Ecological Applications, 31(4), 2021, e02304

Improving the ability of a BACI design to detect impacts within a kelp-forest community

Andrew Rassweiler ,[1,6] Daniel K. Okamoto,[1] Daniel C. Reed,[2] David J. Kushner,[3] Donna M. Schroeder,[4] and Kevin D. Lafferty[5]

[1]*Department of Biological Science, Florida State University, Tallahassee, Florida 32306 USA*
[2]*Marine Science Institute, University of California, Santa Barbara, California 93106 USA*
[3]*Channel Islands National Park, Ventura, California 93001 USA*
[4]*Bureau of Ocean Energy Management, Pacific OCS Region, 760 Paseo Camarillo, Camarillo, California 93010 USA*
[5]*U.S. Geological Survey, Western Ecological Research Center, Marine Science Institute, University of California, Santa Barbara, California 93106 USA*

Citation: Rassweiler, A., D. K. Okamoto, D. C. Reed, D. J. Kushner, D. M. Schroeder, and K. D. Lafferty. 2021. Improving the ability of a BACI design to detect impacts within a kelp-forest community. Ecological Applications 31(4):e02304. 10.1002/eap.2304

Abstract. Distinguishing between human impacts and natural variation in abundance remains difficult because most species exhibit complex patterns of variation in space and time. When ecological monitoring data are available, a before-after-control-impact (BACI) analysis can control natural spatial and temporal variation to better identify an impact and estimate its magnitude. However, populations with limited distributions and confounding spatial-temporal dynamics can violate core assumptions of BACI-type designs. In this study, we assessed how such properties affect the potential to identify impacts. Specifically, we quantified the conditions under which BACI analyses correctly (or incorrectly) identified simulated anthropogenic impacts in a spatially and temporally replicated data set of fish, macroalgal, and invertebrate species found on nearshore subtidal reefs in southern California, USA. We found BACI failed to assess very localized impacts, and had low power but high precision when assessing region-wide impacts. Power was highest for severe impacts of moderate spatial scale, and impacts were most easily detected in species with stable, widely distributed populations. Serial autocorrelation in the data greatly inflated false impact detection rates, and could be partly controlled for statistically, while spatial synchrony in dynamics had no consistent effect on power or false detection rates. Unfortunately, species that offer high power to detect real impacts were also more likely to detect impacts where none had occurred. However, considering power and false detection rates together can identify promising indicator species, and collectively analyzing data for similar species improved the net ability to assess impacts. These insights set expectations for the sizes and severities of impacts that BACI analyses can detect in real systems, point to the importance of serial autocorrelation (but not of spatial synchrony), and indicate how to choose the species, and groups of species, that can best identify impacts.

Key words: before-after-control-impact (BACI) analysis; impact detection; informedness; long-term monitoring; serial autocorrelation.

Introduction

After the T/V Exxon Valdez spilled oil over 2,000 km of wild Alaskan coastline in 1989, the parent corporation argued to reduce their liability to a fraction of the billions of dollars in damages originally assessed. Their arguments succeeded, in part, because the lack of baseline data in Prince William Sound, Alaska, made it difficult for scientists to assess natural resource injuries stemming from the spill with enough certainty to assign culpability (Jewett et al. 1999, Skalski et al. 2001, Fukuyama et al. 2014). In response to the need for more rigorous impact assessments, many government agencies have initiated long-term monitoring programs (Field et al. 2007, Fancy et al. 2009). Using such monitoring data, a before-after-control-impact (BACI) statistical design (Stewart-Oaten et al. 1986, Underwood 1992, 1994), comparing affected sites with unaffected sites before and after an event, can help test for impacts such as occurred after the 2010 Deepwater Horizon spill (Dietl and Durham 2016, Lauritsen et al. 2017). Recent simulation studies have reemphasized the advantages of BACI over alternative approaches such as before-after, control-impact and randomized control trials (Christie et al. 2019). However, the classic BACI analysis makes several assumptions that are often violated by spatially replicated time-series data, and the unbalanced

Manuscript received 12 June 2020; revised 4 September 2020; accepted 27 October 2020. Corresponding Editor: Brice X. Semmens.
[6] E-mail: rassweiler@bio.fsu.edu

statistical designs resulting from unexpected impacts can have low replication and poor power to discriminate between human impacts and natural confounding variation. The degree to which these violated assumptions compromise BACI approaches can be hard to assess from simulation studies. Here, we use real-world data to assess how well a BACI design can detect various hypothetical impacts, and identify ways to improve its performance.

By definition, a BACI design compares control (i.e., non-impacted) and impacted sites. When it is not clear where an impact might occur, monitoring sites cannot be planned to sample both impacted and non-impacted areas. In some cases, as in Prince William Sound, an event might fail to impact any monitoring sites, at which point, control and impact sites can be surveyed only after the impact occurs (e.g., a control-impact study). The risk that an impact occurs at a site where no monitoring data are being collected can only be reduced by adding sites or placing them more strategically. Furthermore, power to detect an impact is, all things being equal, highest when one-half of the sites are impacted (Shaw and Mitchell-Olds 1993). Therefore, monitoring programs usually aim for dispersing sites broadly, choosing areas that differ in their exposure to impacts, or choosing sites that managers care about most. Regardless, the number and distribution of monitoring sites will affect the utility of a BACI design and the extent to which it is balanced.

The other main limitation to BACI designs relates to natural temporal and spatial variation in species abundances. Random variation reduces statistical power, whereas nonrandom variation can either increase the false impact detection rate by creating confounding impacts or decrease power by altering the signal to noise ratio. In many cases, nonrandom variation occurs due to natural environmental drivers. For instance, substrate type, ocean temperature and wave action each drive substantial spatial and temporal variation in species abundances in benthic marine ecosystems (Caselle et al. 2015, Castorani et al. 2018, Miller et al. 2018). Although this variation may in some cases be reduced through the use of paired BACI designs (Stewart-Oaten et al. 1986), in most cases, it must be accounted for through the use of covariates. Including such covariates when testing for diverging population trajectories at impacted sites can partially account for serial autocorrelation and spatial synchrony caused by environmental drivers (Rost et al. 2012, Martínez-Abraín et al. 2013, Lamy et al. 2019). Consequently, false impact detections due to temporal and spatial autocorrelation may be reduced by incorporating underlying environmental drivers as covariates (Parker and Wiens 2005, Kalies et al. 2010). When environmental drivers are not known, random effects for year or site can account for site-specific environmental effects. However, increasing model complexity, whether by adding covariates or random effects, can also reduce power and inflate uncertainty in estimated effects if models are overspecified (i.e., overfitting) or variables/random effects are confounded (i.e., strong collinearity).

Large-scale drivers of kelp forest systems such as sea surface temperature and the North Pacific Gyre Oscillation affect regions rather than individual sites (Reed et al. 2011, Bell et al. 2015, Lamy et al. 2018, Okamoto et al. 2020), and may cause species at nearby sites to fluctuate synchronously rather than independently. For instance, populations may collectively decline due to regional collapses in productivity (Okamoto et al. 2012, 2016), a host might decline due to widespread disease (Harvell et al. 2019), or the relative abundance of cold and warm water species might respond to climate trends (Holbrook et al. 1997). When the drivers of such temporal synchrony are not known (or have not been measured), it can confound before-after comparisons. Most monitoring programs do not have the time series needed to make before-after comparisons, let alone account for temporal correlation structure (Ramsey and Schafer 2002). Even when long time series are available, it can be difficult to disentangle impacts and recovery from natural trends. Here, we consider the extent that it would be possible to control for serial autocorrelation using autoregressive models with reasonably long-term ecological data (Stewart-Oaten 2003).

The ability to account for autocorrelation depends on how the spatial scale of those correlations relates to the spatial scale of an impact. Some environmental covariates vary from site to site, whereas others affect a few neighboring sites at a time. For instance, storms, disease outbreaks, and large temperature anomalies can affect a whole region, or just a small stretch of coastline. Such spatially concentrated effects lead to patchy temporal changes that might appear to be anthropogenic. For instance, if a warm-water event affects species in one part of the monitored region more than another, then this natural change might be misinterpreted as a human impact. Because such natural changes might coincide with hypothesized impacts, we might expect such spatial synchrony would increase the false detection rate, a possibility we explore here.

In environmental assessments, BACI designs often compare abundances in time and space, focusing on one species at a time. However, different species may vary dramatically in their utility for detecting impacts, even if they are similarly vulnerable to disruption by human activities (Roberge and Angelstam 2006, Meyer et al. 2010). Species differ in their distributions, their characteristic temporal variability and sensitivity to environmental drivers, and the degree to which regional populations vary synchronously. An ideal species for impact detection would be widespread with a strong sensitivity to impacts relative to its year-to-year variation under natural conditions. Unfortunately, many species may not be ideal for BACI analyses. Therefore, we evaluated how species varied in their ability to indicate impacts, and how analyzing populations of multiple

species simultaneously might improve BACI performance while reducing false impact detections.

We investigated statistical power and false impact detection for a BACI design when applied to monitoring data for reef fishes, invertebrates, and macroalgae collected by the Channel Islands National Park (CINP) and the U.S. Geological Survey (USGS). The CINP has been monitoring 16 sites at five islands (San Miguel, Santa Rosa, Santa Cruz, Anacapa, and Santa Barbara Islands) off California annually for more than 30 yr (Fig. 1). The USGS has annually monitored seven sites at nearby San Nicolas Island (SNI) over a similar time. Although there are some differences between the programs' methods, both programs count mobile invertebrates, fish, and kelp within permanently marked areas, and sample the percent cover of sessile invertebrates and smaller macroalgae using point contact methods (full details on sampling methods are given in Kenner et al. [2013] and Kushner et al. [2013]). In this marine system, natural variability occurs in space and time (Reed et al. 2016, Lamy et al. 2019). On shallow reefs in southern California, for example, ecological communities can change abruptly (Ebeling et al. 1985, Harrold and Reed 1985, Rassweiler et al. 2010), and are influenced by environmental gradients combined with year-to-year variation in recruitment or temporally correlated environmental forcing (e.g., the El Nino Southern Oscillation or Pacific Decadal Oscillation; Kenner and Tinker 2019). BACI has been used successfully with these data sets, even for single species (Schroeter et al. 2001, Parnell et al. 2005), suggesting these data are suited for evaluating the strengths and weaknesses of this statistical approach.

We applied hypothetical simulated impacts that differed in severity (e.g., percent reduced abundance) and spatial scale (impacts varied in their radius). Importantly, although the impacts were simulated, the underlying data were otherwise real, and thus introduced all the complications present in real-world impact analyses. After controlling for serial autocorrelation in the data, we asked how rarity, spatial synchrony, and variability in time affected the power to detect impacts and false impact detection rates (type I error). We also asked how each affected "informedness" (Youden 1950, Powers 2011), a metric that simultaneously measures capacity to accurately detect real impacts and avoid false impact detections. In addition to analyzing individual species, we explored whether analyzing species groups improved or diminished informedness. These analyses set expectations for the utility of BACI analyses when applied to marine monitoring data, both in this system and in similar ecological contexts, and suggest best practices for design and implementation of such analyses.

Methods

Data used

We considered 28 common species (5 macroalgae, 10 fish, 7 mobile invertebrates, and 6 sessile invertebrates) that were sampled in compatible ways for the whole length of both monitoring programs. We used data from the 16 CINP sites that have been monitored annually since 1985 and the 7 SNI stations monitored over a similar timeframe (all data used here are published in Kenner et al. 2013, Kushner et al. 2013). CINP samples once per year in summer to early autumn, while the SNI sites are sampled twice per year, once in late summer to early autumn and once in spring. We excluded the spring samples from the SNI data to maintain compatibility in temporal resolution between the two data sets. For each site and sampling year, we aggregated the data for each species to produce a measure of abundance based on total count (individuals or points observed) along with the total area or number of points sampled.

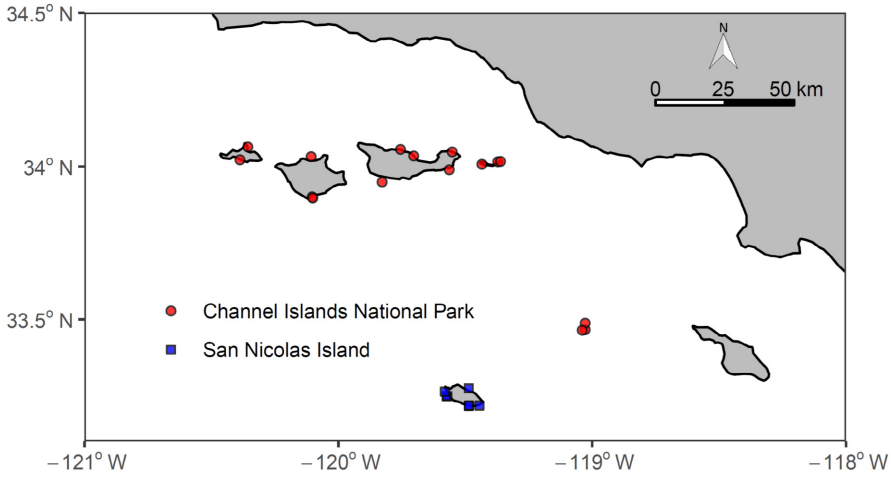

Fig. 1. Study sites in the Southern California Bight, USA. Shapes indicate the programs, designated as either San Nicolas Island or the Channel Islands National Park.

Simulating human impacts

We simulated a hypothetical human impact by reducing a species' abundance over a period of time at "impacted" sites, and then analyzed this altered data set to determine whether the reduction in abundance was statistically detectable and how the likelihood of detection varied for each species and across different spatial scales and severities of impact. We focused here on impacts that reduce the abundance of one or more species. We acknowledge that impacts often have cascading consequences on non-impacted species, occurring through trophic or other interactions. However, such interactions are likely to be complex in real food webs, and difficult to predict without a full model of community interactions that includes its own assumptions about system dynamics.

We simulated circular impacts of varying radii, centered at random between a latitude of 33.2° and 34.1° N and a longitude between 119.0° and 120.4° W, a domain that included all sites in the data set. The location and radius determined which monitoring sites were impacted. We then selected an impact year at random between 1996 and 2006. The 10 yrs before the selected impact year were used as pre-impact data and the 5 yrs after were used as our post-impact period. By doing so, we manipulated the spatial balance between the number of impact and control sites, but fixed the temporal balance between the number of pre- and post-impact years.

For each simulated impact, we constructed a new data set in which the actual observed abundances of each species were retained for all sites in the pre-impact period and the control sites during the post-impact period, but data from the impacted sites in the "post" period were replaced with proportionally reduced abundances. The degree to which impacted abundances were reduced was based on impact *severity*, which was the average fraction by which the abundance of each species within the affected area was reduced (so a severity of 0 meant no effect was simulated; i.e., zero change in abundance at all sites), while a severity of 100% meant complete absence of species at all the impact sites. Within a given simulation, we assumed all monitoring sites within the chosen radius of the center of the impact were similarly reduced over the post-impact period (so effects were spatially uniform and persisted over time). Because the data represent counts (either of whole organisms or of points where organisms occur), simulated severity was applied probabilistically by rounding to the neighboring integers with probability reflecting the remaining decimal (e.g., 10.25 was rounded to ten 75% of the time and to eleven 25% of the time). This probabilistic rounding was done to maintain the integer nature of the data while maintaining overall mean impact severity. For each impact level and each species/functional group, we ran 1,000 simulations and summarized statistics across runs. A bootstrap resampling of our simulations with replacement indicated that 1,000 simulations was sufficient to reduce error and achieve convergence (Appendix S1: Fig. S1). To simplify interpretation, we assumed species impacts were independent (e.g., a reduction in a predator did not result in an increase in its prey).

Statistical models

We used generalized linear mixed effects models for all analysis with a before/after (BA) effect, a control/impact (CI) effect, and an interaction between BA and CI terms representing the BACI effect. This model structure hypothesized a step-change in the abundance of the evaluated species at the impacted sites, and was consistent with the actual structure of our simulated impacts. Alternative forms of BACI, for example those assuming the impact causes a change in population trend over time (Thiault et al. 2017), were not evaluated here.

We used a Poisson likelihood for count data and binomial likelihood for point contact data. We attempted to keep the models simple to minimize convergence issues and model fitting errors in the simulation algorithm. We avoided both the negative binomial and beta-binomial, which also account for additional dispersion, because these implementations converged less reliably over many randomization runs. Instead, to account for both over dispersion and zero-inflation, we included a random log-normal error dispersion term in the model. We also excluded sites from our analysis if the focal species was present at fewer than 15% of sampling events, to reduce zero-inflation and constrain data sets for reasonable comparisons. Rather than add site-level covariates (not all of which were known), to account for among-site variation, we added a random effect of site in the single-species models, and site within species for the multi-species models. We used this random-effects structure because the spatial distribution of abundances among sites is expected to vary by species. Serial autocorrelation was calculated as the within-site empirical partial autocorrelation function, averaged across all sites. To account for serial autocorrelation in our analyses, we used a first-order autoregressive model (AR(1)) on the within-site error terms. Inclusion of the AR(1) model occasionally yielded identifiability challenges, particularly for species that were only intermittently present at sites, and thus we also report scenarios where models did not converge. We estimated statistical models using glmmTMB (Magnusson et al. 2016) in R (R Core Team 2017) with the default algorithms for generality and because they provided flexibility and performance over alternatives. Code is available in a Git repository archived on Zenodo (Okamoto 2020).

Performance metrics

Performance metrics for each run included the following: (1) one-sided significance tests (negative and positive) for impact (from which we calculated power or false impact detection rates, depending on whether an

impact was simulated), (2) whether or not the model converged (indicating when a model failed despite having sufficient data for analysis), (3) data sufficiency, in this case requiring at least two sites in each of the control and impact categories with at least 15% of the years with the species present in the original data (this threshold was chosen for efficiency as data below this threshold consistently provided unreliable model estimates), (4) relative error from the true effect measured as (estimated impact – true impact)/(true impact) (this was only calculated when we simulated a non-zero impact) where the among-simulation mean provides an estimate of bias, and (5) the Youden's J informedness statistic, which measures the probability of a correct inference, assuming an even prior probability of no impact vs. a true impact, thereby assessing both false negative and positive errors (Youden 1950, Powers 2011). Informedness is calculated as the sensitivity (i.e., the probability that a true impact is detected) plus the specificity (the probability that no significant effect is detected in cases where there was no impact) minus one.

Additional statistics

To explore the impact of species rarity, temporal variation and spatial synchrony, we calculated various properties for each species and assessed how they affected model performance. To describe rarity, we considered two related metrics: (1) fraction of simulations in which there was insufficient data to conduct the analysis (i.e., the species was absent from control sites, absent from impacted sites, absent before, or absent after), and (2) fraction of sites in which the species was present. For temporal variation, we calculated the coefficient of variation as the within-site standard deviation divided by the within-site mean, averaged across all sites for the 10 yrs before the impact. Spatial synchrony among sites was estimated as the variance of the mean time series divided by the sum of the covariance matrix of the group of individual time series (Loreau and de Mazancourt 2008). We compared whether these metrics could explain power and false impact detection rates using generalized linear models with a beta-binomial likelihood in a multiple regression framework. To make coefficients comparable, we centered and standardized all variables by subtracting the mean and dividing by the standard deviation. Predictor variables that had only positive values (i.e., coefficient of variation, fraction of sites with sufficient data, and synchrony; all must be >0) were log-transformed.

Species combinations

Because averaging among species can help account for balance and natural temporal and spatial heterogeneity, we assessed power and false impact detection rates for different species combinations. Combinations included all species of the same functional group (macroalgae, fish, mobile invertebrates, sessile invertebrates), and a series of species subsets generated by sequentially adding species ranked in informedness from highest to lowest and lowest to highest (based on single-species analysis). For the latter analysis, we combined only species with counts (not point contact data) in order to maintain a consistent response variable and likelihood structure.

Results

Impact severity and area

As expected, the likelihood of detecting an impact with a BACI design (i.e., statistical power) increased with impact severity. This is illustrated by the top panels of Fig. 2, which are representative of the pattern across species. Furthermore, power had a hump-shaped relationship with the area impacted, as illustrated for the black surfperch, *Embiotoca jacksoni* and the giant kelp, *Macrocystis pyrifera* (Fig. 2A and B). This reflects the probability of obtaining a balanced statistical design with a similar number of impacted and control sites. Statistical balance was most likely for impacts about 65 km in radius (~13,000 km^2); very small-scale impacts were unlikely to impact monitoring sites, whereas large-scale impacts were unlikely to exclude many control sites. The probability of having occupied control and impact sites at a particular scale varied based on the distribution of each species, but only slightly (Fig. 3). Although the general shape of this pattern and location of the peak in statistical power was consistent across most species analyzed, the peak's height varied from species to species, as discussed in more detail below. Although statistical power peaked at intermediate impact areas, precision in estimates of impact severity increased monotonically with impact area, with relative error in estimates showing a consistent decline (and no change in bias) as area increased (Fig. 2C and D). Estimates of impact severity were not consistently biased, with about half the species' estimates above and half below the true severity for an 80% simulated reduction (Appendix S1: Fig. S2). Importantly, decreases in impact severity resulted in increases in among-run variation (i.e., decreased statistical efficiency) and bias in the estimate (Appendix S1: Fig. S3), although the degree of bias and variation differed by species. In short, BACI failed to assess small-scale impacts, and had low power, but high precision when assessing large-scale impacts.

Variation in power and false impact detections

False impact detections were higher than the target false positive rate of 5%, power was generally low, and both varied substantially among species and impact scenarios. In simulations where one-quarter of sites were impacted (five or six sites), both power and false impact detection rates differed among species (Fig. 4). Additionally, we found that false impact detection rates were

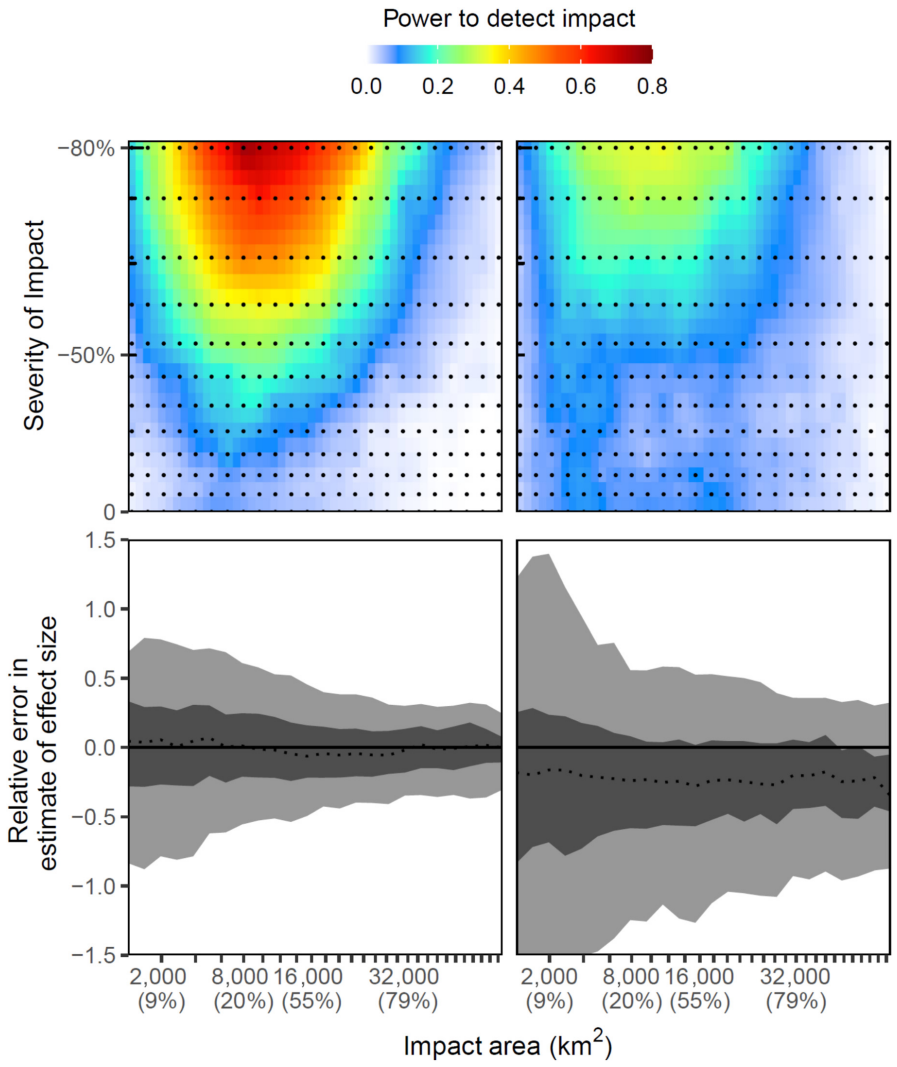

FIG. 2. Top panels illustrate detection probability as a function of the spatial scale and severity of an impact for *Embiotoca jacksoni* (black surfperch, left) and *Macrocystis pyrifera* (giant kelp, right). Bottom panels show relative error in the estimate of effect size as a function of spatial scale of impact (shown for an impact severity of 80%). Bands are the 90% quantiles (light gray) interquartile range (dark gray), and median error (dotted black line). Error is standardized as the log-estimate minus the log-true value divided by the log-true value. Percentages in parentheses indicate percent of sites typically impacted by impacts of that area.

often more common than expected based on the nominal alpha level of 0.05. In scenarios with no simulated impact (zero severity, right panels of Fig. 4), only 11 out of 28 species had <5% false impact detection while 23 out of 28 had a <10% false impact detection. Elevated false impact detection rates were attributable (at least in part) to the spatial clustering of impacts. If impact sites were scattered across the region instead, as might occur with a change in coastal zoning such as implementation of a network of marine protected areas, false impact detections were less common (Appendix S1: Fig. S4).

If serial correlation had not been accounted for, as we did with an AR(1) model, false impact detections would have been much more frequent (Fig. 5), particularly for species with positive year-to-year correlations (i.e., higher levels of correlation). This effect was stronger for scenarios where impacted sites were spatially aggregated than when they were isolated (i.e., impacts were scattered across sites). Although helpful, a first-order autoregressive model did not always eliminate serial autocorrelation, probably due to higher order correlations, variance among sites in the degree of autocorrelation, or heteroskedastic error variances.

Power to detect even strong impacts was variable among species in part because some species were present at few sites in the region or exhibited enormous year-to-year variation in abundance. Results of post-hoc regressions on statistical power and false impact detections showed that power was lower for rare species (those present at few sites) and was also lower for species with

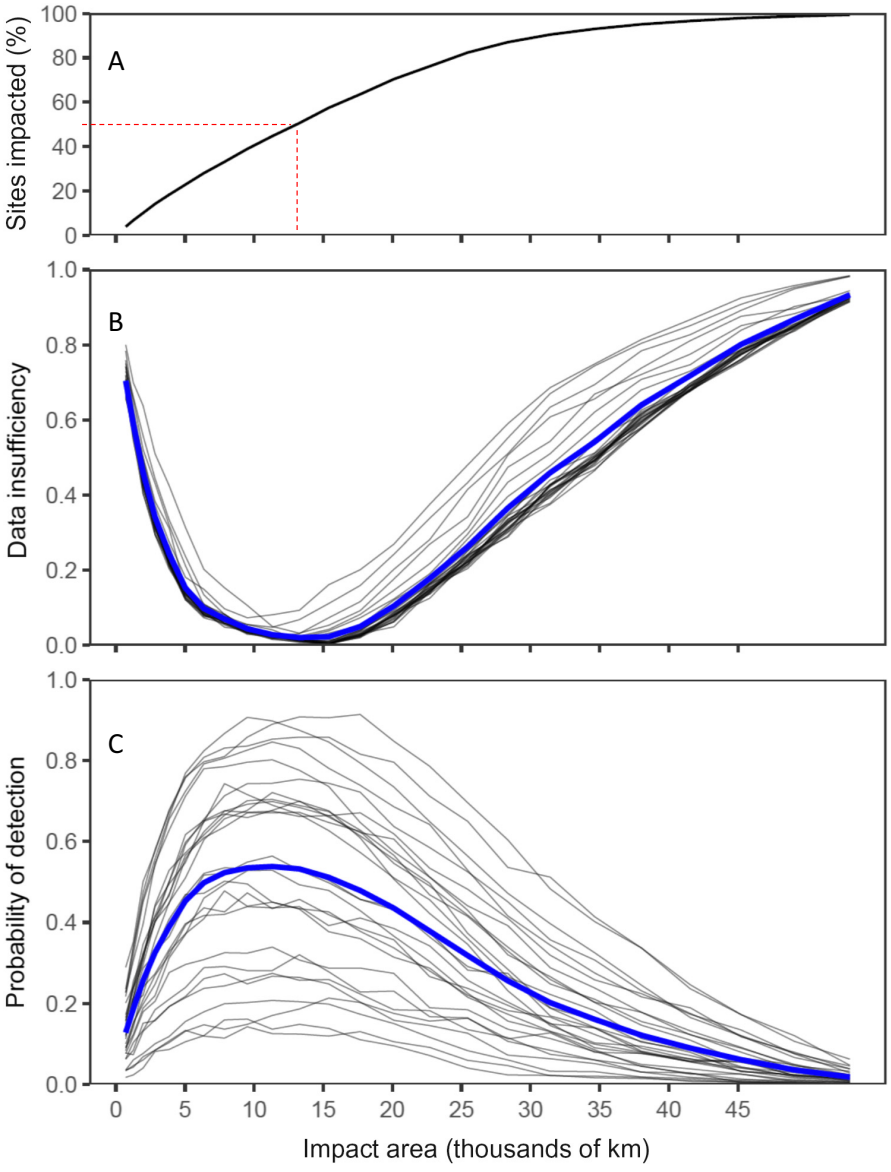

FIG. 3. Effect of impact area on (A) mean percentage of sites impacted (the red dashed line indicates the point at which statistical balance is most likely), (B) probability of data insufficiency for analysis for each species (thin lines) and the mean probability (thick blue line), and (C) detection probability for before-after-control-impact (BACI) estimates for each species (thin gray lines) and the mean probability (blue line). Results are shown for an 80% impact. Note that closely spaced grey lines appear as a heavier black line in panels A and B.

higher temporal variability and higher serial correlation (even after applying an AR(1) error model). However, spatial synchrony did not have the predicted effect of increasing power or false negative detection rates (Table 1, Fig. 6). Unfortunately, the same properties that provided high statistical power also increased false impact detections (Figs. 6, 7). However, some species yielded low false impact detection with only modest statistical power (e.g., the giant kelp, *Macrocystis pyrifera*), whereas others exhibited similarly low false impact detection, with high statistical power (e.g., the black surf perch, *Embiotoca jacksoni*).

We calculated informedness of each species-specific analysis, which is the likelihood of detecting a true impact in a species discounted by the rate at which false impact detections occur. Species varied across a wide range of informedness from near 0 (the sandcastle worm, *Phragmatopoma californica*) to >0.75 (California sheephead, *Semicossyphus pulcher*; cup corals, *Balanophyllia elegans* and *Astrangia lajollaensis*; Fig. 7, Appendix S1:

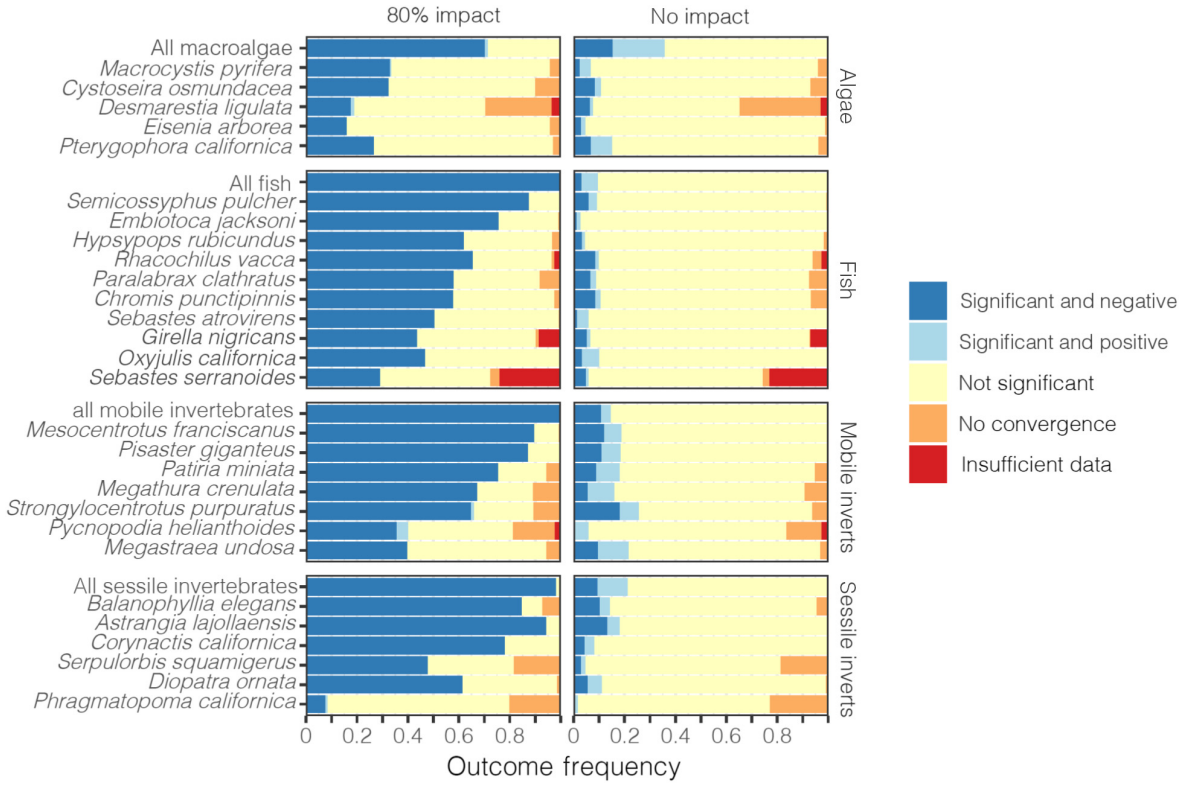

FIG. 4. Detection probabilities for hypothetical impacts as a function of species (rows) and scenario (columns). Left panels represent a large impact (i.e., an 80% reduction), whereas right panels represent scenarios where no impact was imposed. Individual species are plotted as well as results using all species within a taxonomic group in a nested framework (the top bar in each panel, excluding species sampled by point contact for macroalgae). Colors represent proportions of different outcomes from the simulations: "significant and negative" and "significant and positive" indicate a significant effect at alpha = 0.025 (without accounting for multiple comparisons) along with the direction of the estimated BACI effect. No convergence indicates models that failed to converge despite having sufficient data for analysis, where data sufficiency is at least two sites in each of the BA and CI categories with at least 15% of the years with the species present. The BACI was informative for a species when it (1) detected a significant negative impact (dark blue) in scenarios where we reduced species abundance by 80% or (2) found a nonsignificant impact (yellow) when the species abundance was not altered. Species are sorted according to their informedness.

Table S1). In a beta regression (intercept = -1.5 ± 0.82 [mean $\pm$ SE], $P = 0.0664$), three factors predicted a species' informedness, the fraction of sites with sufficient data (4.61 ± 0.61, $P < 0.0001$), followed by the mean CV of the control sites (-1.73 ± 0.46, $P = 0.002$), and the partial autocorrelation coefficient (-3.98 ± 0.87, $P < 0.0001$).

Species combinations

Statistical power in the nested BACI analysis increased substantially when species within a taxonomic group were analyzed collectively relative to when species were analyzed separately (Fig. 4, indicted by "all macroalgae," "all sessile inverts," "all mobile inverts," "all fish"). For three of the four taxonomic groups, false impact detections in the combined analyses resembled the mean independent false impact detection rate (a regression to the mean); macroalgae proved an exception to this pattern as combining the three species increased the probability of both false impact detections and the statistical power beyond what was observed for any of the component species (Fig. 4). Only three macroalgal species were pooled here because of differences in measurement types (point contact vs. counts).

Individual species informedness had a large impact on grouped informedness. When adding species starting with those with the highest individual informedness, grouped informedness rapidly saturated with the five or six most informed species, peaking after six to nine species were included and even falling slightly when species with lower informedness were added to the group (Fig. 8, blue points). In contrast, grouped informedness saturated much more slowly when adding species from lowest to highest informedness, and peaked only as the most effective 25% of species began to be added (Fig. 8, red points). Thus, although grouping species for analysis is promising, efficiency depended largely on which species were used in the analysis.

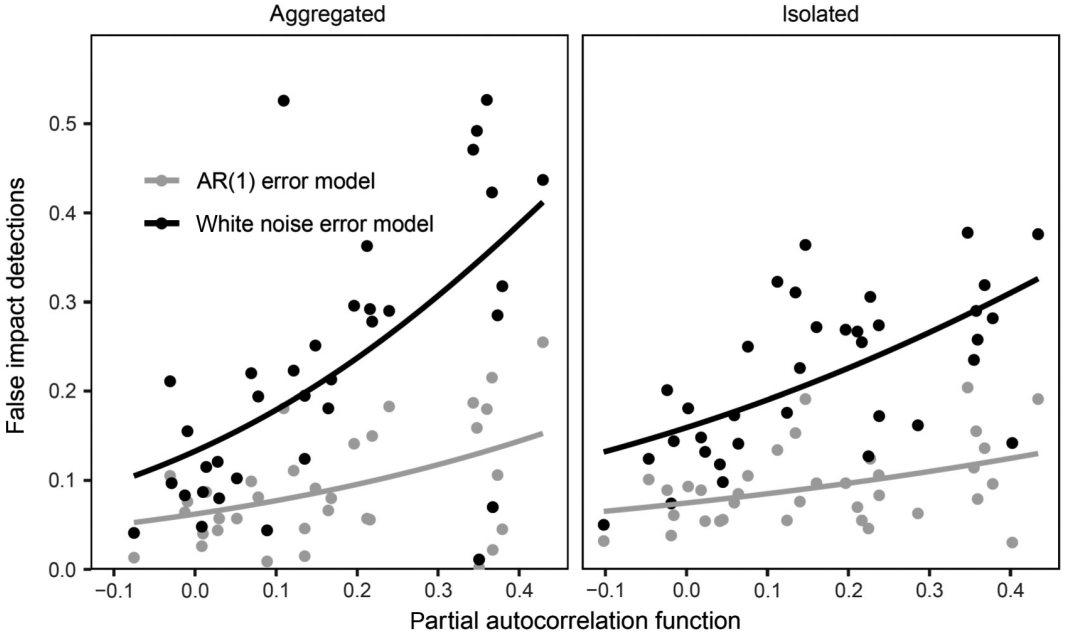

FIG. 5. Temporal autocorrelation and aggregated impacts increased false impact detections (black points), with more problematic rates for spatially aggregated impacts (left) than for impacts that occurred in isolated locations scattered across the region (right). Gray dots and lines indicate results accounting for first-order residual correlations in the models, black results neglect temporal autocorrelation. Each point represents the mean partial autocorrelation for an individual species across 1,000 simulations. Panels represent scenarios where the impacts are spatially aggregated or isolated (i.e., no spatial pattern in the impact).

TABLE 1. Standardized coefficients and 95% confidence intervals from the beta-binomial GLM.

	Power		False impact detections	
Source	Estimate	95% CI	Estimate	95% CI
Coefficient of variation	−0.71	[−1.01:−0.42]	−0.24	[−0.49:0.01]
No. sufficient sites	0.99	[0.68:1.31]	0.37	[0.06:0.69]
Spatial synchrony	0.15	[−0.13:0.43]	−0.18	[−0.45:0.08]
Serial correlation	−0.51	[−0.85:−0.17]	0.28	[−0.02:0.59]
Time series length	0.12	[−0.15:0.39]	−0.08	[−0.40:0.24]

Notes: Models for both power and false impact detections included all shown variables. For interpretation, the coefficients represent the rate of change of the response to a single standard deviation change in the predictor. Thus, all coefficients within a model are directly comparable.

DISCUSSION

BACI analyses were most effective at detecting severe impacts of intermediate spatial scales. Accounting for temporal autocorrelation resulted in unbiased estimates, but the power to detect impacts was quite variable across species. Observed power was often lower than that predicted by simulation studies with similar sample sizes (Christie et al. 2019), but better than assumed from short-term studies (Schroeter et al. 1993). Informedness metrics, which discount statistical power based on the false detection rate, revealed that the species that best inform impacts in this system are ubiquitous, have low year-to-year variation in abundance, and exhibit little temporal autocorrelation. Grouping species into broader taxonomic groups increased statistical power and reduced false impact detection rates, primarily due to including species with high informedness, highlighting how species selection can improve impact detection.

Substantial 5-yr impacts were surprisingly difficult to detect using a BACI design. The most common cause of analysis failures resulted from imbalance, high natural variability, and serial autocorrelation, but not spatial asynchrony. Many simulated impacts led to unbalanced designs, which reduced statistical power. Complicating matters further, some species abundances varied substantially from year to year, reducing the power to detect change, while other species experienced long-term natural trends that were difficult to differentiate from an impact, even after controlling for serial autocorrelation.

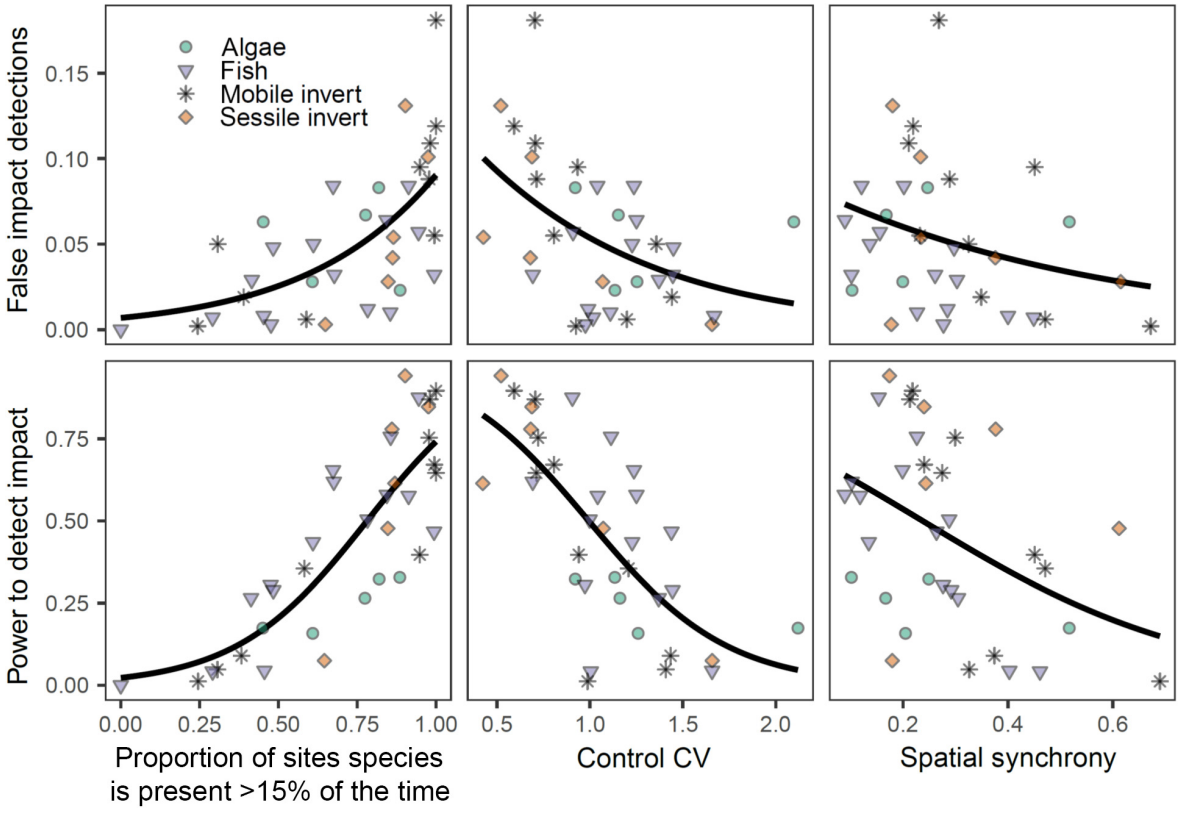

FIG. 6. False impact detections (top row) and statistical power to detect a hypothetical impact (bottom row) as a function of the proportion of sites where the species was present more than 15% of the time (left), of natural temporal variability (mean CV) measured at control sites (center), and of the degree of spatial synchrony of a species' population (right).

It is important, therefore, to account for serial correlations in BACI analyses, consider environmental drivers, and limit assessment to those species that are best suited for indicating impacts. Only by understanding sources of error can we improve impact detection.

As expected, we found that more severe impacts were more detectable, but that the relationship between impact spatial scale and detectability was more complex, and depended on the spatial extent of monitoring locations and their density. Impacts that were small relative to the average spacing between monitoring sites were often not detectable because there were few, if any, monitoring sites in the impacted area. Under real-life circumstances when only one monitoring site is impacted, an alternative approach is to determine if other monitoring sites are available, unimpacted, and possess coherent dynamics with the impacted site. If such a situation exists, one may investigate whether a BACI Paired-Series (BACIPS) design (Osenberg and Schmitt 1996) may have an enhanced ability to detect an impact compared to the region-wide BACI. In any case, description of natural variability obtained from long-term monitoring of the ecosystem can assist in developing a control-impact study should the small-sized perturbation completely miss any of the monitored sites.

By contrast, impacts that approach the spatial scale of the whole monitoring program often had few or no non-impacted control sites that could be used for control-impact comparisons, limiting impact detection to before-after impact analysis, which would only be appropriate for species where serial autocorrelation and environmental covariates could be controlled for statistically or vital rates can be measured directly. This result argues for scientists and resource managers to consider the spatial scale of potential threats that may endanger the managed area of interest when designing a monitoring program. Given that before-after designs often give biased estimates of impacts (Christie et al. 2019), setting up monitoring sites or collaborating with other monitoring groups outside the jurisdiction of the managed area should be considered if potential impacts are anticipated to have large spatial scales.

Managers can improve balance by choosing particular species and sites. Although expected sensitivity to potential impacts is of obvious importance, among similarly sensitive taxa, broadly distributed species (or species groups) are more likely to result in a balanced design than rare species, and focusing on such species is akin to increasing the number of monitored sites. Similarly, choosing some monitoring sites that are likely and unlikely to be impacted will increase balance. Such strategic

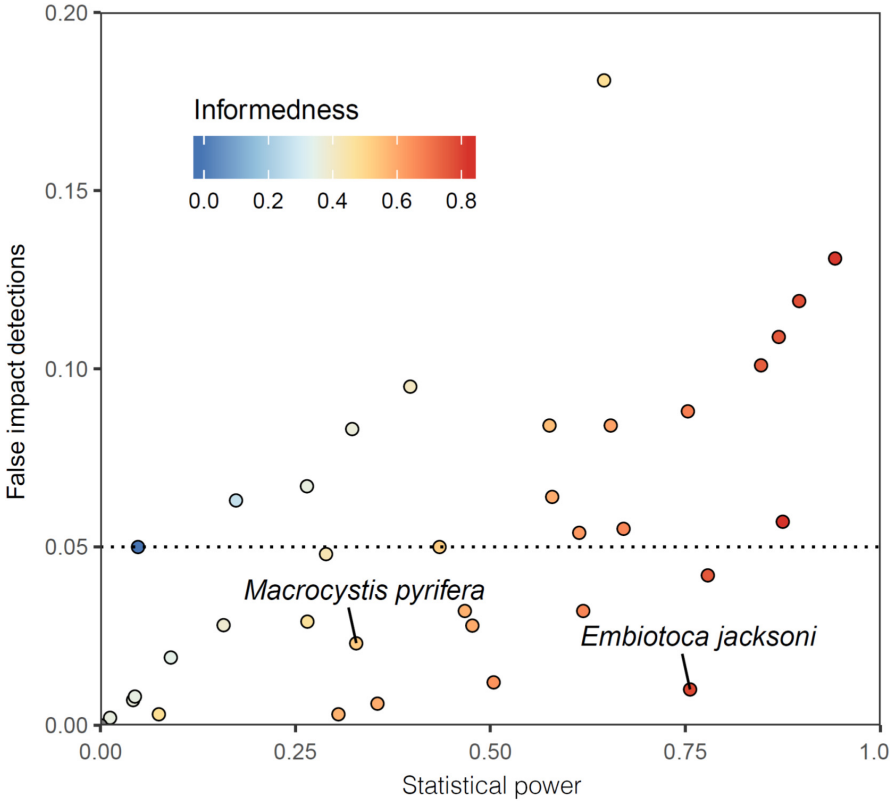

FIG. 7. False impact detections plotted against the statistical power to detect a hypothetical impact. Points represent individual species and colors represent informedness.

site selection is often advocated in textbook examples of BACI designs when the impact has a known spatial and temporal scope, such as might result from a new power plant or marine reserve (Schmitt and Osenberg 1996). When impacts are unplanned, such as an oil spill or a species introduction, it might still be possible to pick sites based on risk models. For example oil spill models use information about potential sources and ocean currents to identify sites at high and low risk of spilled oil (e.g., the General NOAA Operational Modeling Environment; Al Shami et al. 2017, Guo 2017, Li and Johnson 2019), which could form the basis for choices about where to monitor. However, when impacts are not predictable, or when a monitoring is not intended to address a specific anticipated impact (e.g., Davis 2005), then managers might wish to select sites that are either particularly sensitive, or of particular interest, as this is where impact assessments will be most relevant. This was the case for the data evaluated here. The area was mostly within a national park, and the sites were chosen for their high biodiversity value and favorable physical context (e.g., rocky reefs rather than soft sediment). These decisions increase the coverage of kelp forests within the data, but also limit the degree to which observations at these sites are representative of the coastline as a whole.

Temporal variation, reduced power, and serial autocorrelation increased false impact detection rates. False impact detections were particularly apparent for localized spatial impacts (the focus of this analysis), but also occurred when impacts were scattered across the region. Fortunately, a first-order autoregressive model with a common parameter among sites reduced the false impact detection rate by more than one-half (though it did not solve it completely), and suggests that this approach should be included in BACI designs, especially when the data include several pre-impact data points for assessing trends. The strength of this problem varied across species, a pattern to consider when choosing which species to monitor. We found that statistical power was highest and false detection rates lowest for species with low temporal variation, serial correlation, and spatial synchrony. Given ever-present constraints on resources, it is tempting to focus sampling on such tractable species. However, if more variable species are more sensitive to anthropogenic impacts, such a limited sampling regime might systematically underestimate the severity of these impacts, posing difficult trade-offs for program design.

Often, false impact detections were not due to error, but instead were real population trends that differed across the study region, violating the assumption that, in

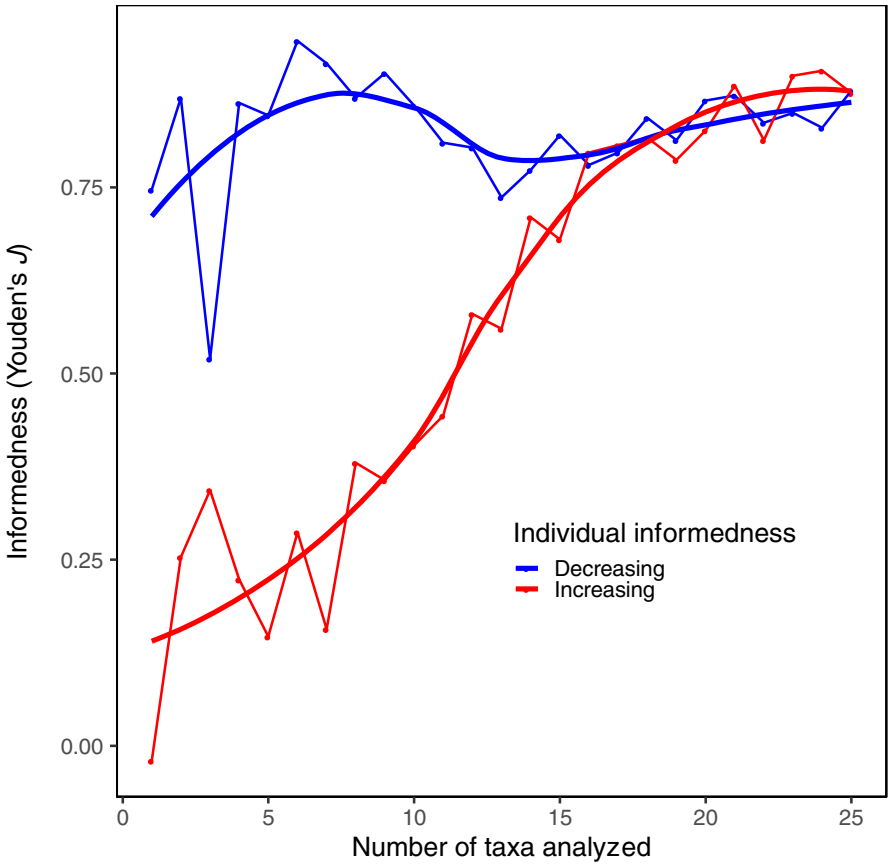

FIG. 8. Informedness as a function of the number of taxa in the grouped analysis and order in which species are added by their individual informedness. Informedness is the probability of correct inference under equal prior odds of impact or no impact.

the absence of an impact, species change in a similar way at both control and impact sites (Stewart-Oaten et al. 1986, Underwood 1992). Although our models accounted for temporal autocorrelation, it is considerably more complicated to adjust for patterns of spatial autocorrelation. The 23 monitoring sites in our analysis are irregularly distributed across a broad area, with large gaps between sites, and the spatial structure in populations of most species in this system are driven by complex environmental, biological, and anthropogenic factors that are often not well represented by geographic distance (Watson et al. 2011, Lamy et al. 2018). When impacts were randomized in space, such that distant sites were as likely to be co-affected as nearby ones, false impact detections were reduced (Appendix S1: Fig. S4). In this case, 15 of 28 species rather than 11 of 28 had a <5% false impact detection rate. Therefore, when assessing localized impacts, it is important to evaluate how much spatial autocorrelation is a problem, and identify species that are particularly sensitive to spatial autocorrelation. Simulations like the ones done here are one route to obtaining such knowledge. These results also emphasize the importance of including key environmental covariates whenever possible.

Impacts on some species may be difficult to detect even with an extensive monitoring program. The macroalgae in this data set provide a striking example of this; a severe impact would be detected less than half the time for any of the five species examined, either because the species had high unexplained variance, or because it was not found at many sites. For example, the five species with the lowest Informedness were a mix of algae and invertebrates (*Phragmatopoma californica*, *Eisenia arborea*, *Desmarestia ligualata*, *Megastraea undosa*, and *Pterygophora californica*), but all had temporal coefficients of variation much larger than their mean abundances, and each species was observed at fewer than 77% of the sites. Not only do such species make poor indicators, but the false discovery rate increases with each species analyzed.

We suggest three approaches to the species problem. The first is to accept that some species are poor indicators and limit the expectations that impacts on such species will be detected. Indicator value, which is best measured as informedness, was predictable from information on prevalence (% times seen), average (unexplained) CV, and the partial autocorrelation coefficient. Therefore, the species that best inform impacts in this

system are ubiquitous, have little year-to-year variation in abundance and little temporal autocorrelation. The second approach is to group species according to ecological role, morphological traits or taxonomic affinities. The average across species should have higher prevalence and a lower CV, and a lower partial autocorrelation coefficient than the average within species. We grouped species within broad taxonomic categories, but it might be more efficient to average species that correlate negatively with one another. Although our analysis did not simulate the cascading effects of species interactions such as competition and predation, interactions of this type might lead to negative correlations of this sort. Averaging always improved statistical power, and usually reduced false impact detections. However, averaging was most effective when poor indicators were excluded. In other words, one bad alga could spoil the bunch. A third approach is needed when variable species are a high priority for monitoring, or when dropping species might bias interpretations about impact severity. For such species, managers might consider changing the sampling technique, scale, or frequency to increase a species' informedness for impact detection.

Due to natural temporal trends, and patchy distributions, ecological data often fail to meet BACI assumptions, and this can lead to low power to detect an impact and high false impact detection rates. Such errors compromise efforts to accurately attribute changes to impacts, and thus make it tenuous to assign culpability or test hypotheses about complex system responses to perturbations. However, strategic site placement, accounting for serial autocorrelations, and analyzing species, or species groupings, with high informedness could overcome many limitations in this large data set from kelp forest communities. Such findings may be useful to other settings. In particular, we hope it stimulates coordination among similar monitoring programs (as has been done here through cooperation between the National Park Service and the U.S. Geological Survey) to maximize the ability to understand the ecological consequences of catastrophic events such as the *Exxon Valdez* oil spill. Such understanding will not only help to assign culpability of impacts through informed natural resource damage assessments should an accident happen, but will also enable a fuller understanding of ecosystems and the natural resources that comprise them.

Acknowledgments

We thank Jim Estes, Tim Tinker, Mark Novak, and Mike Kenner for comments on this work. The manuscript was also substantially improved by constructive comments from Juniper Simonis and another reviewer. Funding for fieldwork and data collection was provided by the United States National Park Service, U.S. Geological Survey, US Fish and Wildlife Service and University of California Santa Cruz. Data set integration across monitoring programs and subsequent analyses were supported with funds from the U.S. Department of the Interior, Bureau of Ocean Energy Management, Environmental Studies Program, Washington, D.C. under Cooperative Agreement Number M11AC00012, and the National Science Foundation via its support of the Santa Barbara Coastal LTER (OCE-1831937), and from the Santa Barbara Channel Marine Biodiversity Observation Network (NASA Grant NNX-14AR62A and BOEM Cooperative Agreement MC15AC00006). All authors planned the research and contributed to the manuscript. D. Okamoto led analyses with help from A. Rassweiler and K. Lafferty; A. Rassweiler, D. Okamoto, K. Lafferty, and D. Reed jointly wrote the manuscript. This product has been peer reviewed and approved for publication consistent with USGS Fundamental Science Practices (http://pubs.usgs.gov/circ/1367/). Mention of trade names or commercial products does not constitute their endorsement by the U.S. Government.

Literature Cited

Al Shami, A., G. Harik, I. Alameddine, D. Bruschi, D. A. Garcia, and M. El-Fadel. 2017. Risk assessment of oil spills along the Mediterranean coast: a sensitivity analysis of the choice of hazard quantification. Science of the Total Environment 574:234–245.

Bell, T. W., K. C. Cavanaugh, D. C. Reed, and D. A. Siegel. 2015. Geographical variability in the controls of giant kelp biomass dynamics. Journal of Biogeography 42:2010–2021.

Caselle, J. E., A. Rassweiler, S. L. Hamilton, and R. R. Warner. 2015. Recovery trajectories of kelp forest animals are rapid yet spatially variable across a network of temperate marine protected areas. Scientific Reports 5:1–14.

Castorani, M. C. N., D. C. Reed, and R. J. Miller. 2018. Loss of foundation species: disturbance frequency outweighs severity in structuring kelp forest communities. Ecology 99:2442–2454.

Christie, A. P., T. Amano, P. A. Martin, G. E. Shackelford, B. I. Simmons, and W. J. Sutherland. 2019. Simple study designs in ecology produce inaccurate estimates of biodiversity responses. Journal of Applied Ecology 56:2742–2754.

Davis, G. E. 2005. National Park stewardship and "vital signs" monitoring: a case study from Channel Islands National Park, California. Aquatic Conservation: Marine and Freshwater Ecosystems 15:71–89.

Dietl, G. P., and S. R. Durham. 2016. Geohistorical records indicate no impact of the Deepwater Horizon oil spill on oyster body size. Royal Society Open Science 3:160763.

Ebeling, A. W., Laur, D. R., and Rowley, R. J. 1985. Severe storm disturbances and reversal of community structure in a southern California kelp forest. Marine Biology 84:287–294.

Fancy, S. G., J. E. Gross, and S. L. Carter. 2009. Monitoring the condition of natural resources in US national parks. Environmental Monitoring and Assessment 151:161–174.

Field, S. A., P. J. O'Connor, A. J. Tyre, and H. P. Possingham. 2007. Making monitoring meaningful. Austral Ecology 32:485–491.

Fukuyama, A. K., G. Shigenaka, and D. A. Coats. 2014. Status of intertidal infaunal communities following the Exxon Valdez oil spill in Prince William Sound, Alaska. Marine Pollution Bulletin 84:56–69.

Guo, W. 2017. Development of a statistical oil spill model for risk assessment. Environmental Pollution 230:945–953.

Harrold, C., and Reed, D. C. 1985. Food availability, sea urchin grazing, and kelp forest community structure. Ecology 66:1160–1169.

Harvell, C. D., et al. 2019. Disease epidemic and a marine heat wave are associated with the continental-scale collapse of a pivotal predator (*Pycnopodia helianthoides*). Science Advances 5:eaau7042.

Holbrook, S. J., R. J. Schmitt, and J. S. Stephens. 1997. Changes in an assemblage of temperate reef fishes associated with a climate shift. Ecological Applications 7:1299–1310.

Jewett, S. C., T. A. Dean, R. O. Smith, and A. Blanchard. 1999. "Exxon Valdez" oil spill: Impacts and recovery in the soft-bottom benthic community in and adjacent to eelgrass beds. Marine Ecology Progress Series 185:59–83.

Kalies, E. L., C. L. Chambers, and W. W. Covington. 2010. Wildlife responses to thinning and burning treatments in southwestern conifer forests: a meta-analysis. Forest Ecology and Management 259:333–342.

Kenner, M. C., J. A. Estes, M. T. Tinker, J. L. Bodkin, R. K. Cowen, C. Harrold, B. B. Hatfield, M. Novak, A. Rassweiler, and D. C. Reed. 2013. A multi-decade time series of kelp forest community structure at San Nicolas Island, California (USA). Ecology 94:2654–2654.

Kenner, M. C., and M. T. Tinker. 2019. Stability and change in kelp forest habitats at San Nicolas Island. Western North American Naturalist 78:633.

Kushner, D. J., A. Rassweiler, J. P. McLaughlin, and K. D. Lafferty. 2013. A multi-decade time series of kelp forest community structure at the California Channel Islands. Ecology 94:2655.

Lamy, T., D. C. Reed, A. Rassweiler, D. A. Siegel, L. Kui, T. W. Bell, R. D. Simons, and R. J. Miller. 2018. Scale-specific drivers of kelp forest communities. Oecologia 186:217–233.

Lamy, T., S. Wang, D. Renard, K. D. Lafferty, D. C. Reed, and R. J. Miller. 2019. Species insurance trumps spatial insurance in stabilizing biomass of a marine macroalgal metacommunity. Ecology 100:1–10.

Lauritsen, A. M., P. M. Dixon, D. Cacela, B. Brost, R. Hardy, S. L. MacPherson, A. Meylan, B. P. Wallace, and B. Witherington. 2017. Impact of the Deepwater Horizon oil spill on loggerhead turtle *Caretta caretta* nest densities in northwest Florida. Endangered Species Research 33:83–93.

Li, Z., and W. Johnson. 2019. An improved method to estimate the probability of oil spill contact to environmental resources in the Gulf of Mexico. Journal of Marine Science and Engineering 7:41.

Loreau, M., and C. de Mazancourt. 2008. Species synchrony and its drivers: neutral and nonneutral community dynamics in fluctuating environments. American Naturalist 172:E48–E66.

Magnusson, A., H. A. Skaug, C. Nielsen, K. Berg, M. Kristensen, K. Maechler, V. Bentham, B. Bolker, and M. Brooks. 2016. glmmTMB: generalized linear mixed models using Template Model Builder. R package version 0.0 2. https://cran.r-project.org/web/packages/glmmTMB/index.html

Martínez-Abraín, A., C. Viedma, J. A. Gómez, M. A. Bartolomé, J. Jiménez, M. Genovart, and S. Tenan. 2013. Assessing the effectiveness of a hunting moratorium on target and non-target species. Biological Conservation 165:171–178.

Meyer, C. F. J., et al. 2010. Long-term monitoring of tropical bats for anthropogenic impact assessment: gauging the statistical power to detect population change. Biological Conservation 143:2797–2807.

Miller, R. J., K. D. Lafferty, T. Lamy, L. Kui, A. Rassweiler, and D. C. Reed. 2018. Giant kelp, *Macrocystis pyrifera*, increases faunal diversity through physical engineering. Proceedings of the Royal Society B 285:20172571. http://dx.doi.org/10.1098/rspb.2017.2571

Okamoto, D. K. 2020. Code for BACI with long term monitoring data. Code archive on Zenodo. https://doi.org/10.5281/zenodo.4265329

Okamoto, D. K., R. J. Schmitt, and S. J. Holbrook. 2016. Stochastic density effects on adult fish survival and implications for population fluctuations. Ecology Letters 19:153–162.

Okamoto, D. K., R. J. Schmitt, S. J. Holbrook, and D. C. Reed. 2012. Fluctuations in food supply drive recruitment variation in a marine fish. Proceedings of the Royal Society B 279:4542–4550.

Okamoto, D. K., S. C. Schroeter, and D. C. Reed. 2020. Effects of ocean climate on spatiotemporal variation in sea urchin settlement and recruitment. Limnology and Oceanography 65:2076–2091.

Osenberg, C. W., and R. J. Schmitt. 1996. Detecting ecological impacts caused by human activities. Pages 3–16 *in* R. J. Schmitt and C. W. Osenberg, editors. Detecting ecological impacts: concepts and applications in coastal habitats. Academic Press, San Diego, CA, USA.

Parker, K. R., and J. A. Wiens. 2005. Assessing recovery following environmental accidents: environmental variation, ecological assumptions, and strategies. Ecological Applications 15:2037–2051.

Parnell, P., C. Lennert-Cody, L. Geelen, L. Stanley, and P. Dayton. 2005. Effectiveness of a small marine reserve in southern California. Marine Ecology Progress Series 296:39–52.

Powers, D. M. W. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies 2:37–63.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org

Ramsey, F. L., and D. W. Schafer. 2002. The statistical sleuth—a course in methods of data analysis.

Rassweiler, A., R. J. Schmitt, and S. J. Holbrook. 2010. Triggers and maintenance of multiple shifts in the state of a natural community. Oecologia 164:489–498.

Reed, D. C., A. Rassweiler, M. H. Carr, K. C. Cavanaugh, D. P. Malone, and D. A. Siegel. 2011. Wave disturbance overwhelms top-down and bottom-up control of primary production in California kelp forests. Ecology 92:2108–2116.

Reed, D. C., A. R. Rassweiler, R. J. Miller, H. M. Page, and S. J. Holbrook. 2016. The value of a broad temporal and spatial perspective in understanding dynamics of kelp forest ecosystems. Marine and Freshwater Research 67:14–24.

Roberge, J. M., and P. Angelstam. 2006. Indicator species among resident forest birds—a cross-regional evaluation in northern Europe. Biological Conservation 130:134–147.

Rost, J., M. Clavero, L. Brotons, and P. Pons. 2012. The effect of postfire salvage logging on bird communities in Mediterranean pine forests: the benefits for declining species. Journal of Applied Ecology 49:644–651.

Schmitt, R. J., and C. W. Osenberg. 1996. Detecting ecological impacts: concepts and applications in coastal habitats. Pages 401. Environmental Science and Engineering. First edition.

Schroeter, S. C., J. D. Dixon, J. Kastendiek, and R. O. Smith. 1993. Detecting the ecological effects of environmental impacts—a case-study of kelp forest invertebrates. Ecological Applications 3:331–350.

Schroeter, S. C., D. C. Reed, D. J. Kushner, J. A. Estes, and D. S. Ono. 2001. The use of marine reserves in evaluating the dive fishery for the warty cucumber (*Parastichopus parvimensis*) in California, U.S.A. Canadian Journal of Fisheries and Aquatic Sciences 58:1773–1781.

Shaw, R. G., and T. Mitchell-Olds. 1993. ANOVA for unbalanced data: an overview. Ecology 74:1638–1645.

Skalski, J. R., D. A. Coats, and A. K. Fukuyama. 2001. Criteria for oil spill recovery: a case study of the intertidal community of Prince William Sound, Alaska, following the Exxon Valdez oil spill. Environmental Management 28:9–18.

Stewart-Oaten, A. 2003. On rejection rates of paired intervention analysis: comment. Ecology 84:2795–2799.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67:929.

Thiault, L., L. Kernaléguen, C. W. Osenberg, and J. Claudet. 2017. Progressive-change BACIPS: a flexible approach for environmental impact assessment. Methods in Ecology and Evolution 8:288–296.

Underwood, A. J. J. 1992. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. Journal of Experimental Marine Biology and Ecology 161:145–178.

Underwood, A. J. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4:3–15.

Watson, J. R., C. G. C. Hays, P. T. P. Raimondi, S. Mitarai, C. Dong, J. C. J. McWilliams, C. A. C. Blanchette, J. E. J. Caselle, and D. A. DA Siegel. 2011. Currents connecting communities: nearshore community similarity and ocean circulation. Ecology 92:1193–1200.

Youden, W. J. 1950. Index for rating diagnostic tests. Cancer 3:32–35.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/eap.2304/full

Data Availability

Code is available in Zenodo (Okamoto 2020): https://doi.org/10.5281/zenodo.4265329.