The importance of open science for biological assessment

Marcus W. Beck ([marcusb@sccwrp.org](mailto:marcusb@sccwrp.org)), Casey O’Hara ([ohara@nceas.ucsb.edu](mailto:ohara@nceas.ucsb.edu)), Julia Stewart Lowndes ([lowndes@nceas.ucsb.edu](mailto:lowndes@nceas.ucsb.edu)), Raphael D. Mazor ([raphaelm@sccwrp.org](mailto:raphaelm@sccwrp.org)), Susanna T. Theroux ([susannat@sccwrp.org](mailto:susannat@sccwrp.org)), David J. Gillett ([davidg@sccwrp.org](mailto:davidg@sccwrp.org)), Belize Lane ([belize.lane@usu.edu](mailto:belize.lane@usu.edu)), Greg Gearheart ([greg.gearheart@waterboards.ca.gov](mailto:greg.gearheart@waterboards.ca.gov))

Version Date: Sun Aug 4 13:01:10 2019 -0700

# Abstract

Open science principles that seek to democratize science can effectively bridge the gap between researchers and environmental managers. However, widespread adoption has yet to gain traction for the development and application of bioassessment methods. At the core of this philosophy is the concept that research should be reproducible and transparent, in addition to having long-term value through effective data preservation and sharing. In this paper, we review core open science concepts that have recently been adopted in the ecological sciences and emphasize how adoption can benefit the field of bioassessment for both prescriptive condition assessments and proactive applications that inform environmental management. An example from the state of California demonstrates effective adoption of open science principles through data stewardship, reproducible research, and engagement of stakeholders with multimedia applications. We also discuss technical, sociocultural, and institutional challenges for adopting open science, including practical approaches for overcoming these hurdles in bioassessment applications.

# Introduction

Bioassessment is an essential element of aquatic monitoring programs and helps to establish a foundation of decisions for managing the ecological integrity of environmental resources. Legal mandates to assess biological condition have set a precedent for developing bioassessment methods in the United States (Clean Water Act, CWA), Canada (Canada Waters Act), and Europe (Water Framework Directive). Decades of research to meet these mandates have supported the development of methods for multiple assemblages with regional applications in streams, rivers, lakes, and marine environments (Karr et al. [1986](#ref-Karr86), Kerans and Karr [1994](#ref-Kerans94), Fore and Grafe [2002](#ref-Fore02), Beck and Hatch [2009](#ref-Beck09), Borja et al. [2009](#ref-Borja09)). This body of applied tools represents substantial progress in understanding how biological organisms can be used as accurate and interpretable sentinels of environmental condition. Monitoring programs in the United States and internationally have collected millions of records of biological data spanning decades and hundreds of assessment methods have been developed from these data. The explicit link to environmental management distinguishes bioassessment from basic ecological research. Although bioassessment can and has been used to inform basic research, the intended use of these tools is to inform the protection and restoration of ecological integrity. As such, environmental managers require additional tools that transform bioassessment results into actionable information.

Bioassessment products will have limited use if they do not meet the needs of management and regulatory communities (Bain et al. [2000](#ref-Bain00), Stein et al. [2009](#ref-Stein09), Kuehne et al. [2017](#ref-Kuehne17)). In the United States, the CWA gives power to states, tribes, and territories for method development, which in turn requires federal approval to be implemented in a regulatory framework (e.g., Total Maximum Daily Load reporting, stormwater permitting). Currently, there is an imbalance between the methods developed for bioassessment and those that are approved for regulatory use. A recent review of assessment methods for ecological integrity in the United States showed that few were explicitly connected to freshwater policy (Kuehne et al. [2017](#ref-Kuehne17)). Biological indices are typically used to develop post-hoc diagnoses to trigger remediation or restoration actions, or to serve as early warning indicators of environmental change (Niemi and McDonald [2004](#ref-Niemi04)). Bioassessment tools for guiding planning activities, such as identifying conservation priorities (Linke et al. [2011](#ref-Linke11), Howard et al. [2018](#ref-Howard18)), could extend the application of bioassessment products beyond regulatory contexts. To these ends, more open sharing and discoverability (i.e., improved database connectivity and ease of access) of bioassessment data, methods, and tools is essential to minimize duplication of effort and to foster collaboration towards shared goals (Nielsen [2011](#ref-Nielsen11), Hampton et al. [2015](#ref-Hampton15)).

An effective bioassessment product must jointly address the technical challenges of developing a sensitive index and the implementation challenges of using an index to address the needs of environmental managers and regulators (Jackson and Davis [1994](#ref-Jackson94), Dale and Beyeler [2001](#ref-Dale01)). For decades, research in the bioassessment community has focused on addressing the former. Substantial technical advances have been made in predicting biological responses to environmental change, how these responses can be distinguished from natural environmental variation, and determining the impacts of these changes. Standardized protocols for sampling and producing assessment tools have also been developed (McDonald et al. [2004](#ref-McDonald04), Stoddard et al. [2008](#ref-Stoddard08)). A more urgent need is to address implementation challenges associated with how developed bioassessment products can be integrated into management or regulatory frameworks. Characterizing how an index could be used in practice to inform decisions and prioritize management actions is often opaque relative to why an index may have been originally developed. Hundreds of assessment methods have been developed for specific regional applications (Birk et al. [2012](#ref-Birk12)) and concerns about redundancy, comparability, duplicated effort, and lack of coordinated monitoring have recently been discussed within the research community (Cao and Hawkins [2011](#ref-Cao11), Poikane et al. [2014](#ref-Poikane14), Kelly et al. [2016](#ref-Kelly16), Nichols et al. [2016](#ref-Nichols16)). Moreover, existing indices may not be easily replicated beyond initial research applications (Hering et al. [2010](#ref-Hering10), Nichols et al. [2016](#ref-Nichols16)) or may be incorrectly applied based on differences between goals for developing an index and the needs of management programs (Dale and Beyeler [2001](#ref-Dale01), Stein et al. [2009](#ref-Stein09)). The abundance of available methods can be a point of frustration for managers given a lack of guidance for choosing an appropriate method among alternatives, particularly as to how a method may relate to specific management, monitoring, or policy objectives (Dale and Beyeler [2001](#ref-Dale01), Stein et al. [2009](#ref-Stein09)).

To address these challenges, a new mode of operation is needed where method development is open and transparent, developed products are discoverable and reproducible, and most importantly, implementation in the management community is intuitive and purposeful. Open science principles that democratize all aspects of the scientific method can meet these needs, yet bioassessment research and its application to better serve the environment has not yet fully embraced these principles. Others have advocated more broadly for inclusion of open science principles in the ecological sciences (Hampton et al. [2015](#ref-Hampton15), [2016](#ref-Hampton16), Lowndes et al. [2017](#ref-Lowndes17)) and a growing wave of momentum has influenced how scientists conceptualize research in other disciplines (e.g., archaeology, Marwick et al. [2016](#ref-Marwick16), behavioral ecology, Ihle et al. [2017](#ref-Ihle17), hydrology, Slater et al. [2019](#ref-Slater19), vegetation sciences, Collins [2016](#ref-Collins16)). Adopting an open science paradigm in bioassessment is particularly relevant compared to other fields given the explicit need to develop tools that are accessible to the management community. Legal and ethical precedents in bioassessment may also necessitate open data sharing given that environmental monitoring programs are often publicly funded.

This review demonstrates tools and approaches for open science, which will empower the research and management community to embrace a new mode of thinking for bioassessment applications. These approaches are expected to benefit the bioassessment research community by augmenting existing workflows for developing assessment tools and improve the ability of these methods to address environmental issues by bridging the gap between the scientific, management, and regulatory communities. As such, this paper describes open science tools that can be adopted in research practices with an emphasis on communication, but we also stress the value that these tools have for the broader community of environmental managers. An overview of the general principles of open science is provided, followed by a discussion of specific benefits and how these principles can be applied to bioassessment. We use examples from the state of California to demonstrate how existing tools can be tailored to address legislative mandates for free and open sharing of data, especially by directly engaging stakeholders that require practical approaches for using bioassessment tools in planning activities. We conclude with a discussion of technical, sociocultural, and institutional hurdles that have, thus far, prevented widespread adoption of open science and provide recommendations for the bioassessment community to address these challenges.

# Principles of open science

Conventional modes of creating scientific products and more contemporary approaches that align with open science principles share the same goals. Both are motivated by guiding principles of the scientific method that make the process of discovery transparent and repeatable. Where the conventional and open science approaches diverge is the extent to which technological advances are leveraged as instrumental tools that facilitate the entire research process. Distinction between the two approaches can be conceptualized as the “research paper as the only and final product” for the conventional approach, whereas the open science approach is inherently linked to advances in communication and analysis that have been facilitated by the Internet and computer sciences (Table 1). As a result, the open science approach can enhance all aspects of the scientific process from initial conception of a research idea to the delivery and longevity of a research product (Figure 1). The process is iterative where products are improved by the individual and/or others, facilitated by open science tools that enhance access and reproducibility of data.

The paradigm of the research paper as a final scientific product can inhibit the uptake of methods and findings by environmental managers. The research paper is conventionally viewed as a communication tool for scientists to report and share results among peers. Researchers access periodicals to stay informed of scientific advances and use the information to replicate methods for follow-up analysis. Although the primary literature continues to provide these fundamental services, this workflow is problematic when scientific products are needed to serve interests outside of the research community. For example, the paper as an endpoint for environmental managers fails to deliver tools that are easily accessible from the practitioners perspective, both in application and interpretation. A research paper is less likely to effect environmental change because it does not provide a mechanism to transfer actionable information to those that require scientific guidance for decision-making, such as sharing analysis code or results that describe output from assessment tools. Numerous studies have documented implementation failures as a result of siloing among research communities where the flow of information does not extend beyond institutional walls (e.g., Mitchell [2005](#ref-Mitchell05), Liu et al. [2008](#ref-Liu08)). Information loss over time is another well-known flaw associated with the paradigm of research paper as final product (Michener et al. [1997](#ref-Michener97)).

## Open data as a component of the open science process

Open data is a fundamental component of the broader open science process in Figure 1. Under this mode of thinking, the research team becomes stewards of its data. The responsibility for data stewardship also extends beyond the research arena to public institutions that maintain large datasets or synthesize smaller datasets from multiple sources. Stewardship allows the data to be treated as a living product with a traceable and replicable provenance (i.e., origin), rather than proprietary and serving only the internal needs of an immediate research goal. Metadata that describe the structure and history of a dataset are essential tools that ensure the data have an identity. Metadata also encourage adoption of core data structures that allow integration across different sources, which is critical for collaboration across institutional boundaries (Horsburgh et al. [2016](#ref-Horsburgh16), Hsu et al. [2017](#ref-Hsu17)). Other open science practices, such as integration of data with dynamic reporting tools or submitting data to a federated repository (i.e., a decentralized database system for coordination and sharing), can be applied to any and all of these data with the end goal of facilitating communication for researchers and those for which the research was developed (Bond-Lamberty et al. [2016](#ref-BondLamberty16)). These tools also benefit the individual researcher by providing information and version tracking to recreate a past working environment. In all cases, the data are openly accessible and documented for reproducibility and discovery using technologies that facilitate communication and sharing.

Open data can benefit research by contributing to an increase in novel products created through collaboration. Collaborative publications have increased in the environmental sciences as research teams leverage open data to create synthesis products that allow novel insights from comparisons across multiple datasets. Quantitative meta-analyses and systematic reviews are increasingly used to extract information from the primary literature (Lortie [2014](#ref-Lortie14)). In addition, open data products can increase efficiency of the individual researcher and a collective research team by encouraging collaborators to adopt an open science workflow. Many tools developed within the software and computer science community to facilitate open process and the creation of open data are now easily accessible to environmental scientists. Version control software (e.g, Git, GitHub), open source programming languages (e.g, R, Python), and integrated development environments (IDEs, e.g., RStudio, Spyder) can all be leveraged by applied ecologists to dynamically create and share open data products that build institutional memory. These tools promote deliberate and shared workflows among researchers that can lead to better science in less time (Lowndes et al. [2017](#ref-Lowndes17)) and have proven useful in recent applications in the hydrologic sciences (Idaszak et al. [2017](#ref-Idaszak17), Slater et al. [2019](#ref-Slater19)).

Open access to data can also benefit management and regulatory communities. Although many monitoring datasets are publicly available, the data quality, level of documentation, and ease of use varies widely. This can create research and implementation challenges as datasets may be difficult to locate, data contexts may be misunderstood, and data from different sources require synthesis. Many open science tools can improve the accessibility of data from monitoring programs by establishing workflows for data discovery and synthesis, often through the adoption of a common metadata structure (e.g., Ecological Metadata Language Standard, or [EML](https://knb.ecoinformatics.org/external//emlparser/docs/index.html)) and integration of data within federated data networks (e.g., [DataONE](https://www.dataone.org/), [iRODS](https://irods.org/)). Open data maintained by management or regulatory communities benefits the research community, which in turn benefits the data maintainers that require scientific products to inform decisions. Open data can also improve public trust in scientific findings by exposing the underlying information used to develop a research product (Grand et al. [2012](#ref-Grand12)). Similar concepts are used in “blockchain” technologies that allow public financial transactions in an open, distributed format, as for trading in cryptocurrencies (Pilkington [2016](#ref-Pilkington16)). Increased trust could facilitate eventual adoption of proposed rules or regulations that are based on research products created from open data.

# Applying open science principles to bioassessment

Here we provide a detailed description of open science processes that the bioassessment community could leverage to adopt a philosophy of creating reproducible, transparent, and discoverable research products for environmental managers. To emphasize the value of open science for specific steps of the scientific process, we first describe a “conventional” linear process (Figure 2a), which is then contrasted with a process that adopts open science tools to facilitate feedback between the management community and stakeholders (Figure 2b). The critical differences are the iterative flow of ideas and products between the management community and stakeholders, the research team developing the bioassessment product, and the broader community that provides data and guidance documents in the primary and secondary literature. In both cases, the technical and implementation phases of a bioassessment product are acknowledged as distinct steps of analysis that describe the entire process from idea conception to adoption in management or regulatory applications. We use this paradigm to jointly demonstrate how open science tools can be applied beyond the research phase, but also reinforce the concept that a bioassessment product is only as relevant as its management application (e.g., a bioassessment product will not have applied value if its final home is the primary literature).

Efforts to formally recognize and integrate open science in bioassessment are needed now more than ever. The transition of bioassessment from taxonomic-based indices to molecular approaches presents novel challenges that will only increase in severity as researchers continue to refine methods for molecular applications (Baird and Hajibabaei [2012](#ref-Baird12)). Although molecular-based indices share similar assessment objectives as conventional indices, the data requirements and taxonomic resolution are substantially more complex. Bioassessment research teams developing molecular methods are and will continue to be inundated with data from high-throughput DNA sequencers. Systematic approaches to document, catalog, and share this information will be required to advance and standardize the science. Molecular approaches are also dependent on existing reference libraries for matching DNA samples for taxonomic identification. The integrity of reference libraries depends greatly on the quality of metadata and documentation for contributed samples. Open science principles should be leveraged in this emerging arena to ensure that new bioassessment methods continue to have relevance for determining the condition of aquatic resources.

The below examples require understanding the distinction between the general open science process in Figure 1, open data as an individual component of the open science process, and the technology-based tools that can be used to achieve these ends. We emphasize that both the tools and open data are critical components that facilitate the broader process to achieve the principles outlined in Table 1. However, open science tools can also be used by individual researchers in an entirely closed workflow with no collaboration or discoverability by others. Similarly, open data can be created through an entirely closed process even though it may appear as an open science product. “Openness” of process, tools, and data exists on a continuum, and incremental improvements can transform an individual’s and research group’s practice over time. We encourage awareness that an open process adopts the open science tools that are appropriate for a research question, the creation of open data can be a fundamental component of the process, and acceptance by the research team and collaborators of the concepts described in Table 1 is critical to achieving openness.

Concepts of openness in Table 1 also have analogs in the conventional paradigm that are worth noting to better distinguish between open and conventional workflows. First, an open process, as noted above, encapsulates concepts of transparency, discoverability, and availability, whereas a closed process is marked by isolation from any external feedback between researchers and outside institutions. Closed processes are also opaque because they are concealed from outsiders and cannot be easily replicated. A common example is manual spreadsheet manipulation where the steps are neither explicitly nor implicitly documented (as through scripting or pseudocode), such that others and the individual researcher may be unable to replicate the workflow in the future. Further, analysis tools that are created with open source platforms may also be opaque if there is no documentation or description of the underlying analyses to illuminate the “black box” of the internal code. Finally, open processes can make data available both from an accessibility and “readability” perspective, as for both human- and machine-readable formats. Inaccessible data by contrast have none of these properties and may be proprietary or unreadable for analysis (e.g., an image of a data table in a PDF).

## Conventional bioassessment

First we consider a hypothetical process by which many bioassessment products are created to highlight challenges to openness with conventional approaches. A typical workflow for developing a bioassessment product is not entirely dissimilar from a conventional scientific process. The primary investigator begins by identifying the research goal, developing methods to achieve that goal, and then identifying the data needs and analyses based on the existing science. What distinguishes bioassessment from traditional scientific studies is that research goals are often defined externally through legislative requirements. This means that the required datasets may be decentralized across institutions or data collection may be contracted to a third party. This separation between consumers and creators of bioassessment products can be a contributing factor for adopting closed workflows given communication barriers between institutions. For these reasons, the conventional process is often linear and opaque by lack of communication and transparency between parties (Figure 2a).

A conventional process for a bioassessment study can also lead to inaccessible data with inadequate documentation. Many bioassessment products begin by synthesizing data from existing monitoring programs, although the original goals for a monitoring program may not align with the needs of a bioassessment product. The data may also have limitations that affect performance or even prevent their use entirely for assessment. For example, a fisheries monitoring program may only sample select locations and may have inadequate coverage of aquatic biodiversity. In these scenarios, supplementary data must be gathered from other sources. The end product is a synthesized dataset that may be appropriate to meet the requirements of a bioassessment product, but its origin and workflow to achieve the necessary format may be unknown except to the individual that compiled the data. This creates significant challenges for future products or policies that depend on these products and may even erode institutional trust if the provenance of the dataset is unknown.

Insufficient tools to facilitate use of bioassessment products by managers and policy-makers, both in calculation and interpretation, represents a critical limitation of conventional workflows. The worst case scenario is a spreadsheet-oriented approach to converting raw data into a synthesized index. Not only is this a major challenge for the researcher that develops the index, but its methods and order of operations are opaque, preventing others from applying the tool to reproduce the results and apply it in novel situations. There are many reasons for these challenges (e.g., lack of documentation, unknown data requirements, no interpretation guidance), but the fundamental issue is inaccessibility of the product. A management or regulatory agency will have no ability to interpret these data and the final product will not pass independent review if the results cannot be reproduced. Recent workflows have been proposed as “cradle to dashboard” to remediate these issues, where code development, data, and final products are intimately linked from conception to completion (e.g., for environmental flows, [https://eflows.ucdavis.edu/](https://eflows.ucdavis.edu/hydrology)). Similar approaches could be useful in bioassessment to remediate data accessibility and reproducibility issues.

## Using open science tools to enhance bioassessment

The above example is an extreme hypothetical description where bioassessment products fail to effect any positive change in environmental management due to a complete deficiency in openness, resulting in failed implementation. Many existing bioassessment methods can be described more positively, yet the research community could benefit from adopting a more open approach to creating and delivering assessment tools. This approach is particularly relevant from the perspective of implementation and science translation, as all modes of communication between research, managers, and regulators could be enhanced with open science. The following section describes key components of the open science process that can facilitate the development and implementation of bioassessment products. We focus specifically on applications for data provenance, method development, and method delivery.

The overall process is shown in Figure 2b as an expansion of general concepts in Figure 1, with a specific science application phase for implementation. This iterative flow of information is facilitated by 1) openly sharing planning documents, 2) using established metadata standards to document synthesized data products, 3) hosting data products on open repositories, 4) creating reproducible summary documents that integrate the data and research products, and 5) incorporating the developed product into interactive applications that deliver the results to the managers and stakeholders. The technical phase of defining research goals, collecting and synthesizing data, and developing the bioassessment product are primary tasks of the research team. However, the process is distinguished by the flow of information to and from the research phase that can benefit the specific project and the science of bioassessment as a whole.

### Developing bioassessment goals

In an open science process, the goals identified by the research team for developing a bioassessment product should occur through direct, two-way interaction with the management or regulatory institution that requires the product. This two-way exchange of information can be accomplished through direct communication and sharing of planning documents to ensure all decisions are transparent, i.e., open planning. In person meetings are ideal, but planning documents are dynamic and will require remote sharing and revision as ideas progress. Online tools such as [Google documents](https://www.google.com/docs/about/), [Slack](https://slack.com/) discussion channels, and open lab notebooks can be instrumental for collaboration. More informal approaches, such as blogging and sharing ideas on social media, can expose new concepts to the broader community for guidance (Woelfle et al. [2011](#ref-Woelfle11), Darling et al. [2013](#ref-Darling13)). Overall, the research team should use these tools to identify stakeholder needs while also considering the balance between the research goals and limitations of the data to meet these goals. This will ensure that the needs of the management and stakeholder communities will be consistent with the services provided by the research product.

### Curating bioassessment data

After the goals are established, the research team identifies requirements and sources of data that need to be synthesized to meet the research needs. Under a closed scenario, data flows one way from the source to the researcher and is used only as a means to create the final research product. In the open scenario, the data itself is a product to achieve the research goals and also becomes available to the research and management community as a fully documented source of information that has value beyond the specific project. The openness of the synthesized data product is one of the primary means of facilitating the implementation phase of a bioassessment product. The synthesized data product can be used by the research team to create interactive applications for stakeholders to share and explore the data and is also fully integrated into summary reports using software for generating dynamic documents (e.g, using knitr, Xie [2015](#ref-Xie15), RMarkdown, Allaire et al. [2018](#ref-Allaire18), Jupyter notebooks, Kluyver et al. [2016](#ref-Kluyver16)). The data product also becomes available on an open data repository that is discoverable by other researchers and can contribute to alternative scientific advances beyond the immediate goals (e.g., Hydroshare for the hydrologic sciences, Idaszak et al. ([2017](#ref-Idaszak17))).

Open bioassessment data created by a research team using an open science process clearly describes the history and purpose of a dataset to easily allow additional research and delivery of products to managers. This is particularly relevant given that a synthesized dataset can originate from multiple sources and the combined output may differ substantially from the raw data. Documenting provenance of a data product allows a research team and others to understand the purpose of a combined dataset, how it was created, and how it differs from the original data. For example, a multimetric index may require taxonomic data collected at multiple sites by different institutions, whereas the output data may include summary scores, individual metrics, and any additional supporting information to assess the quality of the output. A data product can easily be documented using a standardized metadata language (e.g., EML) which describes the who, what, and why to ensure the data have an identity. Adoption of a metadata standard also ensures that a machine-readable file is produced to allow integration into a data repository. This will allow a synthesized data product to be discoverable beyond the specific research application and will provide metadata to help others understand the context of the data. Finally, a dataset can be assigned a unique digital object identifier (DOI, e.g., through [Zenodo](https://zenodo.org/)) that provides a permanent address and is also citable to allow researchers to track usage of a bioassessment data product.

### Using R for bioassessment application

The application of bioassessment products to the management and regulatory community is critical to ensure that research is applied by decision-makers for its intended purpose. Application should be fundamentally linked to open science analysis and development tools that can be used to deliver the products using a reproducible and accessible platform. In particular, the R statistical programming language (RDCT (R Development Core Team) [2018](#ref-RDCT18)) is one of the most commonly used analysis platforms in the environmental sciences for creating specific analysis workflows (Lai et al. [2019](#ref-Lai19), Slater et al. [2019](#ref-Slater19)), yet the availability of existing packages and the ability to create new packages is a strength of R that is under-utilized by the bioassessment community. An assessment index packaged in R can automate the tedious process of converting raw taxonomic data to bioassessment information. A well-designed R package is also modular, meaning it includes all necessary analyses, data, and documentation to allow use by others. This modularity is critical for reaching management and stakeholders by providing a tool that redirects the focus on the output and not the technical details that are less important for decision-making.

Many existing R packages have value for the bioassessment community (Table 2). For managing the day to day tasks of working with multiple datasets, the tidyverse suite of packages provides the necessary tools to import, wrangle, explore, and plot almost any data type (Wickham [2017](#ref-Wickham17b)). These packages are developed around the concept of “tidy” data that provide a common and natural framework for working with data (Wickham [2014](#ref-Wickham14c)). These principles have importance for bioassessment where the synthesized datasets used to create a product should be logically structured to facilitate use by others. Although the time and effort required to produce a tidy dataset can seem excessive relative to native formats (e.g., raw sample data), downstream analyses will be greatly facilitated because a common format can be anticipated rather than having to develop specific workflows to accommodate raw data. The tidyverse also includes the powerful ggplot2 package that is based on a syntactical grammar of graphics for plotting (Wilkinson [2005](#ref-Wilkinson05), Wickham [2009](#ref-Wickham09)). This package provides a set of independent plotting geometries and aesthetics that can be built piecewise and is a departure from other graphics packages that represent a collection of special cases that limit the freedom of the analyst. The structure of ggplot2 makes data visualization an integral part of the analysis workflow, rather than only at the end of an analysis where final graphics are produced. In bioassessment, ggplot2 can be used both in an exploratory role during the development phase and also to create publication quality graphics. More importantly, this package provides the building blocks to create effective data visualizations that convey important components of a bioassessment product to managers and stakeholders.

Bioassessment data are inherently spatial and recent package development has greatly improved the ability to analyze and map geospatial data in R. These tools can readily communicate the spatial context of bioassessment products to managers and stakeholders by mapping index or condition scores to stream flow networks, watersheds, and ecoregions, both for high-level planning and site-specific evaluations. The raster package can used to read/write, manipulate, analyze, and model grid-based spatial data (Hijmans [2019](#ref-Hijmans19)), which are often common supporting layers for bioassessment (e.g., elevation or climate data). For vector data (i.e., points, lines, and polygons), the sf package (“simple features”, Pebesma ([2018](#ref-Pebesma18))) was first released in 2016 and has quickly become the most highly-used approach for working with spatial information in R. The sf package uses principles of data storage that parallel those from the tidyverse by representing spatial objects in a tidy and tabular format. This facilitates analysis by presenting complex spatial structures in a readable format that can be integrated in workflows with existing packages, including other mapping packages (e.g., leaflet, Cheng et al. [2018](#ref-Cheng18), or mapview, Appelhans et al. [2018](#ref-Appelhans18)). This allows the research team to use a workflow that is focused in a single environment, rather than using separate software for statistical and geospatial analysis. This facilitates participation if a research team has a shared analysis platform, while also minimizing loss of information or introducing analysis errors between environments.

Several existing R packages can be used to develop statistical models of bioassessment data that are a necessary component of many analyses. Random forest models have been used to develop predictive bioassessment indices that compare observed taxa to modelled expectations (i.e., O/E indices). The randomForest package (Liaw and Wiener [2002](#ref-Liaw02)) uses an ensemble learning approach that is robust to complex, non-linear relationships and interactions between variables. These models are particularly useful with large, regional datasets that describe natural and anthropogenic gradients in condition (Laan and Hawkins [2014](#ref-VanderLaan14), Mazor et al. [2016](#ref-Mazor16)). Many other modelling packages are available in R that can support index development, such as exploratory analyses to evaluate biological response or identifying significant associations of organisms with stressor gradients. The nlme package can be used to create non-linear mixed effect models that are more flexible than standard regression approaches (Pinheiro et al. [2018](#ref-Pinheiro18)). The nlme package can develop models for nested sampling designs, such as repeat visits to sample sites or otherwise confounding variables that contribute information but are not unique observations (Mazor et al. [2014](#ref-Mazor14)). The mgcv package provides similar functionality as nlme, but uses an additive modelling approach where individual effects can be evaluated as the sum of smoothed terms (Wood [2017](#ref-Wood17)). The mgcv package is often applied to model biological response to stressor gradients (Yuan [2004](#ref-Yuan04), Taylor et al. [2014](#ref-Taylor14))

Other R packages have been developed specifically for bioassessment. For example, the TITAN2 package can be used to develop quantitative evidence of taxon-specific changes in abundance and occurrence across environmental gradients (Baker et al. [2015](#ref-Baker15)). Results from this package can support exploratory analysis for developing bioassessment products, such as identifying indicator species or establishing numeric criteria (Taylor et al. [2018](#ref-Taylor18)). The results can be also be used post hoc to evaluate potential response of a biological index with changing environmental conditions, such as proposed management actions for rehabilitation (King et al. [2011](#ref-King11)). Alternatively, the indicspecies package provides similar functionality but is based only on species occurrence or abundance matrices across sites (De Caceres and Legendre [2009](#ref-DeCaceres09)). This package can be used to identify species that occur at particular sites if continuous environmental data are unavailable, such as those that are representative of reference conditions (Bried et al. [2014](#ref-Bried14)). Finally, the vegan package has been a staple among community ecologists for multivariate analyses in R, such as clustering and ordination (Oksanen et al. [2018](#ref-Oksanen18)). This package has value for bioassessment as an exploratory tool with a variety of applications, such as analyses that identify comparable assessment sites that minimize natural variation (Beck et al. [2013](#ref-Beck13)).

Although the R network includes over 10000 user contributed packages, the most under-utilized aspect of this software for bioassessment is the creation of new packages to support applied research. Community practices have allowed R to reach new audiences where new packages build on the work of others and are transportable between users and operating systems, rather than all researchers reinventing the wheel through duplicated effort. Formalized communities, such as [rOpenSci](https://ropensci.org/), encourage standardization and review of contributed packages within the ecological sciences to make scientific data retrieval reproducible. Several tools have also been developed and published in the last five years that greatly simplify the process of creating new packages in R (Wickham [2015](#ref-Wickham15), Wickham et al. [2018](#ref-Wickham18)). The increasing popularity in the environmental sciences of online repositories for hosting code and software, such as GitHub, also provides a venue for sharing and directly installing these packages.

The advantages of creating and sharing R packages that are specific to bioassessment applications are important for several reasons. First, a well-designed R package provides a compartmentalized set of instructions developed during the technical research phase that can be executed by anyone with access to the software. This allows the developer to include important technical elements required for the execution of a bioassessment product within the package, while allowing the end user to focus on how the output can be used to inform decision-making. R packages also require explicit documentation of the functions and data requirements. As such, package users will not only have access to underlying code but also understand the why and what for different package functions. Detailed vignettes can be included in plain language to describe how to use the package.

Peer-review mechanisms are available for user-contributed packages, such as through rOpenSci or through package-oriented journals (e.g., the R Journal, The Journal of Statistical Software). These review mechanisms are meant to ensure that the underlying code and algorithms are satisfactory for the stated needs that are addressed by the package. However, a majority of user-contributed packages (e.g., on GitHub) are not formally reviewed and users are encouraged to always assess the validity of a package before application. Cursory checks can be made by verifying author credentials or conferring among colleagues as to which packages are commonly used for particular analyses. However, software review is the gold standard and package developers are increasingly encouraged to allow external review of underlying code and theory prior to submitting a package to a public repository (Borregaard and Hart [2016](#ref-Borregaard16)).

Finally, R can be used to create interactive applications that deliver bioassessment products to stakeholders and managers in entirely novel contexts. In particular, the shiny package was first released for R in 2012 and provides programming tools built around concepts of reactivity, where data inputs and outputs can be modified in real time (Chang et al. [2018](#ref-Chang18)). A shiny application is an interactive user interface that is developed with R code, but is a standalone product that can be used without any programming experience. These applications are deployed online and can extend the reach of bioassessment tools to those that require the information for decision-making but otherwise do not have the time or resources to learn R. Applications built in shiny can also be easily linked to other R packages. For example, a shiny website could be created to allow users to upload raw data and estimate and report bioassessment scores using an R package developed externally. This can extend the accessibility of a bioassessment product while maintaining the technical integrity of the original tool. Moreover, shiny applications are completely customizable and can be tailored by the developer to the specific needs of any user. This distinction separates shiny from other web-based analysis platforms.

# Open science in practice

Although bioassessment methods have been sufficiently developed in California, there are no narrative or numeric criteria in place to support designated aquatic life uses in wadeable streams, nor are bioassessment data actively used to support conservation or watershed management. Indices using benthic macroinvertebrates and algae have been developed that provide consistent indications of biological condition across the diverse geography and climates in the state (Fetscher et al. [2013](#ref-Fetscher13), Mazor et al. [2016](#ref-Mazor16), Ode et al. [2016](#ref-Ode16)). A physical habitat index has also been developed that provides complementary information supporting bioassessment data (Rehn et al. [2018](#ref-Rehn18)). Combined, these indices represent significant achievements in overcoming technical challenges for developing accurate and interpretable bioassessment methods. However, these products are not used at a statewide scale to inform decisions and past efforts for stream management have only used a fraction of available methods. A synthesis of condition assessments from multiple methods is needed to effectively implement bioassessment methods in California and data must be presented in a context that is relevant to the needs of decision makers.

Recent regulatory initiatives in California (USA) have established a foundation for openness that could greatly improve the application of bioassessment methods to support decision-making. In particular, these initiatives have set a precedent for openly sharing data collected with public funds. The Open and Transparent Water Data Act passed by the state legislature in 2016 requires water quality institutions to “create, operate, and maintain a statewide integrated water data platform that, among other things, would integrate existing water and ecological data information from multiple databases and provide data on completed water transfers and exchanges” ([AB 1755, Dodd, 2015-2016](https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201520160AB1755)). This legislation also calls for state agencies to “develop protocols for data sharing, documentation, quality control, public access, and promotion of open-source platforms and decision support tools related to water data”. These aspirations were further supported by a [resolution](https://www.waterboards.ca.gov/press_room/press_releases/2018/pr_water_data_071018.pdf) on July 10, 2018 that formally committed the State Water Resources Control Board to “provide broader access to data used to make local, regional, and statewide water management and regulatory decisions in California”. These recent initiatives in California have similarly been observed at the national level. For example, the [Data Coalition](https://www.datacoalition.org/) is an advocacy group that operates on behalf of the private and public sector for the publication of government data in a standardized and open format. The [Internet of Water](https://internetofwater.org/) also operates at the national-level by focusing on strengthening connections between data producers and users through centralized data hubs and data standards.

The recent policies and initiatives are a recognition of the disconnect between the state of the science and the level of integration with policy and decision-making. Although these initiatives represent substantial progress, the implementation of existing bioassessment methods to support policy or planning remains uncertain. For example, thresholds for bioassessment indices that are protective of ecosystem integrity are often required for application in a regulatory context. Thresholds can be identified based on variability of reference conditions, yet the justification for use in policy requires a link between the threshold and ecosystem function. Moreover, reference conditions may be an unreasonable management target in developed landscapes where biological communities could be constrained by stressors originating at the watershed scale (Paul et al. [2008](#ref-Paul08), Kenney et al. [2012](#ref-Kenney12), Shoredits and Clayton [2013](#ref-Shoredits13)). These constraints represent significant challenges in setting statewide thresholds for bioassessment indices where goals may be easily met in some areas of the state (e.g., open areas) but not others (e.g., urban areas).

A fundamental concern of regulated parties with the potential adoption of statewide numeric criteria is the likelihood of achieving a one-size-fits-all requirement across the variety of landscapes and channel types in the state (Lane et al. [2018](#ref-Lane18)). Although assessment tools have been developed to account for regional variation in stream communities, there is a disconnect between what biological condition is a reasonable management goal and what might be required under a statewide criteria. For example, a threshold that represents the tenth percentile of bioassessment scores at all statewide reference sites would be relatively easy to achieve in the Sierra Nevada region that is largely undeveloped. Conversely, a similar threshold would be very difficult to achieve in Southern California where heavy urbanization has occurred in the last seventy years and much of the hydrologic network has been modified through channelization or concrete reinforcement. Managers require bioassessment tools that not only identify state of the resource, but what is a reasonable expectation for biological condition given the immutable changes on the landscape that are largely beyond management intervention. Should concrete channels be held to the same standard as streams in more natural settings, and if not, what bioassessment tools are available to support management goals in these different environments?

Open science methods have recently been used in California to address bioassessment implementation challenges in developed landscapes. The Stream Classification and Priority Explorer or SCAPE (Beck [2018a](#ref-Beck18c), Beck et al. [n.d.](#ref-Beckir)) was developed using an open science framework to help identify reasonable management goals for wadeable streams using existing bioassessment and watershed data. The SCAPE tool represents both a modelling approach to identify appropriate management goals and a set of open science products for science application to environmental managers. First, a statewide landscape model was developed that predicts a range of likely biological condition scores using the California Stream Condition Index (CSCI, Mazor et al. [2016](#ref-Mazor16)) from a suite of watershed metrics (e.g., land cover, road density). The model was developed specifically using predictors that are typically beyond the scope of management intervention to describe the range of scores that are likely to be observed under current land use conditions. From this information, a stream classification system was developed to identify sites where biological condition is “constrained” or “unconstrained” by landscape alteration. This classification system can be used to set management priorities based on the constraint class. For example, a monitoring site with an observed biological index score that is above a predicted range could be assigned a higher management priority relative to a site that is scoring within the range that is expected based on landscape development.

An example of how the SCAPE tool can link bioassessment information to managers is demonstrated in Figure 3. The model was developed using statewide data, whereas results were regionally applied to develop management priorities through close interaction with a local stakeholder group from the San Gabriel River Watershed (Los Angeles, California). The San Gabriel watershed represents an extreme landscape gradient with open areas in the upper watershed and heavy urbanization in the lower watershed where almost all channels are engineered for stormwater management or water conveyance. Streams in the lower watershed represent a management challenge where biological potential is limited by landscape development; reference conditions based on a statewide expectation may be difficult to achieve in these areas. The landscape model helps identify channels where biological constraints occur as a function of landscape to help prioritize among locations. All stream segments in the watershed were classified as constrained or unconstrained by landscape and bioassessment sites were then evaluated relative to the class and ranges of scores that were expected given the landscape. Management actions and priorities were then defined based on the stream class and if the observed score was as under-score, as expected, or over-scoring. For example, sites not scoring as expected were assigned high priority for additional investigation, unconstrained sites were identified as high priority for protection, and under-scoring unconstrained sites were assigned high priority for restoration.

Open science products that were created through the life cycle of the SCAPE tool were critically important for translating and delivering bioassessment data to decision-makers. Not only does the model address a critical management need (i.e., where are constrained streams and how can they be prioritized), but all aspects of model development, implementation, and translation followed an open science workflow (Figure 4). Local stakeholder engagement to identify research goals guided the technical development process of SCAPE (Figure 4, box 1). Continued involvement of this group throughout the project ensured the final product addressed all research goals and management needs. All analyses, including model development and validation, were conducted using the R statistical programming language (RDCT (R Development Core Team) [2018](#ref-RDCT18)). Additional R packages that greatly facilitated the process included the tidyverse suite of packages for data import, wrangling and plotting (Wickham [2017](#ref-Wickham17b)), the sf package for spatial analysis (Pebesma [2018](#ref-Pebesma18)), and randomForest for modelling (Liaw and Wiener [2002](#ref-Liaw02)) (Figure 4, box 2). A version control system (Git) and online hosting (GitHub) also allowed full transparency of decisions that were made to create the tool. A permanent DOI was assigned through Zenodo to track downloads and portability of source code (Figure 4, box 3).

A critical component of the SCAPE tool was the use of an online, interactive web page (<https://sccwrp.shinyapps.io/SCAPE>) that was developed entirely using open-source software (i.e., R, RDCT (R Development Core Team) [2018](#ref-RDCT18), shiny, Chang et al. [2018](#ref-Chang18), Figure 4, box 4)). This online application provided two fundamental services that greatly increased the impact and relevance of the bioassessment product for the management community. First, the application improved stakeholder understanding of the technical model components through direct interaction with key decision points that influenced model output. This improved stakeholder comprehension and trust for the intended application of the model by using a hands-on approach for interacting with the landscape model. Second, potential decision points that affected model output were not hard-coded into the application so that the effects of varying parameters (e.g., potential biointegrity thresholds) could be better understood relative to management implications. This second component is a fundamental benefit of open science applications in bioassessment because the tool was developed to allow exploratory analysis, both within the constraints of the model and the needs of the management community. The online tool did not limit managers to a single answer, rather it provided the tools to determine the most appropriate answer for different situations by giving managers the ability to directly interact with the model.

Conventional scientific products for the SCAPE tool were also created as deliverables using open science tools. These were created as products for both the local management community with direct interests in the landscape model and the broader scientific community concerned with bioassessment challenges in developed landscapes. A manuscript describing the model in detail was created using knitr and RMarkdown (Xie [2015](#ref-Xie15), Allaire et al. [2018](#ref-Allaire18)). This increased efficiency of the writing process and also minimized the potential of introducing errors into tables or figures by eliminating the need to copy results between different platforms. The manuscript was also under full version control through GitHub for streamlined collaboration and tracking of revisions (Figure 4, box 5). A geospatial data file from the model was also made public on a federated data repository, which included metadata and plain language documentation to track provenance of the original information (Beck [2018b](#ref-Beck18d)) (Figure 4, box 6). This ensured that the final data could be discovered by others through keyword searches and also provided a platform to track broader interest in the data products through download statistics.

The SCAPE tool is one example of how open science can be effectively integrated with bioassessment data. Most importantly, SCAPE is an interactive decision-support tool that allows stakeholders to more effectively use bioassessment data to prioritize management actions on the landscape. Open science aspects of SCAPE enhanced this functionality by providing a transparent and reproducible product that greatly expanded the utility of existing bioassessment products in California. The open science products described above could be leveraged in other contexts to increase the impact and relevance of bioassessment data for achieving management goals.

# Limitations and opportunities

Although the case for open science in bioassessment is appealing, the widespread adoption of these principles in practice is inhibited by inertia of existing practices, disciplinary culture, and institutional barriers. Conventional and closed workflows used by many scientists are adopted and entrenched because of ease of use, precedence, and familiarity, yet they can be inefficient, inflexible, and difficult to communicate or replicate. Open science tools can improve analysis, documentation, and implementation through greater flexibility, but they expose research teams to entirely new concepts and skillsets in which they may never have been trained (e.g., Idaszak et al. [2017](#ref-Idaszak17)). Many scientists feel they cannot prioritize learning new skills given existing demands on their time, particularly if the benefits of these approaches, such as the value for the research team of sharing their data, may not be apparent or immediate. Not only are the required skillsets demanding, but the open science toolbox continues to expand as new methods are developed and old methods become obsolete. This requires a research team to stay abreast of new technologies as they are developed and weigh the tradeoffs of adopting different workflows for different research tasks.

Advocates for open science are well aware of the technical challenges faced by individuals that have never been exposed to the core concepts. Most importantly, education and training (e.g., through [The Carpentries](https://carpentries.org), [DataCamp](https://www.datacamp.com/)) remain key components for developing skillsets among researchers where the focus is both on learning new skills for transferability and realizing their value for improving science as a whole (Hampton et al. [2017](#ref-Hampton17)). A goal of many training curricula is to instill confidence in new users by developing comfort with new workflows, such as replacing a point-and-click style of analysis with one focused on using a command line through a computer terminal. Other approaches to demonstrate the value of new techniques use a side by side approach of closed vs open workflows to show the increased efficiency and power of the latter. Adoption becomes much more reasonable once users realize the value of investing in learning a new skill.

Advocates of open science also recognize the limitations of teaching in that not all audiences can be reached and not all materials are retained or even used after training. A technique of letting the trainee become the trainer can be used whereby those that successfully learn new skills can teach others at their home institutions. Those that also adopt new workflows through training can also direct their research products to facilitate collaboration with non-adopters rather than the latter synthesizing and analyzing their data in potentially suboptimal ways (Touchon and McCoy [2016](#ref-Touchon16)). These “champions” can be a voice of encouragement for others by demonstrating how new tools can be introduced and learned over time through shared experiences (Lowndes et al. [2017](#ref-Lowndes17)). This also encourages the development of a community of practice that shares and learns together to navigate the collection of existing and developing open science tools.

The scientific culture within a discipline may inhibit the adoption of open science methods. A common argument against open science is the protection of data that an individual research team may view as proprietary or sensitive. There are reasonable arguments to treat data as personal property, particularly if exceptional effort was spent to secure funding for a project and if the data were hard-earned or sensitive, e.g., detailed location data on endangered species or medical/socioeconomic data (Zipper et al. [2019](#ref-Zipper19)). These issues are less of a concern for bioassessment where many datasets are collected by institutions that are publicly funded and data accessibility may be mandated by law. However, an open science process dictates that both interim and completed research products derived from public data should be available to the broader bioassessment community. This raises an additional concern that research teams using transparent workflows could expose themselves to increased criticism by their peers and the public (Lewandowsky and Bishop [2016](#ref-Lewandowsky16)), particularly where the developed products can have important regulatory implications. However, an argument made throughout this paper is that regulatory, management, and stakeholder groups that will both use and be affected by bioassessment products should be integral contributors to the development process. An open science bioassessment process welcomes criticism and feedback as a natural part of development that will facilitate adoption by ensuring the product meets the needs of all parties. Ultimately, these products are developed to improve the environment as a public resource. The ideals promoted by an open science process directly align with these goals.

Institutional barriers can inhibit open science given the scale of change that must occur for adoption. Bureaucratic hurdles can disincentivize initiatives that promote change, particularly if that change originates from researchers not in administrative roles. Regulatory institutions may also prefer some level of opacity for how research products that influence policy are made available during development. The level of transparency advocated by open science could be viewed as opening the floodgates to increased legal scrutiny that can unintentionally hinder forward progress. Despite these reservations, many public institutions now advocate for increased openness because of the benefits that facilitate and engender public trust. Open data initiatives are now fairly common and represent a form of advocacy by public institutions for broader adoption of open science principles. Although these changes represent progress, many institutions have yet to strictly define open science and how it is applied internally and externally. As open science continues to build recognition, means of integrating toolsets that promote openness and transparency beyond publicly shared data will have to be adopted by regulatory and management institutions.

# Conclusions

The relevance of bioassessment applications can be improved with open science by using reproducible, transparent, and effective tools that bridge the gap between research and management. Many open science tools can improve communication between researchers and managers to expose all aspects of the research process and facilitate implementation to support policy, regulation, or monitoring efforts. Communication ensures that the developed product is created through an exchange of ideas to balance the potentially competing needs of different sectors and institutions. The documentation and archiving of data used to create a bioassessment product also ensures that other researchers can discover and build on past efforts, rather than constantly rebuilding the wheel. Incremental improvements of existing products can reduce the proliferation of site- and taxon-specific methods with limited regional applications by exploring new ways to integrate biological indicators across space and time.

# Author contributions

MB and RM performed the research in the case study. All authors contributed to the conceptual development and writing of the manuscript.

# Acknowledgments

The authors acknowledge support from the San Gabriel River Regional Monitoring Program and the Sanitation Districts of Los Angeles County. We thank Eric Stein for reviewing an earlier draft of the manuscript. We thank Mike McManus for assistance with the conceptual diagrams.

# Figures



Figure 1: A simplified workflow of the open science paradigm (adapted from Hampton et al. ([2015](#ref-Hampton15))). All aspects of the research process, from the conception of an idea to publishing a product, can be enhanced using open science tools. The workflow is iterative where products are continually improved through collaborations facilitated through discovery and reproducibility of open data.

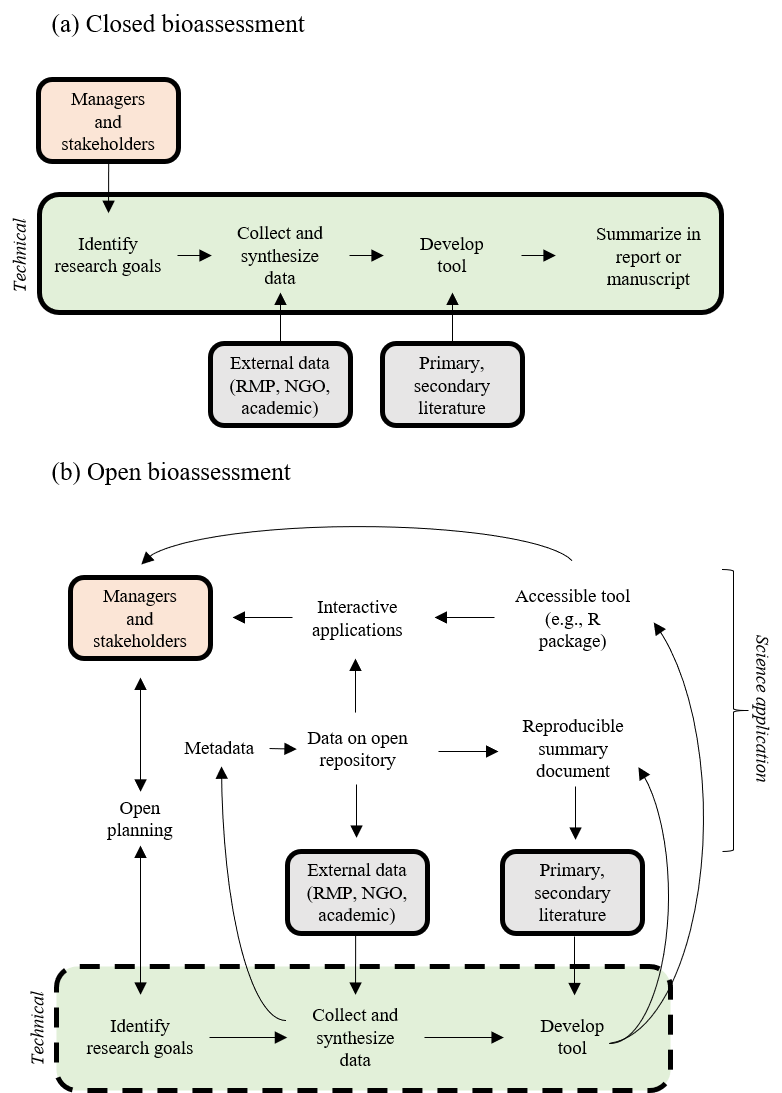


Figure 2: Examples of (a) closed and (b) open approaches for bioassessment applications. The green box in each subfigure represents the technical steps of the individual research team for developing the product, the manager and stakeholder box are those that require or motivate the creation of bioassessment products, the gray boxes indicate sources of external information (data and guidance documents) as input into the technical process, and the open text indicates open components of the planning, application, or implementation phase of a bioassessment product. Figures were adapted from Hampton et al. ([2015](#ref-Hampton15)). NGO: non-government organization, RMP: regional monitoring program.



Figure 3: Schematic demonstrating how the Stream Classification and Community Explorer (SCAPE) can be used to identify potential management actions for stream sites. Stream segment classifications are defined as biologically constrained or unconstrained based on landscape characteristics (left map) and sites with bioassessment scores are evaluated relative to the classifications. Sites can be under-scoring, as expected, or over-scoring relative to the segment classification and expected range of scores (middle plot). Unconstrained sites are those where present landscape conditions do not limit biological potential and constrained sites are those where landscape conditions limit biological potential (right images). Management actions and priorities can be defined based on site scores relative to segment classifications.



Figure 4: The open science workflow for creating the SCAPE tool. The approach was iterative following those shown in Figures 1 and 2b to meet stakeholder needs (box 1; San Gabriel River Regional Monitoring Program, Sanitation District of Los Angeles County, Southern California Coastal Water Research Project). Specific open science or open source applications used at each step are shown in each box with publicly accessible URL links where appropriate.

# Tables

Table 1: Core definitions and principles of open science, from Open Knowledge International, <http://opendefinition.org/>, Creative Commons, <https://creativecommons.org/about/program-areas/open-science/>, and J. Gezelter, <https://www.openscience.org>

|  |  |
| --- | --- |
| Concepts.and.principles | Description |
| Open | Anyone can freely access, use, modify, and share for any purpose |
| Open Science | The practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods |
| Principle 1 | Transparency in experimental methodology, observation, and collection of data |
| Principle 2 | Public availability and reusability of scientific data |
| Principle 3 | Public accessibility and transparency of scientific communication |
| Principle 4 | Using web-based tools to facilitate scientific collaboration and reproducibility |

Table 2: R packages that can be used in the development and application of bioassessment products.

|  |  |  |
| --- | --- | --- |
| Task | Package | Description |
| General | tidyverse (Wickham 2017) | A suite of packages to import, wrangle, explore, and plot data. Includes the popular ggplot2 and dplyr packages. |
| Mapping, geospatial | sf (Pebesma 2018) | A simple features architecture for working with vectorized spatial data, including common geospatial analysis functions |
|  | raster (Hijmans 2019) | Reading, writing, manipulating, analyzing, and modeling gridded spatial data |
|  | leaflet (Cheng et al. 2018) | Integration of R with the popular JavaScript leaflet library for interactive maps |
|  | mapview (Appelhans et al. 2018) | Creates interactive maps to quickly examine and visually investigate spatial data, built off leaflet and integrated with sf |
| Statistical modelling | randomForest (Liaw and Wiener 2002) | Create classification and regression trees for predictive modelling |
|  | nlme (Pinheiro et al. 2018) | Non-linear, mixed effects modelling |
|  | mgcv (Wood 2017) | Generalized additive modelling |
| Community analysis | TITAN2 (Baker et al. 2015) | Ecological community threshold analysis using indicator species scores |
|  | indicspecies (De Caceres and Legendre 2009) | Indicator species analysis |
|  | vegan (Oksanen et al. 2018) | Multivariate analysis for community ecology |
| Science communication | shiny (Chang et al. 2018) | Reactive programming tools to create interactive and customizable web applications |
|  | rmarkdown (Allaire et al. 2018) | Tools for working with markdown markup languages in .Rmd files |
|  | knitr (Xie 2015) | Automated tools for markdown files that process integrated R code chunks |

# References

Allaire, J., Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, and W. Chang. 2018. Rmarkdown: Dynamic documents for r.

Appelhans, T., F. Detsch, C. Reudenbach, and S. Woellauer. 2018. Mapview: Interactive viewing of spatial data in r.

Bain, M. B., A. L. Harig, D. P. Loucks, R. R. Goforth, and K. E. Mills. 2000. Aquatic ecosystem protection and restoration: Advances in methods for assessment and evaluation. Environmental Science and Policy 3:89–98.

Baird, D. J., and M. Hajibabaei. 2012. Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation dna sequencing. Molecular Ecology 21:2039–2044.

Baker, M. E., R. S. King, and D. Kahle. 2015. TITAN2: Threshold indicator taxa analysis.

Beck, M. W. 2018a. SCCWRP/SCAPE: v1.0 (Version 1.0). Zenodo, <http://doi.org/10.5281/zenodo.1218121>.

Beck, M. W. 2018b. Constrained streams for biological integrity in California. Knowledge Network for Biocomplexity. urn:uuid:75411f50-32ed-42a5-bbfd-26833c7a441f.

Beck, M. W., and L. K. Hatch. 2009. A review of research on the development of lake indices of biotic integrity. Environmental Reviews 17:21–44.

Beck, M. W., R. D. Mazor, S. Johnson, K. Wisenbaker, J. Westfall, P. R. Ode, R. Hill, C. Loflen, M. Sutula, and E. D. Stein. (n.d.). Prioritizing management goals for stream biological integrity within the developed landscape context. Freshwater Science.

Beck, M. W., B. Vondracek, and L. K. Hatch. 2013. Environmental clustering of lakes to evaluate performance of a macrophyte index of biotic integrity. Aquatic Botany 108:16–25.

Birk, S., W. Bonne, A. Borja, S. Brucet, A. Courrat, S. Poikane, A. Solimini, W. van de Bund, N. Zampoukas, and D. Hering. 2012. Three hundred ways to assess Europe’s surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. Ecological Indicators 18:31–41.

Bond-Lamberty, B., A. P. Smith, and V. Bailey. 2016. Running an open experiment: Transparency and reproducibility in soil and ecosystem science. Environmental Research Letters 11:1–7.

Borja, A., A. Ranasinghe, and S. B. Weisberg. 2009. Assessing ecological integrity in marine waters, using multiple indices and ecosystem components: Challenges for the future. Marine Pollution Bulletin 59:1–4.

Borregaard, M. K., and E. M. Hart. 2016. Towards a more reproducible ecology. Ecography 39:349–353.

Bried, J. T., S. K. Jog, A. R. Dzialowski, and C. A. Davis. 2014. Potential vegetation criteria for identifying reference-quality wetlands in the south-central United States. Wetlands 34:1159–1169.

Cao, Y., and C. Hawkins. 2011. The comparability of bioassessments: A review of conceptual and methodological issues. Journal of the North American Benthological Society 30:680–701.

Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson. 2018. Shiny: Web application framework for r.

Cheng, J., B. Karambelkar, and Y. Xie. 2018. Leaflet: Create interactive web maps with the javascript ’leaflet’ library.

Collins, S. L. 2016. Vegetation science in the age of big data. Journal of Vegetation Science 27:865–867.

Dale, V. H., and S. C. Beyeler. 2001. Challenges in the development and use of ecological indicators. Ecological Indicators 1:3–10.

Darling, E. S., D. Shiffman, I. M. Côte, and J. A. Drew. 2013. The role of Twiter in the life cycle of a scientific publication. Ideas in Ecology and Evolution 6:32–43.

De Caceres, M., and P. Legendre. 2009. Associations between species and groups of sites: Indices and statistical inference. Ecology.

Fetscher, A. E., R. Stancheva, J. P. Kociolek, R. G. Sheath, E. D. Stein, R. D. Mazor, P. R. Ode, and L. B. Busse. 2013. Development and comparison of stream indices of biotic integrity using diatoms vs. Non-diatom algae vs. A combination. Journal of Applied Phycology 26:433–450.

Fore, L. S., and C. Grafe. 2002. Using diatoms to assess the biological condition of large rivers in Idaho (U.S.A.). Freshwater Biology 47:2015–2037.

Grand, A., C. Wilkinson, K. Bultitude, and A. F. T. Winfield. 2012. Open science: A new "trust technology"? Science Communication 34:679–689.

Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, W. C. Lenhardt, A. MacDonald, W. K. Michener, J. Mudge, A. Pourmokhtarian, M. P. Schildhauer, K. H. Woo, and N. Zimmerman. 2015. The tao of open science for ecology. Ecosphere 6:1–13.

Hampton, S. E., M. B. Jones, L. A. Wasser, M. P. Schildhauer, S. R. Supp, J. Brun, R. R. Hernandez, C. Boettiger, S. L. Collins, L. J. Gross, D. S. Fernández, A. Budden, E. P. White, T. K. Teal, S. G. Labou, and J. E. Aukema. 2017. Skills and knowledge for data-intensive environmental research. Bioscience 67:546–557.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2016. Big data and the future of ecology. Frontiers in Ecology and the Environment 11:156–162.

Hering, D., A. Borja, J. Carstensen, L. Carvalho, M. Elliott, C. K. Field, A. S. Heiskanen, R. K. Johnson, J. Moe, D. Pont, A. L. Solheim, and W. de Bund. 2010. The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. Science of the Total Environment 408:4007–4019.

Hijmans, R. J. 2019. Raster: Geographic data analysis and modeling.

Horsburgh, J. S., M. M. Morsy, A. M. Castronova, J. L. Goodall, T. Gan, H. Yi, M. J. Stealey, and D. G. Tarboton. 2016. Hydroshare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain. Journal of the American Water Resources Association 52:873–889.

Howard, J. K., K. A. Fesenmyer, T. E. Grantham, J. H. Viers, P. R. Ode, P. B. Moyle, S. J. Kupferburg, J. L. Furnish, A. Rehn, J. Slusark, R. D. Mazor, N. R. Santos, R. A. Peek, and A. N. Wright. 2018. A freshwater conservation blueprint for California: Prioritizing watersheds for freshwater biodiversity. Freshwater Science 37:417–431.

Hsu, L., E. Mayorga, J. S. Horsburgh, M. R. Carter, K. A. Lehnert, and s. L. Brantley. 2017. Enhancing interoperability and capabilities of earth science data using the Observations Data Model 2 (ODM2). Data Science Journal 16:1–16.

Idaszak, R., D. G. Tarboton, H. Yi, L. Chistopherson, M. J. Stealey, B. Miles, P. Dash, A. Couch, C. Spealman, D. P. Ames, and J. S. Horsburgh. 2017. HydroShare - a case study of the application of modern software engineering to a large distributed federally-funded scientific software development project. Pages 217–233 *in* J. Carver, N. P. C. Hong, and G. K. Thiruvathukal (editors). Chapter 10 in software engineering for science. Taylor & Francis CRC Press, Boca Raton, Florida, USA.

Ihle, M., I. S. Winney, A. Krystalli, and M. Croucher. 2017. Striving for transparent and credible research: Practical guidelines for behavioral ecologists. Behavioral Ecology 28:348–354.

Jackson, S., and W. Davis. 1994. Meeting the goal of biological integrity in water-resource programs in the US Environmental Protection Agency. Journal of the North American Benthological Society 13:592–597.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser. 1986. Assessing biological integrity in running waters: A method and its rationale. Illinois Natural History Survey, Champaign, Illinois.

Kelly, M. G., S. Birk, N. J. Willby, L. Denys, S.Drakare, M. Kahlert, S. M. Karjalainen, A. Marchetto, J. A. Pitt, G. Urbanič, and S. Poikane. 2016. Redundancy in the ecological assessment of lakes: Are phytoplankton, macrophytes and phytobenthos all necessary? Science of the Total Environment 15:594–602.

Kenney, M. A., P. R. Wilcock, B. F. Hobbs, N. E. Flores, and D. C. Martínez. 2012. Is urban stream restoration worth it? Journal of the American Water Resources Association 48:603–615.

Kerans, B. L., and J. R. Karr. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. Ecological Applications 4:768–785.

King, R. S., M. E. Baker, P. F. Kazyak, and D. E. Weller. 2011. How novel is too novel? Stream community thresholds at exceptionally low levels of catchment urbanization. Ecological Applications 21:1659–1678.

Kluyver, T., B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing. 2016. Jupyter notebooks – a publishing format for reproducible computational workflows. IOS Press.

Kuehne, L. M., J. D. Olden, A. L. Strecker, J. J. Lawler, and D. M. Theobald. 2017. Past, present, and future of ecological integrity assessment for fresh waters. Frontiers in Ecology and the Environment 15:197–205.

Laan, J. J. V., and C. P. Hawkins. 2014. Enhancing the performance and interpretation of freshwater biological indices: An application in arid zone streams. Ecological Indicators 36:470–482.

Lai, J., C. J. Lortie, R. A. Muenchen, J. Yang, and K. Ma. 2019. Evaluating the popularity of R in ecology. Ecosphere 10:e02567.

Lane, B. A., S. Sandoval-Solis, E. D. Stein, S. M. Yarnell, G. B. Pasternack, and H. E. Dahlke. 2018. Beyond metrics? The role of hydrologic baseline archetypes in environmental water management. Enviromental Management 62:678–693.

Lewandowsky, S., and D. Bishop. 2016. Research integrity: Don’t let transparency damage science. Nature 529:459–461.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2:18–22.

Linke, S., E. Turak, and J. Nel. 2011. Freshwater conservation planning: The case for systematic approaches. Freshwater Biology 56:6–20.

Liu, Y., H. Gupta, E. Springer, and T. Wagener. 2008. Linking science with environmental decision making: Experiences from an integrated modeling approach to supporting sustainable water resources management. Environmental Modelling & Software 23:846–858.

Lortie, C. J. 2014. Formalized synthesis opportunities for ecology: Systematic reviews and meta-analyses. OIKOS 123:897–902.

Lowndes, J. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O’Hara, N. Jiang, and B. S. Halpern. 2017. Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1:1–7.

Marwick, B., J. d’Alpoim Guedes, C. M. Barton, L. A. Bates, M. Baxter, A. Bevan, E. A. Bollwerk, R. K. Bocinsky, T. Brughmans, A. K. Carter, C. Conrad, D. A. Contreras, S. Costa, E. R. Crema, A. Daggett, B. Davies, B. L. Drake, T. S. Dye, P. France, R. Fullagar, D. Giusti, S. Graham, M. D. Harris, J. Hawks, S. Heath, D. Huffer, E. C. Kansa, S. W. Kansa, M. E. Madsen, J. Melcher, J. Negre, F. D. Neiman, R. Opitz, D. C. Orton, P. Przystupa, M. Raviele, J. Riel-Salvatore, P. Riris, I. Romanowska, N. Strupler, I. I. Ullah, H. G. V. Vlack, E. C. Watrall, C. Webster, J. Wells, J. Winters, and C. D. Wren. 2016. Open science in archaeology. SAA Archaeological Record 17:8–14.

Mazor, R. D., A. C. Rehn, P. R. Ode, M. Engeln, K. C. Schiff, E. D. Stein, D. J. Gillett, D. B. Herbst, and C. P. Hawkins. 2016. Bioassessment in complex environments: Designing an index for consistent meaning in different settings. Freshwater Science 35:249–271.

Mazor, R. D., E. D. Stein, P. R. Ode, and K. Schiff. 2014. Integrating intermittent streams into watershed assessments: Applicability of an index of biotic integrity. Freshwater Science 35:459–474.

McDonald, M., R. L. Blair, D. Bolgrien, B. S. Brown, J. J. Dlugosz, S. S. Hale, S. F. Hedtke, D. T. Heggem, L. E. Jackson, K. B. Jones, B. M. Levinson, R. A. Linthurst, J. Messer, A. R. Olsen, J. F. Paul, S. G. Paulsen, J. L. Stoddard, K. Summers, and G. D. Veith. 2004. The US Environmental Protection Agency’s Environmental Monitoring and Assessment Program. Pages 649–668 *in* G. B. Wiersma (editor). Environmental monitoring. CRC Press, Boca Raton, Florida, USA.

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7:330–342.

Mitchell, B. 2005. Integrated water resource management, institutional arrangements, and land use planning. Environmental Planning A: Economy and Space 37:1335–1352.

Nichols, S. J., L. A. Barmuta, B. C. Chessman, P. E. Davies, F. J. Dyer, E. T. Harrison, C. P. Hawkins, I. Jones, B. J. Kefford, S. Linke, R. Marchant, L. Metzeling, K. Moon, R. Ogden, M. Peat, T. B. Reynoldson, and R. M. Thompson. 2016. The imperative need for nationally coordinated bioassessment of rivers and streams. Marine and Freshwater Research 68:599–613.

Nielsen, M. 2011. Reinventing discovery: The new era of networked science. Page 272Reprint. Princeton University Press, Princeton, New Jersey, USA.

Niemi, G. J., and M. E. McDonald. 2004. Application of ecological indicators. Annual Review of Ecology, Evolution, and Systematics 35:89–111.

Ode, P. R., A. C. Rehn, R. D. Mazor, K. C. Schiff, E. D. Stein, J. T. May, L. R. Brown, D. B. Herbst, D. Gillett, K. Lunde, and C. P. Hawkins. 2016. Evaluating the adequacy of a reference-site pool for ecological assessments in environmentally complex regions. Freshwater Science 35:237–248.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O’Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. 2018. Vegan: Community ecology package.

Paul, M. J., D. W. Bressler, A. H. Purcell, M. T. Barbour, E. T. Rankin, and V. H. Resh. 2008. Assessment tools for urban catchments: Defining observable biological potential. Journal of the American Water Resources Association 45:320–330.

Pebesma, E. 2018. Sf: Simple features for r.

Pilkington, M. 2016. Blockchain technology: Principles and applications. Pages 225–253 *in* F. X. Olleros and M. Zhegu (editors). Research handbook on digital transformations. Edward Elgar Publishing, Cheltenham, United Kingdom.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team. 2018. nlme: Linear and nonlinear mixed effects models.

Poikane, S., N. Zampoukas, A. Borja, S. P. Davies, W. van de Bund, and S. Birk. 2014. Intercalibration of aquatic ecological assessment methods in the European Union. Environmental Science & Policy 44:237–246.

RDCT (R Development Core Team). 2018. R: A language and environment for statistical computing, v3.5.1. R Foundation for Statistical Computing, Vienna, Austria.

Rehn, A. C., R. D. Mazor, and P. R. Ode. 2018. An index to measure the quality of physical habitat in California wadeable streams. California Water Boards, Surface Water Ambient Monitoring Program, California Department of Fish; Wildlife, Southern California Coastal Water Research Project, Sacramento, California.

Shoredits, A. S., and J. A. Clayton. 2013. Assessing the practice and challenges of stream restoration in urbanized environments of the USA. Geography Compass 7:358–372.

Slater, L. J., G. Thirel, s. Harrigan, O. Delaigue, A. Hurley, A. Khouakhi, I. Prosdocimi, C. Vitolo, and K. smith. 2019. Using R in hydrology: A review of recent developments and future directions. Hydrology and Earth System Sciences 23:2939–2963.

Stein, E. D., M. Brinson, M. C. Rains, W. Kleindl, and F. R. Hauer. 2009. Wetland assessment alphabet soup: How to choose (or not choose) the right assessment method. Society of Wetland Scientists Bulletin 26:20–24.

Stoddard, J. L., A. T. Herlihy, D. V. Peck, R. M. Hughes, T. R. Whittier, and E. Tarquinio. 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society 27:878–891.

Taylor, J. M., J. A. Back, B. W. Brooks, and R. S. King. 2018. Spatial, temporal and experimental: Three study design cornerstones for establishing defensible numeric criteria in freshwater ecosystems. Journal of Applied Ecology 55:2114–2123.

Taylor, J. M., R. S. King, A. A. Pease, and K. O. Winemiller. 2014. Nonlinear response of stream ecosystem structure to low-level phosphorus enrichment. Freshwater Biology 59:969–984.

Touchon, J. C., and M. W. McCoy. 2016. The mismatch between current statistical practice and doctoral training in ecology. Ecosphere 7:e01394.

Wickham, H. 2009. Ggplot2: Elegant graphics for data analysis. Springer-Verlag. New York.

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

Wickham, H. 2015. R packages. Page 182. O’Reilly, Sebastopol, California.

Wickham, H. 2017. Tidyverse: Easily install and load the ’tidyverse’.

Wickham, H., J. Hester, and W. Chang. 2018. Devtools: Tools to make developing r packages easier.

Wilkinson, L. 2005. The grammar of graphics. Page 691second. Statistics; Computing, Springer, New York.

Woelfle, M., P. Olliaro, and M. H. Todd. 2011. Open science is a research accelerator. Nature Chemistry 3:745–748.

Wood, S. N. 2017. Generalized additive models: An introduction with r. Page 476, 2nd editions. Chapman; Hall, CRC Press, London, United Kingdom.

Xie, Y. 2015. Dynamic documents with R and knitr, 2nd editions. Chapman; Hall/CRC, Boca Raton, Florida.

Yuan, L. L. 2004. Assigning macroinvertebrate tolerance classifications using generalised additive models. Freshwater Biology 49:662–677.

Zipper, S. C., K. S. S. Whitney, J. M. Deines, K. M. Befus, U. Bhatia, S. J. Albers, J. Beecher, C. Brelsford, M. Garcia, T. Gleeson, F. O’Donnell, D. Resnik, and E. Schlager. 2019. Balancing open science and data privacy in the water sciences. Water Resources Research 55:1–10.