Improving the IEA approach using principles of open data science

Kimberly Bastille[1,7]   Sean Hardison[2]   Lynn DeWitt[3]
Jennifer Brown[4,5]   Jameal Samhouri[6]   Sarah Gaichas[7]
Sean Lucey[7]   Kelly Kearney[8]   Ben Best[9]   Scott Cross[10]
Scott Large[7]   Ellen Spooner[11,12]

Abstract

As implemented so far, integrated ecosystem assessments (IEAs) are often bespoke products specific to regional policy needs at varying spatial scales. Data acquisition, management, processing, analysis, and communication are universal challenges that apply across the disparate applications of IEAs. Embracing open science, which is defined as public access to scientific data, methods, and products, along with the tools facilitating open science, have been suggested as solutions to these data challenges. Here we provide a snapshot of the state of open science practices in IEAs ongoing across the United States. We show that open science has improved the flexibility, reproducibility and efficiency of the scientific workflows within the IEA framework. Each scientific workflow in place today has progressed incrementally with each iteration and still have opportunities for further advancement. Though the initial time investment to learn new open science software skills is substantial the return is far greater. The rapid product turnaround and the highly collaborative nature of IEAs make open science extremely useful in advancing toward Ecosystem Based Management (EBM).

[1] Integrated Statistics, Woods Hole, MA, USA

[2] University of Virginia, Charlottesville, VA, USA

[3] National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Southwest Fisheries Science Center, Environmental Research Division, Santa Cruz, CA, USA

[4] ECOS Consulting LLC, Lafayette, CA, USA

[5] National Oceanic and Atmospheric Administration, Channel Islands National Marine Sanctuary, Santa Barbara, CA, USA

[6] National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Northwest Fisheries Science Center, Conservation Biology Division, Seattle, WA, United States of America

[7] National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Northeast Fisheries Science Center, Woods Hole, MA, USA

[8] University of Washington, Joint Institute for the Study of the Atmosphere and Ocean, Seattle, WA, USA

[9] EcoQuants LLC, Santa Barbara, CA, USA

[10] National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, National Centers for Environmental Information, Charleston, SC, USA

[11] ECS Federal, Fairfax VA, USA

[12] National Oceanographic and Atmospheric Administration, Fisheries Office of Science and Technology, Integrated Ecosystem Assessment Program, Silver Spring, MD, USA

Introduction

As implemented so far, integrated ecosystem assessments (IEAs) are often bespoke products specific to regional policy needs at varying spatial scales. While assessments are typically unique, they generally conform to the IEA framework outlined in (Levin et al. 2009). The IEA framework is an iterative process that consists of a scoping phase to define management objectives, the identification of ecosystem indicators to monitor key ecosystem components, the compilation of ecosystem status reports assessing status and trends of indicators, a risk assessment phase identifying where ecosystem considerations might threaten management objectives, and finally management strategy evaluation (MSE), in which potential management actions are tested using simulation models. As the approach is iterative, any management actions are followed with continued monitoring and reassessment of management goals (Levin et al. 2008; Levin et al. 2009; Samhouri et al. 2013). The elements of the IEA framework are central to its implementation, but its flow is not prescriptive (Harvey et al. 2016). Instead, IEAs are malleable to the needs of stakeholders, meaning that each IEA manifests with unique structural requirements.

Data acquisition, management, processing, analysis, and communication are universal challenges that apply across the disparate applications of IEAs. The number of entry and exit points for data in an IEA is large, and outputs vary by end-user need. For example, data may enter into an IEA during the exploratory phases of the process, where objectives are scoped and representative indicators are developed. Exit points for data products can be found in the derived indicator data informing ecosystem status reports and risk assessments. Simulated products (e.g. model output summaries) and model parameters, such as those resulting from MSE model runs, are also important to disseminate, as these results inform management actions. Any data that leaves or enters an IEA brings with it metadata and code used to analyze, process, and visualize the data. Through the processing of these datasets, accompanying code and metadata documentation become scientific and technical products in their own right, feeding into subsequent IEA outputs (Figure 1). Adding to these complexities are the diversity of audiences to which IEA products appeal, including

Define EBM Goals & Targets

Develop Indicators

Evaluate Strategies
Management Strategy Evaluation, Summary products, Web-based visualizations, Model inputs, outputs and parameters, Analysis/Visualization code, R packages*

Develop Indicators
Indicator data sets, Web-based visualization apps, Metadata documentation, R packages/Shared R scripts

Implement Management Action

Evaluate and Assess Outcomes

Monitoring of Ecosystem Indicators

Evaluate Strategies

Analyze Uncertainty & Risk

Assess Ecosystem

Analyze Uncertainty & Risk
Risk Assessments, Methods documentation, Analysis/Visualization code

Assess Ecosystem
Ecosystem Status Reports, Report templates, Methods documentation, Analysis/Visualization code
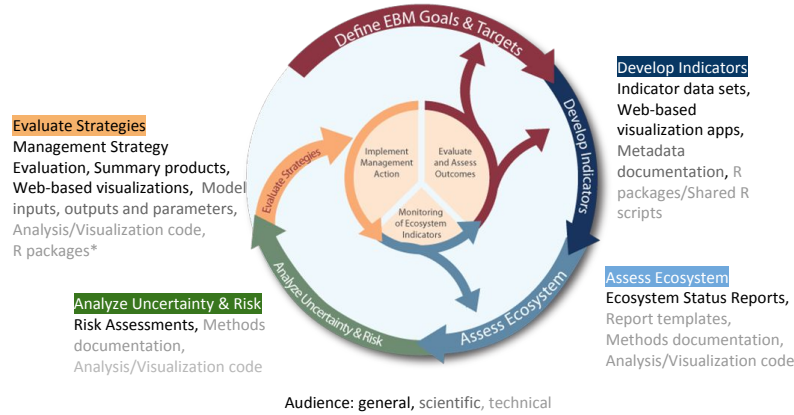
Audience: general, scientific, technical

Figure 1: Products as part of the IEA loop according to audience: general, scientific and technical (*In development).

both technical and non-technical users.

Embracing open science, which is defined as public access to scientific data, methods, and products (Hampton et al. 2015; McKiernan et al. 2016), along with the tools facilitating open science, have been suggested as solutions to these data challenges (Michener and Jones 2012; Hampton et al. 2015; Lowndes et al. 2015; Lowndes et al. 2017; Ma et al. 2018). Michener and Jones (2012) presented examples of how developing data ontologies, which represent formal descriptions of scientific concepts and their relationships, can streamline the synthesis of unique, cross-disciplinary datasets for analysis. Lowndes et al. (2017) demonstrated that embracing open science methods resulted in increased efficiency, transparency, and repeatability in the development of the Ocean Health Index; an iterative framework developed to assess the benefits of marine ecosystems to humans for sustainable management. Case studies for applying open science processes to IEAs have also been developed for ecosystem reporting in the Northeast Large Marine Ecosystem. Specifically, Ma et al. (2018) applied the W3C PROV Ontology (Lebo et al. 2013) to establish machine readable provenance for an ecosystem assessment using IPython, taking a critical advancement towards fully reproducible reporting strategies for IEA.

In this paper, we present the various open science strategies being implemented within different phases of the IEA framework across regions in the United States. While this implementation is still nascent, we show that open science in IEAs has improved the flexibility of assessments for better responsiveness to stakeholder needs, prevented IEAs from becoming "black boxes" through greater transparency, and increased efficiency in data tasks through automated meth-

ods. The accurate capture of data provenance has been recognized as a key component of actionable science (Ma et al. 2014), which IEAs strive towards. Here, we argue that the continued use and improvement of open science techniques in IEAs will further improve their actionability for implementing ecosystem-based management.

Scientific workflow

Within each analytical phase of the IEA framework (ecosystem assessment, risk assessment and strategy evaluation) (Figure 1), the general scientific workflow is similar (Figure 2a). The emphasis may vary between phases of the IEA framework, but the overall workflow (import, analyze, visualize and communicate) remains (Wickham and Grolemund 2016). Adoption of open science principles allows IEA practitioners to efficiently transfer knowledge between phases of the IEA framework (Figure 1). Below we outline the scientific workflow and highlight some open science applications of common software tools used in IEAs (Figure 2b). These examples by no means cover all of the tools available nor all of possible applications of those mentioned, but are representative of the methods currently in use across IEA regions in the United States. Throughout the scientific workflow there are many opportunities to apply open science principles and tools, but there is no prescriptive outline to follow. The iterative nature of exploratory data analysis (Tukey 1977; Wickham and Grolemund 2016) suggests the loop (Figure 2a), whereby upon visualizing outputs, outliers can be handled, models improved, ancillary data imported, visualizations modified and communication strategies updated.

Import

Handling data in a way that is both reproducible and transparent begins with the initial import and good data management practices must start here. IEAs present specific challenges for data import, as indicators are often iteratively updated and may change as management priorities fluctuate through time. The sources of these data are varied and contributors may have their own data management systems, leading to issues when tracking metadata and data provenance. Leveraging open science software tools to control the import, storage, and management of datasets can help address many of these challenges (Yenni 2019).

Two systems for identifying and importing data into any phase of the IEA framework are data catalogs and data services. Data catalogs serve as an inventory of available data. They present metadata so at a glance the data consumer can determine if the full dataset meets their needs. Data services on the other hand provide a standard way of accessing data and allow for machine-to-machine transport. Once a data service is set up, updating a dataset can be as easy as
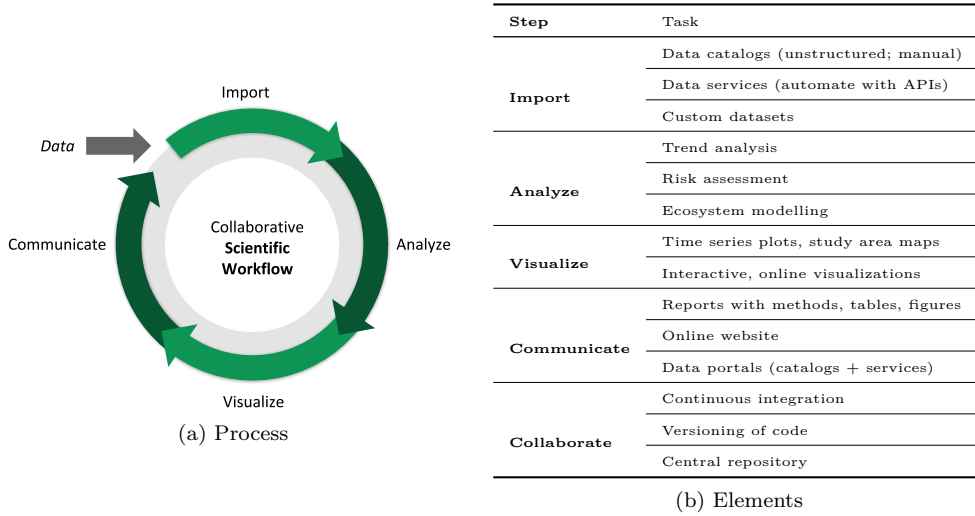
(a) Process

Step	Task
Import	Data catalogs (unstructured; manual)
	Data services (automate with APIs)
	Custom datasets
Analyze	Trend analysis
	Risk assessment
	Ecosystem modelling
Visualize	Time series plots, study area maps
	Interactive, online visualizations
Communicate	Reports with methods, tables, figures
	Online website
	Data portals (catalogs + services)
Collaborate	Continuous integration
	Versioning of code
	Central repository

(b) Elements

Figure 2: Scientific workflow.

dropping the updated file into a directory. The data service ERDDAP (Simons 2019) is a free and open source software tool that can serve gridded and tabular datasets in a variety of formats (.csv, .png, .nc, etc.) through a single interface. ERDDAP provides direct access to indicator data for IEA scientists and the public, and datasets can be imported programmatically (e.g. using `rerddap` (Chamberlain 2019) in R). Both data catalogs and data services provide more transparent and efficient paths of importing data compared to traditional methods like email. As more IEA contributors develop data catalogs and services, the network of publicly available, properly documented IEA source data products will continue to grow.

Data Provisioning

A data provisioning strategy, through which the dissemination of data between collaborators and with the public is defined, is important for the development and execution of the IEA process. Establishing a data service or central repository for indicator data provides access to the data for all collaborators, and in addition provides a single authoritative source for the latest data and metadata. Data services such as ERDDAP provide direct access to data using open science tools and scripts (Figure 2b). DataOne provides a full archive of a dataset in most any format, but datasets are not easily subsettable or translated into alternative formats like ERDDAP. Both ERDDAP and DataOne can have standardized metadata, e.g. in Ecological Metadata Language (EML) associated with datasets that describe the data and its derivation, i.e. methods. Inputting and maintaining metadata is one of the more onerous yet essential tasks of data

management to ensure data integrity and proper usage. A short-term, but less-standardized strategy for provisioning data is to use Github to store the data files, such as done with `ecodata` (github.com/NOAA-EDAB/ecodata) in the Northeast US IEA region. This method is feasible for short tabular datasets (e.g. .csvs), but continually updated, long-term data storage is not the intended purpose for space-limited Github repositories, and a more robust data provisioning software is typically warranted. By utilizing any of these data services, standard methods for data access can be used to import, analyze and visualize the data (Figure 2a). Rather than downloading data and keeping private copies on individual computers, specific versions and data sources can be flagged and products updated in an automated fashion as the data are revised.

Groups may choose between storing data and indicators in a central repository with a data service providing access, or keeping control of the data within the institution implementing the IEA. In the latter case, data can be distributed among institutions, each institution installing their own data service, and a single data service at one of the institutions can act as a virtual central repository. Installing a data service can seem daunting, but there are tutorials and classes available, and institutions with established data services are often more than happy to help other institutions set up their own servers or serve their data. Once a data service is set up, and IEA practitioners have decided on a standard data format, updating a dataset in a data service such as ERDDAP can be as easy as dropping the updated file in a directory. Efficient indicator storage and data management practices such as these translate into quicker product development in the following phases of the IEA framework.

Analyze

The analysis performed in each phase of the IEA framework varies from simple trend analysis to more complex ecosystem modelling and simulation. To accomplish these disparate tasks, software tools have been developed and deployed to streamline each phase of the IEA framework. Software packages used to store related functions specific to the needs of regional IEAs have been built embracing open science principles (for example, several IEA software products are under development at github.com/NOAA-EDAB). These packages handle everything from simple to complex analyses and can be made available to the public which promotes reproducibility and transparency. Standard packaging of functions creates inline documentation and allows others in the open-source community to contribute improvements for approval by the package author. In some cases it is useful to rely on tools already in existence like Ecopath with Ecosim, an ecosystem modelling and simulation tool which is free, publicly available and well documented (Christensen and Walters 2004). The analysis step in the scientific framework is important and decisions made here should be defendable, the application of open science will make this step more transparent and thus trustworthy.

Visualize

Once analyzed, information must be shared in a digestible format, most likely some intuitive visualization. To produce understandable visualizations and make them available to resource managers and the public, IEA practitioners must package complex ecosystem information quickly and repeatedly. Application of software tools have greatly decreased the time spent building visualizations resulting in more rapid and effective communication of IEA products.

Best practices for data visualization are easily implemented in the `ggplot2` R library (Wickham 2016), which implements the "grammar of graphics" (Wilkinson 2012). These static visualizations can further be made interactive, e.g. zoom, pan and turn on/off series in an online time series plot, by simply feeding a `ggplot2` plot object to the `ggplotly` function in the plotly library (Sievert 2019). The plotly library, like most R packages producing interactive online visualizations, uses the htmlwidgets framework (htmlwidgets.org) to wrap JavaScript visualization libraries into easy to use R functions.

Communicate

Perhaps the most important step in the scientific workflow is the communication of products developed for regional IEAs. As shown in Figure 1, the diversity of IEA products is further complicated by the many audiences that must also be appealed to. Several tools have been used by IEA teams to meet this challenge; perhaps most notably the use of interactive websites allowing users to visualize and interact with IEA data at multiple technical levels (e.g. https://mbon.ioos. us/). These websites, with associated reports and data portals, all must be established and maintained throughout the often multi-year IEA process.

Some IEA teams have chosen to use Rmarkdown (Baumer et al. 2014) to build their yearly reports, because scripting-based methods of reporting are typically more reproducible than traditional approaches like Microsoft Word (Lowndes et al. 2017). These methods allow for template development to ensure cohesive products between years, and also for the simultaneous editing of code and text for simplifying workflows. This approach removes the need to copy/paste figures into a document. Programmatic document creation can easily be translated to other communication types (presentations, books, articles, etc.) facilitating the delivery of information to a variety of audiences. All of these features, and especially when combined with public code repositories and continuous integration software, promote transparency and create a balance between automation and flexibility.

Web applications and data portals play an important role within an IEA. Web applications can provide a means of communicating indicator status and trends to managers, fisheries councils, and the public, and can be designed to customize the experience for particular interest groups. The technical barriers to developing useful web tools for communicating IEA products has declined with the

advent of HTML/CSS/JS wrappers for scientific programming languages like R and Python. For example, websites may be generated from static Rmarkdown and Jupyter Notebook documents (Kluyver et al. 2016) for sharing IEA products. Further, web applications can directly access data services such as ERDDAP in order to always serve the most recent indicator data. Data portals act as a tool in indicator data discovery and exploration, a source for data import and a place to view indicators in a central location, which also makes them a useful communication tool.

Collaborate

Collaboration is built into all aspects of the IEA framework and associated scientific workflows. IEAs must include input from stakeholders, management bodies and multiple scientific disciplines (Levin et al. 2009, @samhouri2014), and several open science methods have been used to facilitate this communication. Having a central repository with the latest versions of files is essential for internal collaboration. For code, the most common open-source central repository in use for IEAs is Github. Github works with the version-control software Git to track progress, roll back changes, and prevent loss of work amongst collaborators (Ram 2013; Chacon and Straub 2014).

The institutional memory gained by using Github project management, e.g. with issues and messages from commits, means that new team members can quickly view the progression of a project and contribute. This is especially useful in IEAs in which there are many contributors and those contributors may change through time. Combining version control with a central repository, like Github, means that all interested parties have access to IEA work and team members can collaboratively and simultaneously work on projects while on different machines without losing individual work products.

Internal reproducibility is essential for collaborative work and should be a focal component of IEA development to ensure efficiency. For R users, this may be achieved through the use of packages like `here` (Müller 2017), which enhances the portability of scripts across collaborator computers by managing paths with respect to the root of the repository. Continuous integration (CI) is another software developer concept useful to the IEA process. By utilizing the "freemium" model of CI services such as Travis CI, websites and reports can be automatically updated for free with the latest data and content on a regular basis (e.g. daily or monthly) and when any files in an associated public repository are changed. If anything breaks in the scripts then maintainers can be notified via email. Together these software tools have improved the scientific workflow within all phases of the IEA framework.

Define EBM Goals & Targets, Develop Indicators (IEA Process)

Clear and achievable goals and targets are necessary for priority setting the IEA workflow (Levin et al. 2009), and establishing these targets requires flexible collaboration between many stakeholders. Collaborations between scientists, stakeholders, and managers may be enhanced through the adoption of open science methods. For example, the development of interactive data portal websites, such as those from the Marine Biodiversity Observation Network (https://mbon.ioos.us/), Mid-Atlantic Ocean Data Portal (http://portal.midatlanticocean.org/), or Northeast Ocean Data Portal (https://www.northeastoceandata.org/) make IEA-related science accessible for stakeholders and managers without technical backgrounds. For stakeholders, lowering the barriers to entry for accessing science products will make for more informed IEA goal development.

Embracing open science practices during this phase of the IEA process will benefit scientists as well. As Hampton et al. (2015) advocated, "free and unfettered access" promotes the discovery of relevant datasets that could be used as ecosystem indicators. Further, the amount and spatiotemporal scales of open-sourced information available from a wide variety of stakeholders and public repositories creates a crosswalk between goal setting and indicator development. Open science principles allow data consumers to quickly take inventory of what data exist and at what scale relevant to their data needs (Wilkinson 2016).

Developing indicators that are both of interest to managers as well as relevant in explaining ecosystem dynamics is a vital phase of the IEA framework (Levin et al. 2009). It can be challenging to identify indicator datasets if they are not stored together or if the are not publicly available. As previously stated, IEAs are working with data catalogs and data servers to apply open science principles to overcome these challenges. Data provenance also plays an important role in indicator development and adds transparency to derived datasets (Ma et al. 2018).

Assess Ecosystem (Northeast Ecosystem Assessments)

Regular production of ecosystem assessments are important for the continued tracking of how the system is changing (Levin et al. 2009; Zador et al. 2017) and therefore have become a staple in IEAs. Though intervals between assessments vary by region, the need to reiterate workflows persists, and developing automated workflows for common tasks can greatly enhance efficiency and reproducibility of these workflows. As discussed by Lowndes et al. (2017), automating key tasks allows for quicker product turnaround, increased time spent interpreting indicators, and effectively communicating the findings. In

the Northeast IEA region, the IEA team has developed a workflow to automate many aspects of their ecosystem assessment reports. From initial data import to ecosystem assessment production, each step in the scientific workflow has included streamlining tools to enhance reproducibility and collaboration.

The NE IEA workflow has progressed steadily in its use of open science principles over the past few years. Each iteration of the annual ecosystem assessment brings improvements to the workflow and identification of new opportunities for streamlining. The NE IEA team produce two ecosystem assessment (State of the Ecosystem, SOE) reports each year covering multiple large marine ecosystems. The variety of spatial and temporal scales described by data in these reports create a complex problem for data management and directory organization. The workflow constructed to manage these complexities is as follows.

With each iteration, data contributors deliver a diverse array of datasets, both in structure and scientific domain. Processing code for these data sets, which converts them from native formats to a single, standard format used within ecosystem reporting workflows, is included with an open-source R package containing the derived indicator data (Hardison 2018). Packages associated with ecosystem reporting are continually updated as new indicators are developed, and outputs from these packages are directly incorporated into SOE reports and other IEA products. Report production is facilitated by the use of Rmarkdown for automating the production of ecosystem assessments using report templates created to maintain cohesive report formatting from year to year. Team members are able to collaboratively write and edit both text and code through Github, further expediting the process. Each indicator used in the reports is documented using a comprehensive methods document where all contributors can collaboratively edit and update individual indicator chapters (NOAA-EDAB 2019). This workflow allows for both automation and flexibility where needed, and continues to be improved upon.

All team members working on the reports, packages, or templates use the same coding language and directory structures, which greatly increases the ease of collaboration internally. The efficiency and institutional memory gained through the transition to open science principles has allowed for more reproducible and coherent products from year to year, while also allowing the products to grow and adapt to the needs and requests of the primary audience (Fisheries Management Councils).

Analyze Uncertainty & Risk (California Current Fisheries)

The California Current IEA (CCIEA) is an example of a region that has embraced open science principles and built on them over the years. CCIEA reports are generated from data spanning the entire West Coast, numerous data

providers, and many different ecosystems. Members include scientists from two Fisheries Science Centers, and many private foundations and academic institutions. The CCIEA had foresight early in the development process to agree upon a standard data format for sharing of indicators and a standard plot for indicators that included statistics of interest. Shared R scripts were developed to read the standard format and generate the plots. Today, all CCIEA indicator data is served by ERDDAP, making it available to be used directly by these same R scripts, as well as by an interactive web application that provides time series plots and tables of indicator status and trends.

In 2019, this data management technique was used to conduct an ecological risk assessment (ERA) for nine California fisheries (Samhouri et al. 2019). The open science principles applied to develop other phases of the IEA framework were directly implemented in this ERA. The time spent developing the ERA decreased due to the efficiency built into the data management steps of the scientific workflow, which allowed for more time dedicated to interpretation and dissemination of the results. In the future, this groundwork can be leveraged to spin up new ERAs, as the current structure exists in a repeatable format. Like all phases of the IEA process, risk assessments require the participation of many stakeholders. The assessments are linked to management decisions and management objectives, which are often focused on balancing the needs of a variety of stakeholders. The participatory process and transparency in the development of the ERA may increase stakeholders' trust in the assessment and therefore its application in management (Samhouri et al. 2019). Public cooperation in establishing the priorities and publicly available code and data highlight the adoption of open science principles by the IEA team. This adds transparency, trust, and possibly greater adherence to management decisions built on this work. In addition, it increases the potential for transferability and adaptation to decision contexts in the California Current region and beyond.

As a second example in the California Current, the development of the EcoCast framework (Hazen 2018) and tool (Welch 2018) relies on open science principles to create a user-friendly interface for determining risk to protected species from fishing. Specifically, this approach describes the expected distribution of species targeted by a fishery (e.g., swordfish) in relation to the expected distribution of species caught in that fishery by accident (e.g., leatherback turtles). These predictions of species' distributions are made in near-real time and available online, enabling managers and fishermen to make daily decisions about where it is more or less risky to fish in order to avoid bycatch. This tool therefore informs a dynamic ocean management approach (Lewison et al. 2015) that can improve fisheries sustainability . It was developed in collaboration between SWFSC, NMFS Western Regional Office, Council advisory body staff, reflecting the collaborative approach underlying open science principles. All of the code, both describing the development of EcoCast and its operationalization, are freely available on GFithub.

Evaluate Management Strategies (Pacific Islands Coral Reef Management)

The final phase of the IEA framework is to evaluate potential management strategies (Figure 1). The Puakō coral reef ecosystem management strategy evaluation (MSE) used open science principles at several points of MSE development. Simulation studies such as these carry with them a variety of input, output and parameterization data that can greatly benefit from the application of open science principles. Input data used to conduct the Puakō coral reef ecosystem MSE were sourced from a literature review by scientific partners (Weijerman et al. 2018) and modelled using free and publicly available ecosystem modelling software, Ecopath with Ecosim (EwE; version 6.4.4).

Many other opportunities exist for using open science principles in MSE development. Publicly available code for standardized analysis and visualization could increase reproducibility and incorporation of code and text in the report production could greatly increase communication efficiency. Automation and streamlining of this scientific workflow could lead to faster MSE production in other regions through enhanced reproducibility and institutional memory or could allow for more management strategy simulations as time needed to analyze, visualize and communicate work is reduced. As in all other IEA phases, further adoption of open science is needed.

Discussion

The implementation of IEAs is a challenge in data science as much as it is a challenge in the traditional hard sciences. Each data science challenge faced by IEA teams is unique due to the place-based nature of IEAs and EBM more broadly. However, overlap in workflows between regions implementing IEAs suggests key takeaways from the work accomplished so far.

Due to the diverse nature of datasets used in IEAs, it is crucial to standardize and automate data formatting upon submission whenever possible. The development of tools for massaging data into usable formats may have steep learning curves, but greatly reduce the overall effort in subsequent IEA iterations. However, despite standardization efforts, data and metadata are often inconsistent from year to year. While this can be due to human error or events such as staff turn over, it could also be due to changes in ecosystems and priorities. Naming conventions that worked one year may fail the next due to unexpected changes in sampling regions or species behavior. Ecosystem events (for example the "warm blob") cause changes in survey and reporting priorities. Any IEA data workflow must be flexible enough to handle such changes without becoming unduly labor intensive. Each IEA team must establish the appropriate balance between automation and flexibility.

The use of software development tools within the scope of IEA will only get the IEA practitioner so far. Perhaps the larger challenge in the IEA process is impressing upon collaborators that their contributions are valued and their time well spent. Efficient data pipelines can lower the barrier to entry for scientists wishing to contribute their work to an IEA. However, that does not keep collaborators engaged through what is likely to be a multi-year project spanning several distinct periods. Strategies have been developed to smooth this process. As discussed, some IEA teams install contributor data into publicly accessible data servers with web-based analysis and visualization tools (Figure 2b). Data services such as these can incentivise long-term collaboration by adding value, visibility and proper attribution to both the contributor and their work.

The placed-based nature of EBM determines the spatial scale of datasets used in IEA products. In the case of the Channel Islands National Marine Sanctuary, the use of openly available data lead to a better description of ecosystem state. Three different long term monitoring surveys that collect samples within the sanctuary have regional extents much larger than the sanctuary; the California Cooperative Fisheries investigations Ichthyoplankton surveys and Northwest Fisheries Science Center's Shelf Rockfish Hook and Line surveys cover much of Southern California waters and the Northwest Fisheries Science Center Groundfish Bottom Trawl survey covers the entire U.S. west coast. In each case, the data was subset to include one analysis using only samples collected in the sanctuary paired with another analysis that included samples from the large region in which the sanctuary resides. This is an example of down-scaling from larger regional monitoring to a scale that is relevant for the management area, in this case the marine sanctuary.

The issue with scale can also work in the other direction. On the east coast, where many states are included in the large marine ecosystem management area, smaller state-specific coastal surveys must be combined to cover the total managed area. Finding statistically sound datasets in the appropriate scale is important and sometimes difficult. Public access to data of varying scales can help identify and fill gaps in data of the appropriate scale.

Making all data publicly available may not be the best practice for all datasets. Data can sometimes be sensitive in nature and thus must be handled appropriately before being made available to the masses. This is often the case for fisheries dependent data that uses catch location or in circumstances when knowledge of the exact location could put a species at risk of poaching, for example ESA-listed black abalone, *Haliotis cracherodii*, along the California coast. It is standard to aggregate data in these circumstances and, though resolution is lost, the usability of the dataset increases. Password controls on services such as ERDDAP can provide a place for sensitive data while allowing access to scripts that aggregate and prepare indicators for public access. Individual IEAs can choose a method of data aggregation that suits their unique situation. Though we advocate transparency and publicly accessible data, this may not be

practical for all data types and thoughtful action must be taken when handling sensitive datasets.

Coming together around a single development environment that works for most parties will increase efficiency and minimize pain points in the development process by ensuring that team members feel that they're able to contribute and understand the scripts being shared that concern their work. This also increases the flexibility in which the IEA team can adapt to new projects. Project development workflows, with standardized directory structures containing unique locations for data, processing and analytical scripts, and output documents will greatly enhance productivity. These practices will address challenges in "starting over" for each application of the IEA process. Strive to ensure that your scripts will run with little to no effort on your team member's machines. Embracing relative working directories (e.g. making use of the 'here' package for R (Müller 2017)) will go a long way towards getting you there. This will also help stakeholders/scientists use your work if they choose to. A great launching off point for preparing for the shift to these principles may be the participation in a data literacy workshop (ex. Software Carpentry series) that is focused on reproducibility and open science principles.

Call to action

Each scientific workflow in place today started with small steps and has progressed incrementally with each iteration. We recommend starting small while keeping in mind the possibility for growth. To get started, we suggest three relatively simple action items: prior to beginning a new IEA, choose a single development environment for the group to learn and work in, establish and adhere to project development workflows, and advocate for group participation in a data literacy workshop or other technical trainings (Wilson 2006; Lowndes et al. 2017). These initial steps will help get your group engaged with and (hopefully) excited about open-science techniques.

Each regional IEA team applies open science principles differently, and though the scientific workflow is unifying throughout, its implementation is varied and constantly improving. While there is an upfront cost to learning new technical skills and becoming familiar with the programs discussed in this paper, we suggest that the return on this investment will be worth it. The reproducibility and efficiency gained through applying open science principles is valuable and should be the standard. Adoption of open science principles/tools to support IEAs across the US has dramatically increased the IEA practitioners' efficiency as the need to produce informative, up-to-date, and usable products can be met faster by the application of these principles. The rapid product turnaround and the highly collaborative nature of IEAs make open science extremely useful in advancing the implementation of EBM.

References

Baumer, Ben, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J Horton. 2014. R markdown: Integrating a reproducible analysis tool into introductory statistics. *arXiv preprint arXiv:1402.1894*.

Chacon, Scott, and Ben Straub. 2014. *Pro git.* Apress. https://doi.org/10.1007/978-1-4842-0076-6.

Chamberlain, Scott. 2019. *Rerddap: General purpose client for 'erddap' servers.*

Christensen, Villy, and Carl J Walters. 2004. Ecopath with ecosim: Methods, capabilities and limitations. *Ecological modelling* 172. Elsevier: 109–139.

Hampton, Stephanie E, Sean S Anderson, Sarah C Bagby, Corinna Gries, Xueying Han, Edmund M Hart, Matthew B Jones, et al. 2015. The tao of open science for ecology. *Ecosphere* 6. Wiley Online Library: 1–13.

Hardison, Sean. 2018. *Ecodata: Documentation of ecosystem indicator reporting.*

Harvey, Chris J., Christopher R. Kelble, and Franklin B. Schwing. 2016. Implementing "the IEA": using integrated ecosystem assessment frameworks, programs, and applications in support of operationalizing ecosystem-based management. *ICES Journal of Marine Science* 74: 398–405. https://doi.org/10.1093/icesjms/fsw201.

Hazen, Elliot. 2018. EcoCast. *GitHub repository.* https://github.com/elhazen/EcoCast; GitHub.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. 2016. Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, 87–90.

Lebo, Timothy, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. Prov-o: The prov ontology. *W3C recommendation.* w3c.

Levin, Phillip, Michael Fogarty, Gary Matlock, and Marjorie Ernst. 2008. *Integrated Ecosystem Assessments.* NOOA NMFS-NWFSC-92.

Levin, Phillip S, Michael J Fogarty, Steven A Murawski, and David Fluharty. 2009. Integrated ecosystem assessments: Developing the scientific basis for ecosystem-based management of the ocean. *PLOS Biology* 7. Public Library of Science: 1–6. https://doi.org/10.1371/journal.pbio.1000014.

Lewison, Rebecca, Alistair J. Hobday, Sara Maxwell, Elliott Hazen, Jason R. Hartog, Daniel C. Dunn, Dana Briscoe, et al. 2015. Dynamic ocean management: Identifying the critical ingredients of dynamic approaches to ocean

resource management. *BioScience* 65: 486–498. https://doi.org/10.1093/biosci/biv018.

Lowndes, Julia S Stewart, Benjamin D Best, Courtney Scarborough, Jamie C Afflerbach, Melanie R Frazier, Casey C O'Hara, Ning Jiang, and Benjamin S Halpern. 2017. Our path to better science in less time using open data science tools. *Nature ecology & evolution* 1. Nature Publishing Group: 0160.

Lowndes, Julia S Stewart, Erich J Pacheco, Benjamin D Best, Courtney Scarborough, Catherine Longo, Steven K Katona, and Benjamin S Halpern. 2015. Best practices for assessing ocean health in multiple contexts using tailorable frameworks. *PeerJ* 3. PeerJ Inc.: e1503.

Ma, Xiagang, Peter Fox, Curt Tilmes, Katherine Jacobs, and Anne Waple. 2014. Capturing provenance of global change information. *Nature Climate Change* 4. Nature Publishing Group: 409–413.

Ma, Xiaogang, Stace E Beaulieu, Linyun Fu, Peter Fox, Massimo Di Stefano, and Patrick West. 2018. Documenting provenance for reproducible marine ecosystem assessment in open science. In *Information retrieval and management: Concepts, methodologies, tools, and applications*, 1051–1077. IGI Global.

McKiernan, Erin C, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, et al. 2016. Point of view: How open science helps researchers succeed. *Elife* 5. eLife Sciences Publications Limited: e16800.

Michener, William K, and Matthew B Jones. 2012. Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in ecology & evolution* 27. Elsevier: 85–93.

Müller, Kirill. 2017. *Here: A simpler way to find your files.*

NOAA-EDAB. 2019. Technical documentation, state of the ecosystem. *GitHub repository.* https://github.com/NOAA-EDAB/tech-doc; GitHub.

Ram, Karthik. 2013. Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine* 8. https://doi.org/10.1186/1751-0473-8-7.

Samhouri, Jameal F., Alison J. Haupt, Phillip S. Levin, Jason S. Link, and Rebecca Shuford. 2013. Lessons learned from developing integrated ecosystem assessments to inform marine ecosystem-based management in the USA. *ICES Journal of Marine Science* 71: 1205–1215. https://doi.org/10.1093/icesjms/fst141.

Samhouri, Jameal F., Errin Ramanujam, Joseph J. Bizzarro, Hayley Carter, Kelly Sayce, and Sara Shen. 2019. An ecosystem-based risk assessment for california fisheries co-developed by scientists, managers, and stakeholders. *Biological Conservation* 231: 103–121. https://doi.org/https://doi.org/10.1016/j.biocon.2018.12.027.

Sievert, Carson. 2019. *Plotly for r.*

Simons, R. A. 2019. ERDDAP. Monterey, CA: NOAA/NMFS/SWFSC/ERD.

Tukey, John W. 1977. *Exploratory data analysis.* Addison-Wesley.

Weijerman, Mariska, Jamison M. Gove, Ivor D. Williams, William J. Walsh, Dwayne Minton, and Jeffrey J. Polovina. 2018. Evaluating management strategies to optimise coral reef ecosystem services. *Journal of Applied Ecology* 55: 1823–1833. https://doi.org/10.1111/1365-2664.13105.

Welch, Heather. 2018. *GitHub repository.* https://github.com/HeatherWelch/EcoCast_Operationalization; GitHub.

Wickham, Hadley. 2016. *Ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York.

Wickham, Hadley, and Garrett Grolemund. 2016. *R for data science: Import, tidy, transform, visualize, and model data.* " O'Reilly Media, Inc.".

Wilkinson, Leland. 2012. The Grammar of Graphics. Edited by James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori. Springer Handbooks of Computational Statistics. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-21551-3_13.

Wilkinson, Michel AND Aalbersberg, Mark D. AND Dumontier. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3: 160018 EP.

Wilson, Greg. 2006. Software carpentry: Getting scientists to write better code by making them more productive. *Computing in Science & Engineering* 8. IEEE: 66–69.

Yenni, Erica M. AND Bledsoe, Glenda M. AND Christensen. 2019. Developing a modern data workflow for regularly updated data. *PLOS Biology* 17. Public Library of Science: 1–12. https://doi.org/10.1371/journal.pbio.3000125.

Zador, Stephani G, Kirstin K Holsman, Kerim Y Aydin, and Sarah K Gaichas. 2017. Ecosystem considerations in alaska: The value of qualitative assessments. *ICES Journal of Marine Science* 74. Oxford University Press: 421–430.