DATA MANAGEMENT PLAN

We are aware that the NSF Policy on Dissemination and Sharing of Research requires prompt publication of results, as well as sharing of supporting materials, with other researchers. We will comply enthusiastically and comprehensively. In addition to products to be published in academic journals, the other primary supporting materials of the proposed research consist of (1) data products, (2) additions to open-source software packages that implement new statistical methods, and (3) computer code that lead to manuscripts. Our Data Management Plan is a strategy for archiving and disseminating data, metadata, and code in ways that promote complete reproducibility and transparency, provide wide and easy access to new data and approaches, and facilitate continuing research by ourselves and other researchers. Code will be written primarily in the *R* and MATLAB languages. Manuscripts will be written in LaTeX. We have a track record of open and reproducible science, including publishing data[115,118,172,173,231–236], code[41,221,237–240], open-source software packages[121,208,209,241,221], and open-source repositories[242–244] that reproduce complete analyses leading to a published paper at the click of a button, even for works still undergoing peer review.

(1) Data. ***Kelp biomass and community datasets.*** Our project leverages a 36-year ongoing time series of giant kelp canopy biomass determined from Landsat satellites sensors. These data will be used in Tasks I, II, and III. The dataset was developed by PIs Bell and Cavanaugh, and uses linear spectral unmixing to determine the fractional cover of giant kelp canopy in each Landsat pixel (30 m × 30 m), every 1–2 months, from 1984 to present, across the range of giant kelp dominance in California, USA, and Baja California, Mexico. These estimates of fractional cover are then converted to canopy biomass using an empirical relationship developed with ancillary data from the SBC LTER. These datasets are publicly available on the Environmental Data Initiative (EDI) Data Portal[118]. Data updates produced during the course of the proposed research will be added to the existing public dataset.

For Task I, we will construct a harmonized spatial time series of sea urchin density from several publicly available datasets covering 91 sites across California from 1981 to present[176–185]. This integrated sea urchin dataset will be made available on the EDI Data Portal by the end of the project. For Task II, we will use reef community data that includes the cover and biomass of over 50 understory algal species and will be compiled from two existing datasets. These data include monitoring efforts from the San Onofre Nuclear Generating Station Mitigation Monitoring Program[181–185] (see letter from D. Reed) and community surveys from the SBC LTER, which are already posted on the EDI Data Portal[180]. For Task III, kelp wrack and sandy beach community data include monthly time series of wrack cover and wet biomass at five beaches spanning ~50 km of coastline. These data also include monthly counts of shorebirds and annual surveys of macroinvertebrates. The data are available on the EDI Data Portal[186–188].

Environmental and oceanographic datasets. Spatial time series of sea surface temperature used in Tasks I, II, and III are available from the Scripps Photobiology Group. Seawater nitrate concentration will be derived from empirical temperature-to-nitrate relationships[164]. Modeled significant wave height data are available from the Coastal Data and Information Program[166–168] and NOAA Wave Watch III[192] online dataset repositories. Location-specific time series of temperature, nitrate concentration, and significant wave height will be made available on the EDI Data Portal by the end of the project. Dispersal of giant kelp spores in Task I and giant kelp wrack in Task III will be determined using a publicly available dataset summarizing Lagrangian particle simulations from ocean circulation models (Regional Oceanic Modeling System, ROMS)[195–199]. A spatial time series of minimum transport times between each location through time will be made available on the EDI Data Portal by the end of the project, as we have done in past works on giant kelp dispersal[173]. Several oceanographic climate indices, such as the North Pacific Gyre Oscillation, Multivariate El Niño Southern Oscillation Index, and the North Pacific Gyre Oscillation Index, will be accessed from publicly available sources[140,141,171].

Metadata. For each dataset we will develop and publish metadata files that describe the personnel involved in each step of data collection; the methods, location, and time of data collection; and descriptions about how the data are organized. We will conform to the Ecological Metadata Language (EML) standards and approaches because of their widespread usage and ease of access.

Data re-use policies. No permission restrictions will be placed on the data. We anticipate that students, academic scientists, conservation practitioners, and managers will be the typical data users.

(2) Git and reproducible workflows. Our strategy for the management of the data analysis process provides complete transparency and reproducibility of all analyses. As is increasingly common in modern quantitative ecology, we use the Git version control system to track the complete revision history of all analyses, and make the repository for each project public upon submission of that work for publication (or before). For instance, refs. 242–244 contain the repositories for a few of our recently published or in-press works. Git is used widely by software engineers and more recently by quantitative ecologists. Git stores complete revision histories so information cannot be lost, and effectively manages revision conflicts from multiple users. Data for all analyses will be automatically imported by analysis scripts from the repositories mentioned in *Data*, above. These analysis scripts, as well as the manuscript writing process, will be archived in the same repository. The end result is an open-access repository from which the entire data analysis process and compilation of the written paper can be quickly reproduced.

(3) *R* packages. To promote the easy and wide adoption our new methods by other researchers, we will use our established practice of developing *R* packages that contain unit-tested and well-documented code implementing our new methods. Packages are initially held on Github[209] and are published to the Comprehensive *R* Archive Network (CRAN)[121,208] after a sufficient period of testing. Packages include standard documentation as well as explanatory “vignettes” that serve as manuals to guide readers on how to use the package[121]. When appropriate, we have written an accompanying methods paper[119,241]. We will apply these practices for the dissemination of reusable codes and methods in the proposed research. New or improved methods will be incorporated into our existing *wsyn*[121], *tsvr*[208], or *mms*[209] packages. In our experience, our open-science practices are extremely important for facilitating the uptake of our statistical methods by other researchers from a range of backgrounds. This component of our workflow is distinct from activities described above in *Git and reproducible workflows* because packages are created for widespread use in a variety of different scientific applications.

(4) Linking of data, code, and results. In summary, as a single project develops toward completion, we will create four interacting and interlinked online resources that support the project and also provide independent value. (1) Data on which analyses are based will be permanently archived (with DOI), freely-accessible, and containing descriptive metadata. (2) Reusable code and any new methods will be put into an *R* package hosted on Github (a cloud-hosted Git repository), and eventually on CRAN. (3) A project repository, which imports our archived data and makes use of our package(s), will store the complete revision history of the analysis, and can reproduce all analyses and recompile the paper at the touch of a button. (4) The published paper, which includes links to all items 1–3, above. In fact, these items are as much part of the finished product as is the published paper. As an example, our recent in-press paper constructing a timescale-specific variance ratio[223] comprises not only the paper itself, but also (1) the archived data[236]; (2) the *R* package *tsvr*[208], which contains implementations of new methods developed for the manuscript and an associated vignette, published on CRAN prior to manuscript submission; and (3) a project repository[242] that imports the data and the package, and performs all analyses used in the paper.

(5) Code storage and quality, release of information. We will use a cloud-hosted Git repository (Github) for the duration of the project. Repositories will be made public upon submission of papers for peer review. Cloud storage requirements will not exceed what Github provides free to academic users. Reuman will oversee unit testing of *R* code, with assistance from Walter, Sheppard, and Castorani. Unit tests will be included in *R* packages. The project repositories will be maintained indefinitely by Reuman and Castorani (very little maintenance will be required once the project concludes). All repositories will be public and publications will link to them. *R* packages will be maintained on Github and CRAN. All code will be released under Version 3 of GNU Affero General Public License as developed by the Free Software Foundation, allowing for the source code to be freely distributed and modified.