through customizable mobile applications platforms (e.g., ODK, Fulcrum, ESRI Survey123). Programming and automation tools, such as R and Python, are now readily available to ecologists with a relatively low barrier of entry thanks in part to the Data and Software Carpentries (Teal et al. 2015, Wilson 2016) and other data and code training programs. These software tools increase the speed of data examination, cleaning, and error evaluation. As a result, ecologists can automate traditionally error-prone aspects of the data workflow by restricting data entry to valid ranges and enabling on-the-fly analysis (Yenni et al. 2019). The development of reproducible computing frameworks, including Jupyter Notebooks and R Markdown, and containerization (e.g., Docker, Singularity), allows ecologists to track and easily share analysis processes, thereby reducing errors when replicating analyses (Peng 2011). Standards such as the Ecological Metadata Language, repositories such as the Environmental Data Initiative, and aggregators such as DataOne provide an opportunity for documenting and archiving data long after collection (Fegraus et al. 2005, Michener et al. 2012). For example, NEON uses the Fulcrum app for standard, electronic data collection of observational data, and R scripts managed in Docker containers to automate sensor data processing (Metzger et al. 2019). Cleaned NEON data are then published along with metadata to a data portal.

Technology integration to improve data quality is possible in large organizations and data collection efforts that have dedicated resources to build organized workflows. However, in smaller projects (e.g., long-tail science; Laney et al. 2015), implementing these technologies in a coordinated approach to manage data quality can still be overwhelming without an overarching cultural framework to inform who, how, and why to best implement different technical solutions. It is the experience of the authors in working with NEON, LTER, BLM AIM, and long-tail science data that there is uneven adoption of technologies to prevent errors and few processes available for correcting errors in source data sets, even if they are resolved prior to analyses. Given the rapid growth of data collection, the rising prominence of data aggregation through repositories, and the call for improved synthetic studies that draw from data integration efforts, there is an urgent need for all ecologists (scientists, academics, data managers, data collectors, students) to adopt a more comprehensive framework that incorporates both technological and cultural data quality practices.

Data quality is foundational to improving trust and ensuring the legacy of current ecological research and optimizing management. Following a review of the current data quality approach, encapsulated in the DataOne data lifecycle, we present a conceptual data quality framework that explicitly identifies quality assurance and quality control steps to improve data quality across a range of collaboration models, data types, and ecological studies. Although some of the topics discussed in the present article may be familiar to data managers, designated data managers may not be available in every lab or research partnership (Laney et al. 2015). Data quality is an issue that concerns all ecologists, not just data managers, so we address how all members of a team, regardless of career stage, can participate in improving data quality throughout the data lifecycle. We also discuss, for the benefit of all ecologists, how the framework can be applied to evaluate data quality roles within the data lifecycle and how approaches for ensuring data quality differ among data types. Finally, we explore how the data quality framework can be used to evaluate data quality over time to improve our ability to detect and understand ecosystem trends.

Current data quality approach

The current data quality approach in ecology is focused on improving information management via the data lifecycle, which describes how data are created, preserved, and used. The DataOne lifecycle (figure 1), which includes steps for planning, collecting, assuring, describing, preserving, discovering, integrating, and analyzing, is a common data management approach embraced in ecology (Michener and Jones 2012). Many funding agencies, including the US National Science Foundation, now require data management plans that specifically address the DataOne lifecycle.

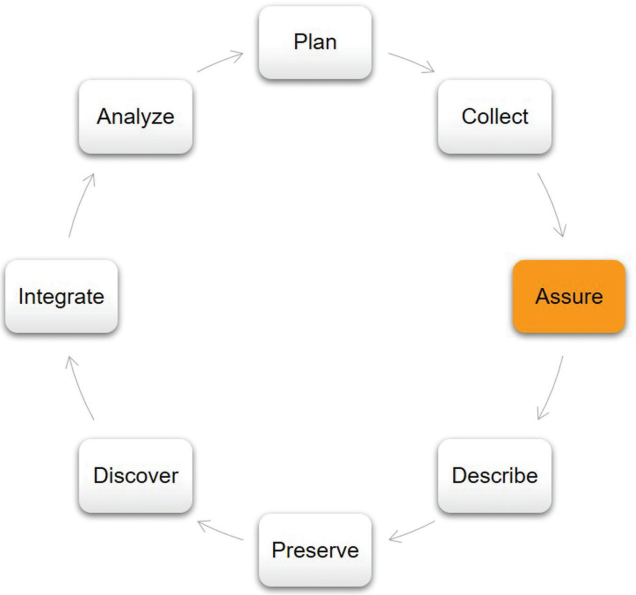

Figure 1. Traditional data lifecycle diagrams isolate quality assurance and quality control at a single stage as Assure or quality assurance/quality control in (QA/QC) the data workflow, generally following data collection. Modified from the DataOne lifecycle (Michener et al. 2012).

Simultaneously, ecologists have developed best practices for navigating the data lifecycle, including building data management plans (Michener 2015), data sharing and reproducibility (White et al. 2013, Powers and Hampton 2019), data reformatting or creating tidy data (Wickham 2014), scientific computing (Wilson et al. 2014, 2017), and working with community scientists (Kosmala et al. 2016). The DataOne lifecycle provides a useful organizational structure for how data moves through the research life cycle. The benefit is that it illustrates how data can be shared through repositories (preserve) and so encourages broader collaboration, use and reuse of data. However, the DataOne life cycle was developed in an era in which broad data sharing was new and it does not capture the extent of active data quality processes needed to support data transfer from one ecologist to another. In the current data sharing environment, the approach of relying on institutional knowledge of data quality processes during a single assurance step is no longer sufficient for ensuring data quality. In the collective experience of the authors, the DataOne lifecycle does not reflect successful data quality practices used by many ecologists such as reviewing data for errors prior to analysis. Therefore, it has become increasingly important for everyone to play a role in ensuring data quality throughout the data lifecycle. A central issue in modernizing the DataOne lifecycle is the need to expand how quality assurance and quality control processes are incorporated into ecological data culture in a coordinated manner that expands on current successful data quality practices and applications of technology.

The principles of quality assurance and quality control can provide a framework for organizing appropriate tools and technologies to ensure data quality. Quality assurance is an active anticipatory process to minimize the chance of an error being inserted into data (Herrick et al. 2018, Michener 2018). Conversely, quality control is a reactive process to detect, describe, and, if possible, address inaccuracies that occur at any point in the data lifecycle (Herrick et al. 2018, Michener 2018). The desired outcome of quality assurance is fewer errors in data or analysis products, whereas quality control provides active validation of quality within data or analysis products, documentation and correction of errors, and an account of any errors that may remain (Zuur et al. 2010). Quality assurance is a continuous process throughout the scientific method and data lifecycle (Herrick et al. 2018, Michener 2018). Data management, written protocols, training, and calibration steps are all components of quality assurance. The driving questions of quality assurance include the following: *What could go wrong? How will we prevent errors? How will we address errors when they do occur?* Quality assurance tasks are often similar among ecological subfields, projects, data types, and career stages. In contrast, quality control tasks are often discipline specific, asking whether the data are complete, correct, and consistent. If the answer is no, then steps are taken to address those issues if possible. Quality control tasks occur at distinct points within the data lifecycle, including immediately after data collection, during archiving, and prior to analysis. Quality control tasks can often be automated to detect missing data and flag erroneous values (Rüegg et al. 2014, Yenni et al. 2019).

The current data quality paradigm, encapsulated in the DataOne lifecycle, inadequately incorporates quality assurance and quality control as it aggregates and isolates quality assurance and quality control to a single Assure or quality assurance/quality control (QA/QC) step within the data lifecycle (figure 1; Michener and Jones 2012, Rüegg et al. 2014). The single Assure stage emphasizes data quality associated directly with data collection, but fails to properly acknowledge opportunities for preventing, introducing, detecting, and addressing errors at other stages of the data lifecycle. Although the data manager and the data collector in the data lifecycle certainly have a responsibility for data quality, every individual who interacts with data has an opportunity to improve or degrade data quality. A new framework would encourage all ecologists and land managers, who increasingly rely on found data and may not have a personal relationship with the study initiators or data collectors (e.g., Poisot et al. 2016) to participate in ensuring data quality.

The second issue with isolating quality assurance and quality control as a discrete step in the data management lifecycle is that quality assurance and quality control are easily conflated. The current framework misses unique opportunities to prevent and detect errors throughout the data lifecycle by treating quality assurance and quality control as a single process. For example, a principal investigator adds a new

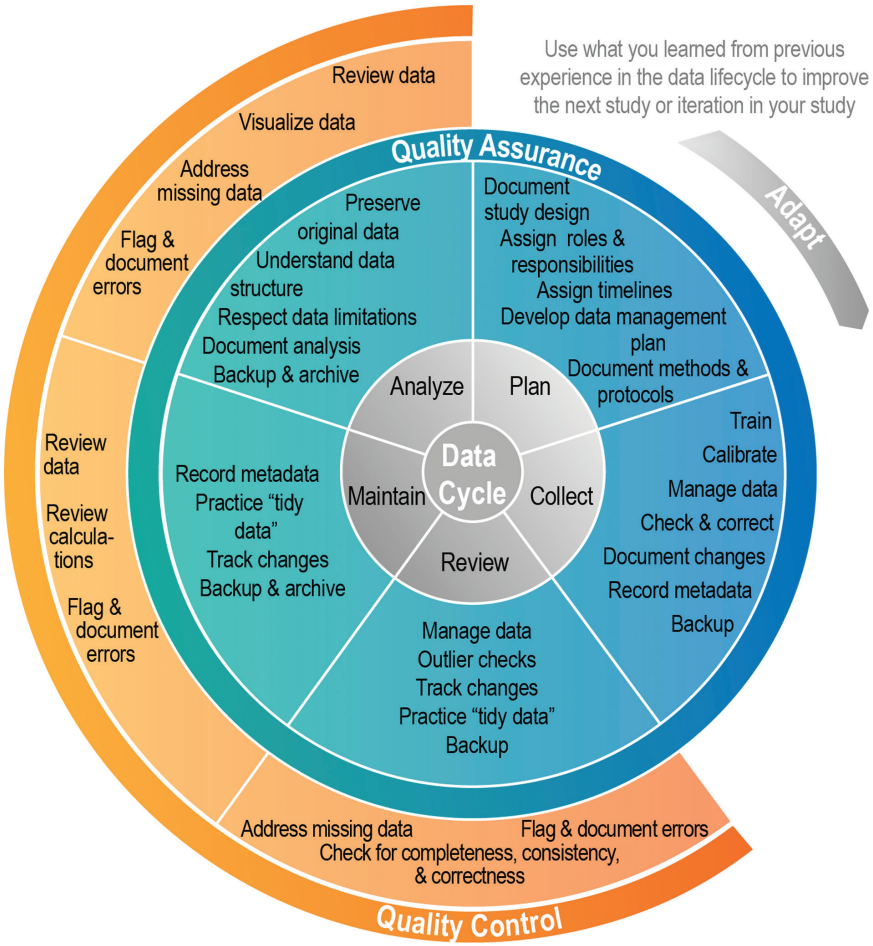

Figure 2. The quality assurance and quality control (QA&QC) framework, which follows the data lifecycle (inner circle) with explicit quality assurance and quality control incorporated at each stage. Quality assurance (middle circle) is a continuous process, with explicit steps at each stage of the data lifecycle. Quality control (outer half circle) processes begin after data are collected. For simplicity we have only identified five lifecycle stages. However, this framework can easily be expanded or contracted to accommodate a different number of lifecycle stages (e.g., figure 1; Michener et al. 2012).

species cover method to a study at the last minute. The data management plan is not updated to include this data type in the study, and the data collectors improvise a data sheet in the field that inadvertently omits key data elements. When the data are digitized, the handwritten data sheet is difficult to read, so a species name is incorrectly entered. The original data collector has left the team and the transcription error is not caught during quality control. The data manager uploads the data to a repository without documentation of the data type in the data management plan and the incorrect version of the field protocol document. The data user discovers the data set and makes an additional data processing error that leads the data user to believe the data is another kind of data (e.g., species presence rather than species cover), and incorrectly parameterizes a model. In this hypothetical study, the DataOne lifecycle accurately describes how the data moved; however, every team member made an error of omission or commission that was not caught during quality control. Communicating data quality steps and detecting gaps in data quality is difficult, especially in large, transdisciplinary teams. The consequences of such errors include erroneous conclusions (Morrison 2016), lack of reproducibility (Peng 2011, Powers and Hampton 2019), retraction (Evaristo and McDonnell 2020), and effects on management decisions (Vauhkonen 2020). A comprehensive data quality approach is needed to adequately represent both technological and cultural aspects of producing and maintaining high quality ecological data.

Effectively separating quality assurance and quality control and ensuring that data quality processes are implemented more widely than the single assurance step requires broader changes than simply splitting quality assurance and quality control within the DataOne lifecycle. These changes include the need to identify successful cultural and technological data quality practices and where they are most appropriately applied, clearly articulate roles and responsibilities for data quality practices beyond the data collector and data manager, and establish approaches for describing data quality shortcomings, reviewing weaknesses as a team, and working to improve existing and future data sets. A cultural change in data quality requires a supporting framework that evolves the DataOne life cycle from a mechanistic description of data movement (e.g., data collector to data repository) to a set of community actions that all ecologists can participate in to ensure data quality.

An improved data quality framework

Although the DataOne lifecycle and other technological advances have improved data quality in the realm of information management, a framework is needed that identifies successful data quality practices, supports research collaboration culture, and addresses all aspects of the research and resource management lifecycle. We present a quality assurance and quality control framework (QA&QC) that builds on previous advances but explicitly considers quality assurance and quality control as distinct and important processes that encompass the data lifecycle (figure 2). In this framework, quality assurance scaffolds the entire data lifecycle to reduce errors from planning to analysis. Quality

control begins after data are acquired and follows both quality assurance and the data lifecycle from data review to analysis. Although we identify example quality assurance and quality control tasks in figure 2, the quality assurance and quality control framework is largely conceptual to provoke discussion among ecologists about how to prevent, detect, and document errors at every data lifecycle stage.

The quality assurance and quality control framework provides a collaborative communication tool to identify data quality actions and improve data-driven ecological research and management. Ecologists can use the framework as an assessment tool to document the relative effort or infrastructure currently in place for their study and to isolate vulnerabilities within current data workflows. The quality assurance and quality control framework can improve the rigor of ecological research and strengthen collaborations by identifying required data quality steps and who will execute those steps throughout the data lifecycle. This framework can also be used to communicate how data quality workflows differ among data types. The final benefit of the framework is that it can be applied retroactively to describe which quality assurance and quality control steps have or have not been taken in longitudinal and found data sets.

Data quality through roles and responsibilities. Ecology is an increasingly collaborative and transdisciplinary science. Although each team member who interacts with data has an opportunity to influence data quality, each person who interacts with data is not equally responsible for both quality assurance and quality control at every stage of the data lifecycle. The quality assurance and quality control framework enables ecologists to examine how quality assurance and quality control responsibilities differ by role within a lab group, interdisciplinary collaboration, or national monitoring program (box 1). Project leaders or principal investigators oversee data quality at all levels and ensure that adequate plans are developed to maintain data quality (figure 3). These tasks may include planning data collection and error checking timelines, organizing observer training and calibration, ordering and calibrating field equipment and sensors, and sample design preparation. The data collector is primarily focused on preventing errors during the data collection and review stages. The data manager is typically engaged with all stages of the data workflow and ensures that adequate data management is planned, verifies that other team members know how to interact with the data management systems, and conducts data review. Analysts lead the final review of the data and maintain error free analysis and interpretation.

The advantage of conceptualizing quality assurance and quality control tasks by roles is that the framework enables communication between roles and leadership and enables opportunity for iterative improvement. For instance, the quality assurance and quality control framework clearly communicates to project leaders that they have responsibility for data quality and oversight at each level of the data workflow (figure 3). Expressing quality assurance and quality control roles through the quality assurance and quality control framework (figure 3a) demonstrates the value of the data management team who plays a critical role in ensuring data quality at all stages. If there are breakdowns in data quality during one field season, the framework can be used to identify communication improvements among personnel or if additional personnel are needed to maintain data workflow and data quality. Although not every team or partnership may have a fulltime data manager, analyst, or data collector, we encourage ecologists to identify the individual who will take on those tasks. Formalizing roles and responsibilities for data quality with this framework is applicable to teams of any size that collect, manage, or analyze data. Successful implementation of this framework will build a culture in which all team members are continuously applying quality assurance and quality control to every aspect of the data lifecycle.

A data quality workflow for different data types. Ecologists often use a mixture of sensor and observational data to understand ecosystem processes. In repeated observational studies (e.g., the North American Breeding Bird Survey), in which an emphasis on quality assurance prior to data collection is critical, the current paradigm can miss opportunities to address data quality at other stages of the data lifecycle. The quality assurance and quality control framework supports developing an integrated approach to data quality that recognizes that there is no global quality assurance and quality control protocol for all data types. Quality assurance is a common element through planning, calibration, and training of the data collection team in observational studies, sensor networks, and remote sensing platforms (box 2). However, there are differences in the amount of quality assurance and quality control effort required between these data types. In observational studies, quality assurance through training and calibration is the primary opportunity to reduce errors, whereas there are few opportunities during quality control (Sauer et al. 1994). Sensors require equal quality assurance and quality control efforts to prevent, detect, and correct anomalous readings (Sturtevant et al. 2018). Differentiating data quality practices by data type is not only important in data collection and data curation but also during analyses in which preprocessing steps, outlier checks, and pathways to resolving errors vary. The quality assurance and quality control framework formalizes management and documentation of different data types, preventing data quality lapses that can have significant financial and scientific costs (e.g., Hossain et al. 2015).

Understanding data quality in longitudinal data. Understanding ecosystem change in response to climatic and anthropogenic drivers is a major focus of contemporary ecological research. Changes in observers or sensors, incomplete digitization, and shifting data management practices can affect apparent trends (box 3). Therefore, it is critical to identify where data quality influences variability in longitudinal studies, to describe how shifts in data management might mitigate

Box 1. Using quality assurance and quality control to manage roles and responsibilities in the Bureau of Land Management's Assessment, Inventory, and Monitoring Program.

One example of how roles and responsibilities vary is in national monitoring programs. The BLM AIM program is a standardized monitoring program that collects data across dryland, aquatic, and wetland ecosystems on federal lands in the United States (figure 3; Toevs et al. 2011). Each year, 3000–5000 monitoring locations are sampled through a federated data collection effort (figure 3b). Sampling is conducted by approximately 400 data collectors and managed by 150 local project leaders at BLM field offices. These project leaders are coordinated through one of 20 monitoring coordinators located at BLM state or regional offices. A national BLM AIM team of natural resource scientists, data managers, analysts, and statisticians manage centralized training, data collection workflows, data management, and support analyses at national, regional, and local scales. Ensuring data quality across all individuals involved in AIM data collection and management is successful because the program clearly articulates the role of each individual who interacts with the data, works to ensure that those individuals are aware and equipped to complete their data quality responsibilities, and iterates on the basis of feedback from team members (figure 3a; Bureau of Land Management 2020). Although not all ecological teams will operate at the scale of the BLM AIM team, the process for clearly identifying team members' roles and ensuring that team members are supported with training and resources to complete their data quality-related tasks can be extended to every ecological team and collaboration.

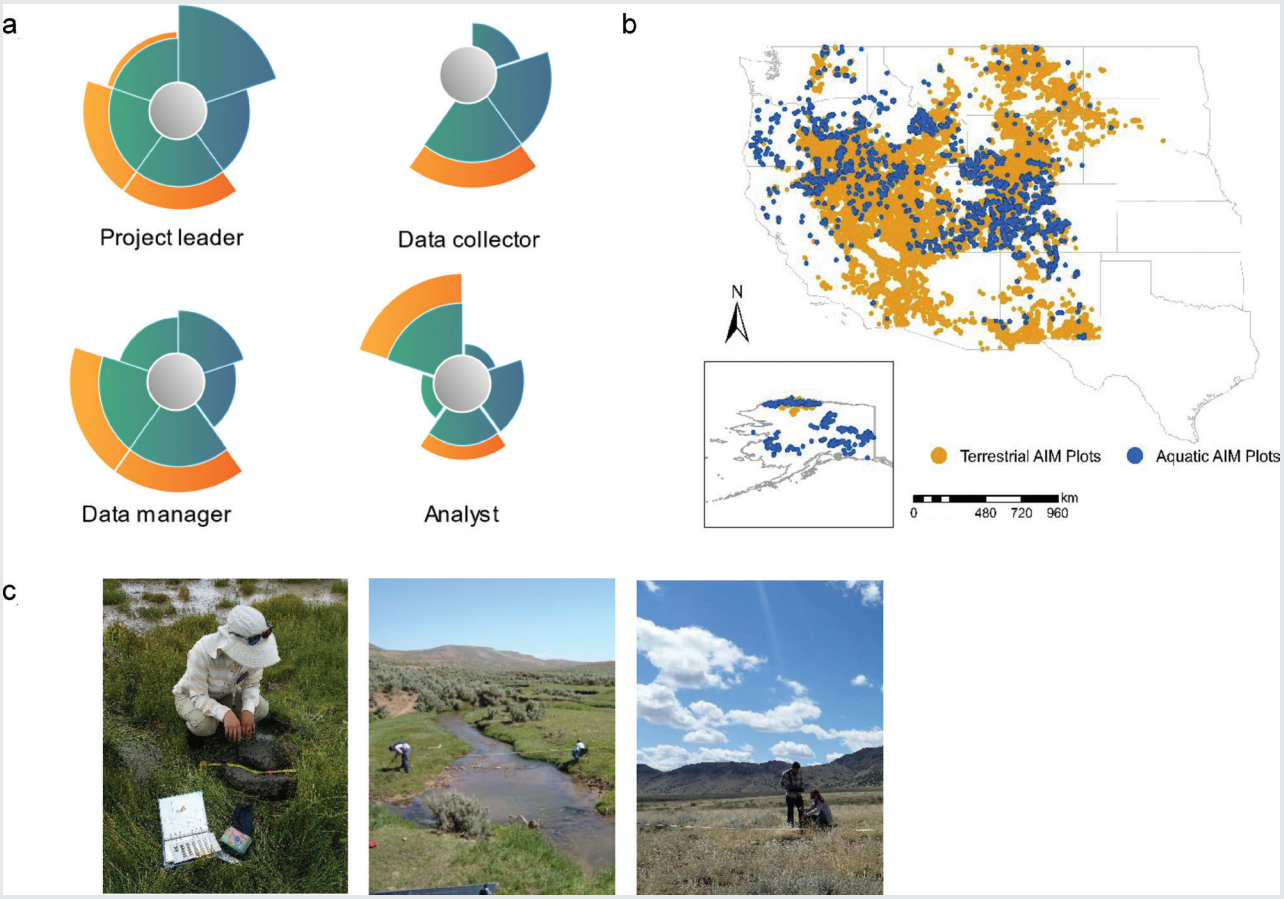

Figure 3. Comparison of data quality roles and responsibility by team member within the quality assurance and quality control framework for the BLM AIM program (a). Because of this collaboration between project leads, data managers, data collectors, and analysts, over 35,000 monitoring locations have been sampled since 2011 (b) in wetland, aquatic, and terrestrial ecosystems (c). Refer to figure 2 for a description of the lifecycle represented in panel (a). Photograph: Bureau of Land Management.

Box 2. Understanding quality assurance and quality control for different data types.

The US National Science Foundation's National Ecological Observatory Network (NEON) is a long-term, continental scale ecological monitoring effort of 81 terrestrial and aquatic sites across the United States (Keller et al. 2008). At each NEON site, biological, chemical, and physical data are collected through monthly observational sampling, continuous *in situ* instrument systems, and from an airborne observation platform (figure 4). NEON collects and manages over 175 data products along with more than 100,000 biological, genomic, and environmental samples collected each year. Although each data type requires different quality assurance and quality control approaches, each system follows the same operational data lifecycle, requiring careful planning and calibration, data collection, initial review, data maintenance, and publication on the NEON Data Portal for open access use in ecological analysis (figure 4b; Sturtevant et al. 2018). NEON also promotes analysis quality assurance through a training series that facilitates the exploration and analysis of NEON data. The challenges of collecting, managing, and using more than one kind of data are common throughout ecological research and land management. Ecologists will benefit from NEON's approach of identifying core data and quality assurance and quality control procedures, but then building parallel workflows that are specific to each data type. When the data are brought together in analysis, it is particularly important that data users understand the differences in data structures and how data errors might manifest differently among data types.

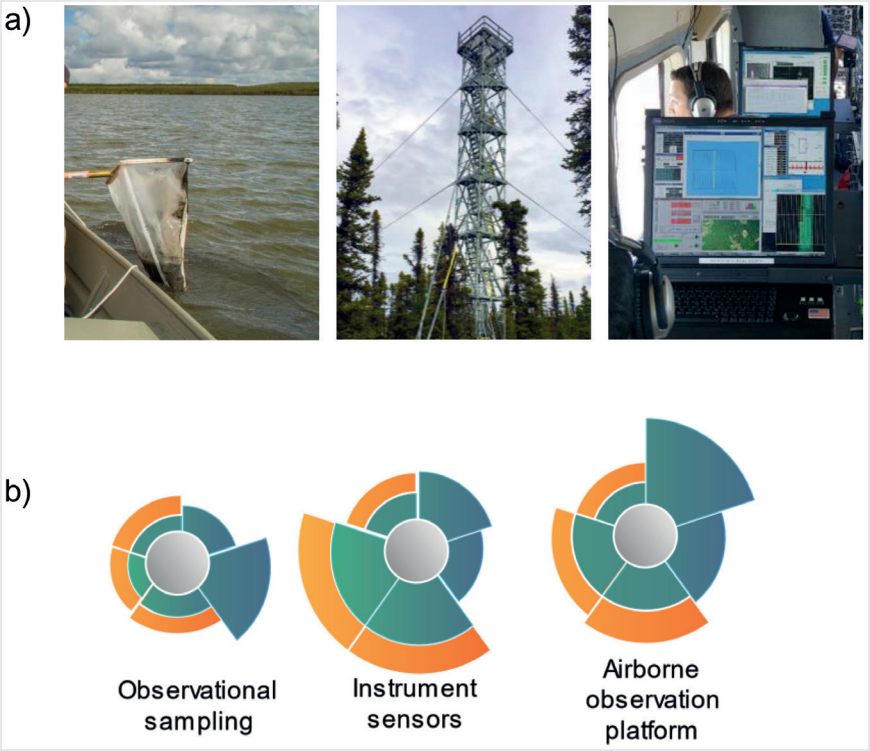

Figure 4. Three types of data are collected at NEON sites, observational, sensor, and airborne remote sensing (a). Each data system follows the same general data lifecycle, including careful planning and calibration, data collection, initial review, data maintenance, analysis, and publication. However, the amount of quality assurance and quality control applied at each step varies by data type (b). Refer to figure 2 for a description of the lifecycle represented in panel (b). Photograph: National Ecological Observatory Network

issues, and to provide detailed documentation to accompany the data. Data providers can use the quality assurance and quality control framework to detect and describe data quality shifts through the data lifespan, whereas data users might leverage the framework to evaluate data for errors, structural problems, and other issues affecting data quality. Often these shifts are known to individuals on the project but not easily accessed by new collaborators. Using the quality assurance and quality control framework, an evolving team can proactively reduce or even eliminate knowledge gaps due to personnel turnover. Detailed lab notes and records are valuable in documenting shifts in data quality, but the quality assurance and quality control framework offers an approach to synthesize the data quality history. Without quality assurance

Box 3. Applying the quality assurance and quality control framework to understand longitudinal data quality.

Consistent application of quality assurance and quality control is especially critical for long-term ecological research. The Jornada Quadrat study (figure 5) is a long-term vegetation study of 122 quadrats established to investigate livestock grazing effects on plant community dynamics as well as vegetation responses to variable climatic conditions in the Chihuahuan Desert (Chu et al. 2016). Quadrats were charted consistently from 1915 to 1947, with only a portion of the quadrats charted intermittently between 1947 and 1979. Sampling resumed in 1995 and continues every 5–6 years (figure 5b). As data collectors change and technology evolves throughout the study, examples of quality assurance and quality control successes and challenges were found during repeat sampling efforts, digitizing historical data sheets, and analyzing long-term trends.

Data quality has varied across the Jornada Quadrat study. An effort is underway to flag data quality issues in the data set to help inference limitations and assumptions necessary in future analyses (figure 5a). Between 1915 and 1947, quality assurance included laying out the sample design and developing a consistent method for charting. Known quality control steps were limited to tracking the chain of custody for errors between data collectors and documented error checking. Quadrat sampling from 1947 to 1979 was sporadic and data quality during this period is the poorest in the record. Woody species cover fluctuated dramatically, which is highly unlikely given shrub encroachment records from the same period (figure 5c). Since 1995, stricter protocols for sampling the quadrats have been implemented and documented. The same set of data collectors have recorded information since 2001, therefore interobserver variability is the lowest for this period of the overall data set. Future data collection events will follow the newly developed documentation to minimize observer variability.

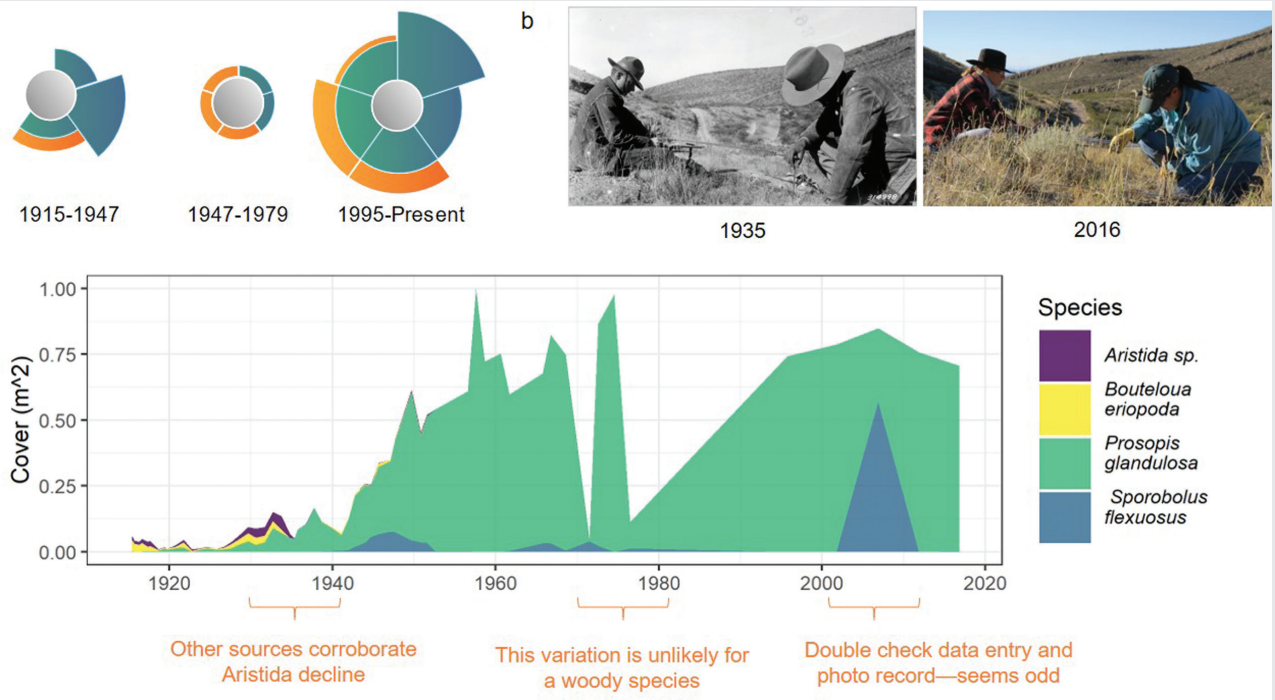

Figure 5. The Jornada Quadrat study is an ongoing longitudinal study of vegetation pattern and trends from 1915 to present. Data quality has varied throughout the data set (a) as different data collectors and data managers participated in the study (b). This has resulted in anomalies in the data set, including an unlikely decline and increase in* Prosopis glandulosa *(c). Refer to figure 2 for a description of the lifecycle represented in panel (a). Photograph: USDA-ARS Jornada Experimental Range.

and quality control documentation published alongside data in repositories, data sets may be lost entirely or become unusable in future ecological research (Laney et al. 2015). This is a significant cost to the ecological community, in terms of wasted resources and unnecessary information gaps critical to understanding rapidly changing ecosystems. Evaluating longitudinal data through the quality assurance and quality control framework will enable data strengths and weaknesses to be communicated to the ecological community to support the use of valuable long-term data sets.

How can ecologists adapt to improve data quality?

In every data set, there are opportunities for ecologists to improve data quality. By working through the quality

assurance and quality control framework, ecologists can identify strengths and weaknesses in their data lifecycle and opportunities for iterative improvement. An assessment of roles and responsibilities may reveal gaps or unbalanced workloads in ensuring data quality. The increasingly integrative nature of ecology means that developing a quality assurance and quality control workflow for one data type may spark ideas for improving another. For example, the standard error checking processes common in sensor data (Rüegg et al. 2014) can be adapted to observational data lifecycles (Yenni et al. 2019). In ongoing longitudinal studies and network research programs, improvements in quality assurance and quality control can be directly applied to the next data collection cycle and to future studies. Future software and hardware advances may change how we interact with data and conduct ecological analyses, which are likely to affect the scientific culture of using data and ensuring data quality. This will require iterative improvement of data workflows, training resources, education, and communication media. Adjusting to these technology shifts is an opportunity to evaluate and document the current data quality regime (box 3) before adopting new hardware and software.

The iterative nature of data quality is a cultural value that the ecological community should embrace. As a data-driven science, we can work to improve the quality of the data that are advancing the field of ecology. We encourage ecologists to use the quality assurance and quality control framework to evaluate their data sets and ecological studies, from planning through analysis. Grant proposal guidelines could provide adequate space for applicants to address quality assurance and quality control, in addition to data management. Project status reports might include data quality issues found during data collection, storage, and analysis and might describe how those issues were overcome. Data users who leverage ecological repositories and other sources of found data can use the quality assurance and quality control framework during initial data exploration to clearly identify data types, describe data provenance, and document assumptions that might affect data quality and subsequent analyses.

Current ecological education could be expanded to include frequent discussions of quality assurance and quality control. For instance, data education resources, such as the Data and Software Carpentries (Teal et al. 2015, Wilson 2016) can include the quality assurance and quality control framework in their data modules together with technical solutions (e.g., coding, reproducibility, data management). In the academic realm, lab exercises could include a reflection section encouraging students to identify what went well and what could be improved from a data quality perspective. In exercises in which data are provided, students should be encouraged to ask questions about the data quality history, structure, and how known errors might affect their results and interpretation. If different kinds of data are presented in a university course, students could be encouraged to compare and contrast data quality challenges and successes among data sets as a final exercise. We also encourage graduate students and advisors to build quality assurance and quality control into graduate education culture, which might include data quality as a topic in reading group discussions, requiring a quality assurance and quality control plan as part of graduate research proposals, and asking thesis defense questions that require students to reflect on quality assurance and quality control. Finally, we call on postdoctoral fellows and faculty to facilitate a supportive data quality culture in which making mistakes is normalized as a learning tool and all members of a lab work together to prevent and correct errors. Expanding ecological education to include the quality assurance and quality control framework in addition to data management will equip the next generation of ecologists to harness the wealth of ecological data available to them.

Evolving the DataOne lifecycle to include the quality assurance and quality control framework, however, requires active engagement in the ecological community beyond ecological education. All ecologists, in the research and management communities, should consider building on existing data management habits by describing their quality assurance and quality control workflow as a critical component of meeting study objectives. When establishing collaborative projects, we encourage ecologists to identify and periodically revisit the quality assurance and quality control tasks and goals of their projects. It is the experience of the authors that clearly defined quality assurance and quality control duties and expectations facilitate a more inclusive environment in which new and junior team members' contributions are broadly recognized for supporting data quality (e.g., in data collection), and there is a defined process for identifying areas of improvement that the entire team should address. Whereas data quality expectations have historically been an unspoken component of ecology, adopting the quality assurance and quality control framework is one way to describe ecological data expectations within the diverse ecological community.

Conclusions

Maintaining trust within the new cultural paradigm of transdisciplinary scientific collaboration requires an effective data quality culture. Continuous quality assurance and active quality control steps need to be included in the scientific process alongside collection, management, and analysis skill sets. Although the DataOne lifecycle has unified the ecological community in preserving and sharing data, it insufficiently represents data quality workflows. The quality assurance and quality control framework presented in the present article provides a much-needed structure for all members of the ecological community to ensure data quality at every data life stage, for every data type, and throughout the lifespan of a data set. This structure enables ecologists to implement practical data quality approaches to different kinds of data, identify roles and responsibilities within a team, and evaluate and improve long-term ecological data sets. Publishing quality assurance and quality

control workflows alongside data and analysis will increase transparency in open, reproducible science thereby increasing trust in the scientific process. Although the next steps of action will be discipline, project, and data set specific, the imperative to take these steps is global. The quality assurance and quality control framework can enhance existing ecological data and collaboration approaches, reduce errors, and increase efficiency of ecological analysis thereby improving ecological research and management.

Acknowledgments

The data for this article are available on request from the corresponding author. Lauren Price developed the map for figure 3. Deana Pennington and Darren James provided valuable review of the ideas presented in the present article. We are especially grateful to the reviewers for their insightful advice that greatly strengthened the manuscript. This research was supported by the USDA NRCS (agreement no. 67-3A75-17-469) and the BLM (agreement no. 4500104319). This research was a contribution from the Long-Term Agroecosystem Research (LTAR) network. LTAR is supported by the United States Department of Agriculture. The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute. This material is based in part upon work supported by the National Science Foundation through the NEON Program. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US government.

References cited

Beck J, Böller M, Erhardt A, Schwanghart W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics 19: 10–15.

Bond-Lamberty B, Peyton Smith A, Bailey V. 2016. Running an open experiment: transparency and reproducibility in soil and ecosystem science. Environmental Research Letters 11: 084004.

Brunialti G, et al. 2012. Can we compare lichen diversity data? a test with skilled teams. Ecological Indicators 23: 509–516.

Bureau of Land Management. 2020. BLM's Terrestrial Assessment, Inventory, and Monitoring (AIM) 2020 Field Season Data Management Protocol. https://aim.landscapetoolbox.org/data-management-project-evaluation/

Campbell LK, Svendsen ES, Roman LA. 2016. Knowledge co-production at the research–practice interface: embedded case studies from urban forestry. Environmental Management 57: 1262–1280.

Carter SK, et al. 2020. Bridging the research-management gap: landscape science in practice on public lands in the Western United States. Landscape Ecology 35: 545–560.

Chu C, Kleinhesselink AR, Havstad KM, McClaran MP, Peters DP, Vermeire LT, Wei H, Adler PB. 2016. Direct effects dominate responses to climate perturbations in grassland plant communities. Nature Communications 7: 11766.

Dietze MC, et al. 2018. Iterative near-term ecological forecasting: needs, opportunities, and challenges. Proceedings of the National Academy of Sciences 115: 1424–1432.

Evaristo J, McDonnell JJ. 2020. Retraction note: global analysis of streamflow response to forest management. Nature 578: 326.

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

Fegraus EH, Andelman S, Jones MB, Schildhauer M. 2005. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America 86: 158–168.

Foster SD, Shimadzu H, and Darnell R. 2012. Uncertainty in spatially predicted covariates: Is it ignorable? Journal of the Royal Statistical Society C 61: 637–652.

Goda K, Kitsuregawa M. 2012. The history of storage systems. Proceedings of the IEEE 100: 1433–1440.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11: 156–162.

Herrick JE, Van Zee JW, McCord SE, Courtright EM, Karl JW, Burkett LM. 2018. Monitoring manual for grassland, shrubland, and savanna ecosystems 2nd ed, vol. 1. US Department of Agriculture, ARS Jornada Experimental Range.

Hossain MS, Bujang JS, Zakaria MH, Hashim M. 2015. Assessment of the impact of landsat 7 scan line corrector data gaps on sungai pulai estuary seagrass mapping. Applied Geomatics 7: 189–202.

Keller M, Schimel DS, Hargrove WW, Hoffman FM. 2008. A continental strategy for the national ecological observatory network. Frontiers in Ecology and the Environment 6: 282–284.

Kosmala M, Wiggins A, Swanson A, Simmons B. 2016. Assessing data quality in citizen science. Frontiers in Ecology and the Environment 14: 551–560.

Laney CM, Pennington DD, Tweedie CE. 2015. Filling the gaps: sensor network use and data-sharing practices in ecological research. Frontiers in Ecology and the Environment 13: 363–368.

Metzger S, et al. 2019. From NEON field sites to data portal: a community resource for surface–atmosphere research comes online. Bulletin of the American Meteorological Society 100: 2305–2325.

Michener WK. 2015. Ten simple rules for creating a good data management plan. PLOS Computational Biology 11: e1004525. https://doi.org/10.1371/journal.pcbi.1004525.

Michener WK. 2018. Quality assurance and quality control (QA/QC). Pages 55–70 in Recknagel F, Michener WK, eds. Ecological Informatics: Data Management and Knowledge Discovery. Springer. https://doi.org/10.1007/978-3-319-59928-1_4.

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. Ecological Informatics, Data Platforms in Integrative Biodiversity Research 11: 5–15.

Michener WK, Jones MB. 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology and Evolution 27: 85–93.

Morrison LW. 2016. Observer error in vegetation surveys: a review. Journal of Plant Ecology 9: 367–379.

Peng RD. 2011. Reproducible research in computational science. Science 334: 1226–1227.

Peters DPC, et al. 2018. An integrated view of complex landscapes: a big data-model integration approach to transdisciplinary science. BioScience 68: 653–669.

Poisot T, Gravel D, Leroux S, Wood SA, Fortin M-J, Baiser B, Cirtwill AR, Araújo MB, Stouffer DB. 2016. Synthetic datasets and community tools for the rapid testing of ecological hypotheses. Ecography 39: 402–408.

Powers SM, Hampton SE. 2019. Open science, reproducibility, and transparency in ecology. Ecological Applications 29: e01822. https://doi.org/10.1002/eap.1822.

Rüegg 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. Frontiers in Ecology and the Environment 12: 24–30.

Sauer JR, Peterjohn BG, Link WA. 1994. Observer differences in the north american breeding bird survey. Auk 111: 50–62.

Sturtevant C, et al. 2018. NEON Science Data Quality Plan. NEON. DOC.004104vA. NEON (National Ecological Observatory Network). https://data.neonscience.org/api/v0/documents/NEON.DOC.004104vA.

Taylor SD, Meiners JM, Riemer K, Orr MC, White EP. 2019. Comparison of large-scale citizen science data and long-term study data for phenology modeling. Ecology 100: e02568. https://doi.org/10.1002/ecy.2568.

Teal TK, Cranston KA, Lapp H, White E, Wilson G, Ram K, Pawlik A. 2015. Data carpentry: workshops to increase data literacy for researchers. International Journal of Digital Curation 10: 135–143.

Toevs, GR, Karl JW, Taylor JJ, Spurrier CS, Bobo MR, Herrick JE. 2011. Consistent Indicators and Methods and a Scalable Sample Design to Meet Assessment, Inventory, and Monitoring Information Needs Across Scales. Society for Range Management.

Van Niel KP, Austin MP. 2007. Predictive vegetation modeling for conservation: Impact of error propagation from digital elevation data. Ecological Applications 17: 266–280.

Vauhkonen J. 2020. Effects of diameter distribution errors on stand management decisions according to a simulated individual tree detection. Annals of Forest Science 77: 21.

Webb NP, et al. 2016. The national wind erosion research network: Building a standardized long-term data resource for aeolian research, modeling and land management. Aeolian Research 22: 23–36.

White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. 2013. Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution 6: 1–10. doi:10.4033/iee.2013.6b.6.f

White EP, Yenni GM, Taylor SD, Christensen EM, Bledsoe EK, Simonis JL, Morgan Ernest SK. 2019. Developing an automated iterative near-term forecasting system for an ecological study. Methods in Ecology and Evolution 10: 332–344.

Wickham H. 2014. Tidy data. Journal of Statistical Software 59: 1–23.

Wilson G. 2016. Software carpentry: Lessons learned. F1000Research 3: 62. https://doi.org/10.12688/f1000research.3-62.v2.

Wilson G, et al. 2014. Best practices for scientific computing. PLOS Biology 12: e1001745. https://doi.org/10.1371/journal.pbio.1001745.

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. 2017. Good enough practices in scientific computing. PLOS Computational Biology 13: e1005510. https://doi.org/10.1371/journal.pcbi.1005510.

Yenni GM, Christensen EM, Bledsoe EK, Supp SR, Diaz RM, White EP, Morgan Ernest SK. 2019. Developing a modern data workflow for regularly updated data. PLOS Biology 17: e3000125. https://doi.org/10.1371/journal.pbio.3000125.

Zuur AF, Ieno EN, Elphick CS. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1: 3–14.

Sarah E. McCord (sarah.mccord@usda.gov), Nicholas P. Webb, Justin W. Van Zee, Erica M. Christensen, Ericha M. Courtright, Amalia Slaughter, and Nelson G. Stauffer are affiliated with US Department of Agriculture ARS Jornada Experimental Range, in Las Cruces, New Mexico, in the United States. Sarah E. McCord and Craig Tweedie are affiliated with the University of Texas—El Paso, in El Paso, Texas, in the United States. Nicholas P. Webb, Erica M. Christensen, and Connie Maxwell are affiliated with New Mexico State University, in Las Cruces, New Mexico, in the United States. Sarah H. Burnett is affiliated with the Bureau of Land Management, National Operations Center, in Denver, Colorado, in the United States. Christine M. Laney and Claire Lunch are affiliated with the Battelle-National Ecological Observatory Network, in Boulder, Colorado, in the United States. Jason W. Karl is affiliated with the Department of Forest, Rangeland, and Fire Sciences, at the University of Idaho, in Moscow, Idaho, in the United States.