Ecological Applications, 15(6), 2005, pp. 2037–2051

ASSESSING RECOVERY FOLLOWING ENVIRONMENTAL ACCIDENTS: ENVIRONMENTAL VARIATION, ECOLOGICAL ASSUMPTIONS, AND STRATEGIES

KEITH R. PARKER[1,3] AND JOHN A. WIENS[2]

[1]*Data Analysis Group, P.O. Box 128, Cloverdale, California 95425 USA*
[2]*The Nature Conservancy, 4245 North Fairfax Drive, Arlington, Virginia 22203 USA*

Abstract. Gauging whether or when a population, species, or community recovers from an environmental accident or disturbance, such as an oil spill or forest fire, is complicated by environmental variation in time and space, and therefore depends on the assumptions one makes about equilibrium. These ecological assumptions about equilibrium affect how one designs and interprets studies to assess recovery from environmental accidents or disturbances. We use examples from studies conducted following the *Exxon Valdez* oil spill to illustrate several approaches to assessing recovery and their sensitivity to the form of equilibrium one assumes. Baseline study designs, which compare levels of a resource after the disturbance to pre-disturbance levels for impact data only, are generally inadequate because they rest on the unrealistic assumption of steady-state equilibrium. Since data for the impact area only are used, recovery and temporal variation are confounded. Unlike baseline designs, before–after control–impact (BACI) designs use impact and reference data, and relax this sensitivity by incorporating both temporal and spatial variation. Studies that compare impacted with reference areas in a single year following the disturbance assume spatial equilibrium and therefore may confound recovery with systematic spatial differences between the areas. Sampling and analytical strategies such as stratified random sampling or the use of environmental measures as covariates may lessen the sensitivity to this assumption. Multiyear studies that include comparisons between impacted and reference areas or that sample areas along a gradient of disturbance rest on the more realistic assumption of dynamic equilibrium.

Understanding the underlying assumptions and how they relate to the approach one uses must be part of assessing the recovery of biological resources from an environmental accident. Because the dynamics of different populations, species, and communities and the environments they occupy vary and exhibit different dependencies on the scale of disturbance (and the scale of analysis), there is no single ''best'' approach to assessing recovery. Discussions about recovery should include an explicit and honest consideration of the underlying ecological assumptions, the likelihood that they hold in the system being studied, and the consequences if the assumptions are violated.

Key words: ecological assumptions; environmental accident; equilibrium; Exxon Valdez; impact; recovery; spatial and temporal variation.

INTRODUCTION

When an environmental accident such as an oil spill occurs, the first question is ''What were the impacts?'' If scientific studies demonstrate that an impact occurred, the next question is ''When is the system recovered?'' The impacts of the accident often appear quickly, but recovery is a process that may take much longer. For example, an oil or chemical spill occurs at a known point in time and the greatest injury to biota typically occurs within days of the accident. In contrast, recovery may depend on a variety of factors, such as rates of decontamination, recruitment, succession and restoration of food sources (Wiens 1995). Time to recovery is therefore indeterminate and may take months or years (Skalski et al. 2001). Gauging when recovery occurs is difficult, but it is critically important for legal, management, aesthetic, scientific, and ethical reasons.

Determining when recovery has occurred requires that the target—the state of the ''recovered'' system—be specified. In this sense, assessing recovery is similar to doing ecological restoration, which requires a clear statement of the restoration goal or endpoint. Restoration ecologists have grappled with this problem, even if they have not entirely resolved it (Bradshaw 2002, Hobbs 2004). In an ecological sense, restoration can be defined as ''the return of an ecosystem to a close approximation of its condition prior to disturbance'' (National Research Council 1992). The goal is to ''emulate a natural, functioning, self-regulating system that is integrated with the ecological landscape in which it

Manuscript received 15 November 2004; revised 9 March 2005; accepted 31 March 2005. Corresponding Editor: F. C. James.

[3] E-mail: krparker@sonic.net

occurs'' (National Research Council 1992) or to reestablish ''a critical range of variability in biodiversity, ecological processes and structures, regional and historical context, and sustainable cultural practices'' (Society for Ecological Restoration 1996). Gauging the recovery of an endangered species involves similar issues, although the goal or endpoint usually is framed in terms of the viability or sustainability of populations of the species.

Although these statements may seem clear enough, problems arise in their interpretation. Should restoration aim to recreate the conditions that existed just before the system was damaged or degraded, or should it be conditions at some time in the past before humans started to modify it? Should the goal be the restoration of a system identical in function and structure to that which existed at some previous time, or to return the system to the envelope of ''natural range of variability'' of the system? What is the natural range of variability of the system? Can studies be implemented to assess recovery? How will decisions be made to determine when recovery has occurred?

For recovery from accidental impacts, the goal is seemingly more straightforward—a return to what the system would have been like had the accident not occurred, taking into account the effects of natural variation (U.S. Code of Federal Regulations 2001). Lurking beneath the surface of this statement, however, are the same problems that restoration ecologists face. The similarities between assessing recovery and doing restoration mean that the issues we discuss here with reference to recovery are equally relevant to restoration efforts.

And there are issues. In the end, these issues relate to how one is to decide when recovery has occurred or restoration goals have been met. It is simple enough to make decisions in a stable, unchanging environment—recovery occurs when the state of the system after the impact matches its state before the impact or that of a similar, unimpacted reference area. This, in fact, is the view of recovery embodied in assessing the damages from environmental accidents by the *Exxon Valdez* Oil Spill Trustee Council (hereafter, ''Trustees''). The Trustee Council was formed to oversee restoration of the injured ecosystem through the use of a $900 million civil settlement. The Council consists of three state and three federal trustees charged with assessing impact and recovery from the *Exxon Valdez* oil spill. Scientific work is carried out under the aegis of federal and state agencies). For example, the *Exxon Valdez* Oil Spill Trustee Council (2002) defines recovery of sea otters (*Enhydra lutris*) and eight other taxa as having occurred when the population in oiled areas returns to its pre-spill levels, and Peterson (2001, Peterson et al. 2004) used a failure of populations to return to pre-spill levels as evidence of continuing impacts.

This equilibrium view reflects the notion of a ''balance of nature'' that long dominated thinking among both ecologists and the general public (Wiens 1977, Worster 1977, Pimm 1991). Most ecologists, however, now believe that such strict equilibrium is rare in biological systems (Wiens 1984, Chesson and Case 1986, Giller and Gee 1987), primarily because it fails to take into account natural variation. Systems change over time, and even seemingly similar places may differ from one another in important attributes. This means that ''recovery'' must be assessed against a backdrop of both temporal and spatial variation, some of which is a natural part of the system, some of which may be anthropogenic in origin, and some of which is seemingly random (largely due to the influences of factors operating at geographic scales other than those being considered). This variation creates a paradox in assessing recovery. Recovery is likely to take a long time, so the longer one evaluates a system, the more likely one is to document recovery and verify its persistence. With a longer time frame, however, more things happen to the biological resource being assessed and the effects of natural variation aggregate and cascade. With more time, the goal of recovery increasingly becomes an unpredictably moving target, creating both conceptual and analytical challenges.

Our objective here is to review the ways in which assumptions about natural variation affect how one thinks about and assesses recovery from environmental impacts. In a previous paper (Wiens and Parker 1995), we focused on how ecological assumptions and natural variation relate to the sampling methodology and statistical analysis used for assessing accidental impacts. That analysis was based on our experience in assessing environmental impacts from the *Exxon Valdez* oil spill. We paid little attention to assessing recovery, primarily because at the time the paper was written there was incomplete information on recovery. Since 1995, several investigators have used long-term studies to assess the recovery of impacted biota. These studies illustrate the importance of separating the recovery signal from natural variation and of verifying the ecological assumptions on which detecting recovery depends.

Of course, everything in ecology is sensitive to scale, and assessments of recovery are no exception. Whether one sees recovery or not depends on the time scale of reference and the spatial scale or resolution of analysis, and because different species in a system function at different scales of time and space, the scale of multispecies assessments may be appropriate for some species but not for others (e.g., intertidal invertebrates vs. migratory birds). We do not explicitly consider these scaling issues here; they are important, but we believe that our comments and guidance on assessing recovery hold regardless of the scale of a disturbance or of the systems being investigated.

As we did before (Wiens and Parker 1995), we illustrate issues and strategies in assessing recovery with studies from the *Exxon Valdez* oil spill. The *Exxon Valdez* ran aground on Bligh Reef in Prince William

Sound, Alaska (hereafter, PWS) on 24 March 1989, spilling 41 000 m^3 of North Slope crude oil. The spill affected ~2100 km of shoreline and was observed as far away as 970 km from the spill site (Neff et al. 1995). The chemical and biological effects of this spill have been discussed extensively (e.g., Wheelwright 1994, Wells et al. 1995, Rice et al. 1996, Irons et al. 2000, Peterson 2001, Wiens et al. 2001, 2004), with no overall resolution of the recovery status of many species or of the amount of residual oil remaining in shoreline sediments and its bioavailability (Boehm et al. 2004, Short et al. 2004).

For convenience, we adopt several terminological conventions. ''Biological resources'' (sometimes shortened to ''resource'') are quantifiable components of the systems such as organisms, populations, species, and communities. ''Levels'' of a resource are measures such as abundance, diversity, community structure, reproductive rates, mortality, or age. Hence, levels are quantifiable on an objective scale and can be used to estimate means and variance and to test hypotheses. ''Natural factors'' are the physical and chemical features of the environment that affect the level of a resource at a given time and location (e.g., temperature, substrate, dissolved oxygen, wave energy, total organic carbon). On occasion, we refer to impacted resources as ''injured.'' We use ''gradient'' analysis and ''dose–response regression'' interchangeably, in which dose is a measure of exposure to oil and response is a measure of the biological system.

Defining Recovery

Recovery is a temporal process in which impacts (e.g., contamination or physical alterations of habitat) progressively lessen through natural processes and/or active restoration efforts and natural factors regain their influence over the biological resource being assessed. There also is a spatial dimension to recovery, because locations differ in magnitude of impact and have different dynamics of natural factors. Although the temporal dynamics of the impacted environment and rates of recovery may be more similar at finer spatial scales, this depends on the inherent spatial heterogeneity of the system.

We define recovery as occurring when the injured resource reaches the level which it would have been, had it not been injured in the first place. After recovery occurs, the influence of impact-related factors will have diminished to the point where levels of the resource vary temporally in a natural way. This definition is based on Federal NRDA regulations under CERCLA (U.S. Code of Federal Regulations, Revised 2001, Volume 43, Sections 11:14 and 11:72), in which the recovery period is defined as the length of time required to return the services of the injured resource to their baseline condition. Baseline is ''the conditions that would have been expected at the assessment area had the discharge of oil or release of hazardous substance under investigation not occurred, taking into account both natural processes and those that are the result of human activities.'' Importantly, CERCLA recognizes that natural variation and anthropogenic factors are also at play during the period of decontamination and recovery; the recovered state is not necessarily a previous condition but an expected condition based on the recovery process, natural variation, and anthropogenic effects. Further, by using ''level,'' CERCLA recognizes that the recovered state of a resource can be measured in ways other than abundance and leaves open the possibility of assessing recovery in terms of diversity, species richness, variability, or other metrics.

In the operational phase of determining recovery, one compares impacted and non-impacted estimates of levels of the resource. Such comparisons are inferential and rely on statistical tests of hypotheses. Thus, statistically based measures of precision or uncertainty need to be considered to determine the confidence with which recovery can or cannot be inferred. The degree of such confidence depends on both the length of time to recovery and the variable temporal and spatial dynamics of the environment in which recovery takes place. These dynamics usually are unknown. Clearly, varying levels of impacts across locations that differ in environmental dynamics will complicate assessments of recovery and can lead to incorrect conclusions about if and when recovery occurs.

Ecological Assumptions

Assessments of recovery make assumptions about the nature of temporal and spatial variability of the systems being studied, but such assumptions often are neither acknowledged nor tested. Our point in this paper is that the assumptions one makes about ecological variability make all the difference in how or whether one can assess recovery (Green 1979, Wiens and Parker 1995).

Consider, for example, coastal marine ecosystems such as those impacted by the *Exxon Valdez* oil spill. Seasonal changes can be dramatic, especially at higher latitudes (e.g., Prince William Sound, Alaska) where ice, wind, and waves can ''reset'' shoreline ecosystems annually. Multiyear trends in climatic and oceanographic conditions affect temperature and salinity (Hare and Mantua 2000, Peterson and Schwing 2003). Populations of migrating seabirds, marine mammals, and fish are further subjected to temporal variations in the environments they encounter during migration. The complex mix of shoreline substrates, landforms, islands, currents, weather patterns, and past history of disturbance create considerable spatial variation in environmental conditions as well. These temporal and spatial variations affect the distribution and abundance of marine organisms, from algae to grazers and filter-feeders to top predators.

Following Wiens and Parker (1995), we categorize assumptions about the temporal and spatial equilibrium

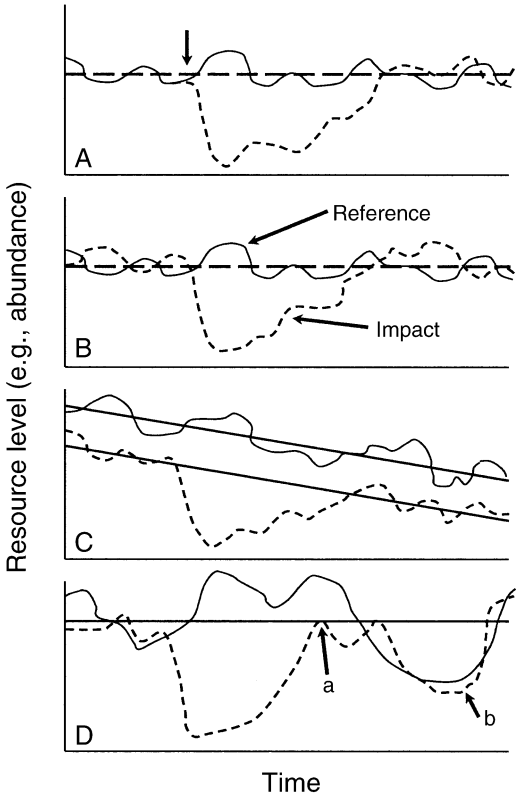

FIG. 1. Ecological assumptions affecting the assessment of recovery from an environmental accident. (A) Steady-state equilibrium. The vertical arrow indicates the accident, and the dotted line the state of the affected system; the solid curve indicates the dynamics of the system in the absence of a perturbation, and the solid horizontal line the steady-state mean level of the resource. Recovery occurs when the impacted system returns to the mean steady-state equilibrium. (B) Spatial equilibrium. In this case, two different areas have similar long-term dynamics. The unaffected system serves as a reference for the impacted system, and recovery occurs when the impacted system returns to the point where its dynamics are again similar to those of the reference site. (C) Dynamic equilibrium. The reference and impact areas have different levels of the resource (spatial variation), but their temporal dynamics are similar (in this example, a long-term decline). Recovery occurs when the dynamics of the impacted system once again parallel those of the reference system, even though levels of the resource differ between the areas. (D) In this example, there is considerable natural variation about the long-term steady-state mean. If this long-term variation is not considered or is unknown, the impacted system may erroneously be deemed recovered when it is not (point a) or may be considered still to be impacted when its dynamics in fact match those of an unimpacted system (point b).

conditions of a system that has not been impacted by an environmental perturbation such as an oil spill as steady-state equilibrium, spatial equilibrium, and dynamic equilibrium.

In the strict sense, steady-state equilibrium implies that features of the system being considered do not vary in time. In practice, this assumption is usually applied to a single location (e.g., the area impacted by an oil spill or fire) or species. Levels of the resource, and the natural factors controlling them, have a constant mean value over time (Fig. 1A). The resource at a given location has a single long-term equilibrium to which it will return if perturbed. Spatial equilibrium occurs when two (or more) sampling areas (e.g., impact and reference) have equal natural factors and, consequently, similar levels of a resource (Fig. 1B). Spatial equilibrium implies that mean levels of a resource are equal among areas; in the absence of an impact event, differences in means are due to sampling error and stochastic variations. The assumption of dynamic equilibrium recognizes variation in both space and time. Natural factors and levels of resources normally differ between two or more areas being compared, but the differences between mean levels of the resource remain constant over time (Fig. 1C). Mean levels can change over time for the areas, but such changes are similar among the areas. In the absence of an impact, the systems in different areas will track one another over time.

We have framed these assumptions as they relate to attempts to assess recovery in real-world situations. It may be useful, however, to digress briefly to examine how these different assumptions relate to a theoretical model such as that for logistic population growth. In the logistic model

$$dN/dt = rN\,(1 - N/K) \qquad (1)$$

changes in population size (or resource level), N, are a function of the intrinsic rate of change (r) and the carrying capacity of the environment (K; see Kingsland 1985 for an historical perspective). There are two equilibria in this model, when $N = 0$ and when $dN/dt = 0$ (i.e., $N = K$). The former is uninteresting from our perspective; the latter corresponds directly to steady-state equilibrium. Spatial equilibrium implies that equilibrium population sizes (K) are equal in the different areas or systems being compared (e.g., impact [i] and reference [r] areas). Note that this assumption does not require that $r_i = r_r$ or that the rate of change following a disturbance be the same (i.e., $dN/dt_i = dN/dt_r$), only that the eventual equilibrium be the same in all places being considered. Under the assumption of dynamic equilibrium, K may differ among areas but $r_i = r_r$, so (because r is a per-capita rate), $dN/dt_i = dN/dt_r$; growth trajectories of the systems in different places will parallel one another over time, but at different levels (N).

Even (or perhaps especially) such a simple theoretical model of system dynamics is operationally unrealistic in most real-world situations. A theoretical ecologist's interest in the logistic model may be in the way it portrays process, through the influences of different values of r or the effects of time lags on system trajectories (e.g., limit cycles). An applied ecologist's interests, in contrast, are phenomenological, focusing on the outcomes of the dynamics rather than the dynamics themselves. We make comparisons of N over time, between areas, or both, and the different assumptions lead us to hypothesize similar or different values of N. Alternatively, we might focus on dN/dt, again with our expectations differing depending on the underlying assumptions. We do not focus on r or K, largely because in practice neither their values nor their variances are known or, in many cases, knowable. In a sense, we are suggesting that, operationally, it may not really matter much if the components of r (rates of birth, death, emigration, and immigration) are still affected by a disturbance, so long as N meets some hypothesized state that we can label ''recovery.''

Given this phenomenological focus, how do the different assumptions about the temporal and spatial variability of a system lead operationally to different ways of assessing recovery? For steady-state equilibrium, ongoing impacts occur when mean pre- and post-event levels differ; recovery occurs when mean levels no longer differ (Fig. 1A). If natural factors vary temporally at the impacted sites, a relaxed form of the assumption of steady-state equilibrium may still hold if the variance about the mean remains within a constant envelope. Under these conditions, however, attempts to assess recovery may be confounded with natural changes over time. As a result, an impacted system may mistakenly be deemed ''recovered'' or ''not recovered'' depending on when its natural temporal variations intersect some long-term mean that is presumed to be the steady-state equilibrium (Fig. 1D). For a system in decline due to long-term environmental changes or anthropogenic factors, such as depletion of prey resources by overfishing, recovery based on an assumption of steady-state equilibrium will never be achieved, whereas for a resource experiencing a long-term increase, recovery may appear to occur sooner than it actually does.

For spatial equilibrium, impacts are ongoing as long as mean levels at impact and reference areas differ; recovery occurs when the means no longer differ (Fig. 1B). If natural factors at impact sites differ from those at reference sites, however, or if some important factors covary with exposure, it will be difficult to determine whether unequal means for impact and reference areas result from a lack of recovery or from some natural difference between the two areas being compared (i.e., a violation of the spatial equilibrium assumption).

For dynamic equilibrium, departures from a constant difference between impact and reference areas indicate ongoing impact. As the effects of exposure to the perturbation diminish, the system will return to a constant difference between impact and reference areas, signaling recovery (Fig. 1C). Dynamic equilibrium assumes that temporal dynamics at impact and reference areas are similar. This assumption may not hold, especially over long periods of time. Unlike treatments in an experimental study, the distribution of contaminants in an accidental impact is not randomized (Wiens and Parker 1995). For example, easterly and northerly facing bays in PWS were more heavily oiled than were westerly and southerly facing bays. Consequently, natural factors are likely to differ between impact and reference areas. The assumption of dynamic equilibrium requires that, in the absence of an impact, the difference between impact and reference means remains consistent over time. Operationally, assessments of dynamic equilibrium are also compromised when mean measures (e.g., abundance) have zero values. Thus, if values in either ''reference'' or ''impact'' (or both) drop to zero, the assumption of a constant difference between the areas will fail.

All three equilibrium assumptions also require the assumption that the perturbation did not push the resource past some threshold, thereby moving levels of the resource toward a new, different equilibrium. Such a scenario could occur when the impact event alters the physical environment in which the injured resource lives. Intense cleanup activities that permanently alter substrates or mixing of hazardous substances with fine substrates could provide such a push. From a restoration perspective, decades of grazing could push a resource to such a new threshold; this is the basis of various ''state-and-transition'' models (Bestelmeyer et al. 2003, 2004, Peters et al. 2004). Under these conditions, an impacted system would ''recover'' to a different state from that before the impact or that in reference locations. Recovery to different states will complicate attempts to relate conditions to the three equilibrium assumptions. The original steady state would not reoccur, spatial disequilibrium could dominate impact and reference areas, and dynamics of temporal changes between the two areas would be altered. Systems with such multiple or alternative stable states have received considerable theoretical and applied attention (e.g., Gunderson and Holling 2002). In the following examples, there is no evidence that resources were pushed to a new equilibrium state, at least on the geographic scale of sampling used to assess impacts.

Operationally, these assumptions of steady-state, spatial, and dynamic equilibrium require one to make statistical inferences about the state of recovery. Statistically significant differences (e.g., lower abundance at impact than at reference areas) suggests ongoing impact. The disappearance through time of a previously documented significant difference (i.e., a failure to reject the null hypothesis of no impact) signals recovery (Wiens 1995). Inferences about recovery are strength-

ened if there is clear evidence for an initial impact whose effects measurably decreased over time. Environmental data showing decreased contamination or increased dominance in natural factors over impact-induced factors help support a conclusion of recovery, as does a time-series of data showing stability in the recovered state. Such a hypothesis-testing approach incorporates natural variability of the systems into assessments of impact and recovery and formalizes such decisions in commonly-used statistical criteria. Of course, a null hypothesis (e.g., of no continuing impact) is not ''accepted'' simply because it is not ''rejected.'' Conclusions about ''recovery'' need to be couched in the context of type II errors (declaring recovery when it has not yet occurred) and analyses and interpretations should be conducted to minimize such errors. In the end, of course, well-planned studies and statistical analysis of recovery will always be superior to subjective decisions, as long as the results are interpreted with appropriate caution.

Examples

In the following examples from the *Exxon Valdez* oil spill, we show how investigators dealt with equilibrium assumptions using different study designs to assess the recovery of various taxa. We organize design and analytical strategies under three general categories (Wiens and Parker 1995): baseline, single year, and multiyear. We define a baseline study as one that compares pre- and post data from the impact area only. This definition is analogous to Green's (1979) Main Sequence 2, where impact is inferred from temporal change alone. Under steady-state equilibrium, as defined previously in *Ecological assumptions*, recovery occurs when pre- and post-spill means are equal. Because natural factors for most biological resources vary temporally, however, results from baseline studies are usually insufficient by themselves to assess recovery status. Single-year studies compare impact and reference areas in a single year. They rely on sampling and analytical strategies to reduce differences in natural factors among areas in order to approximate spatial equilibrium. Recovery occurs when impact and reference means are equal. Multiyear studies reduce the effects of temporal and spatial variation by subtracting out naturally varying temporal effects. If impact and reference areas are in dynamic equilibrium, recovery occurs when differences in annual means become constant; that is, trend lines in means become parallel. Table 1 compares features of these three design strategies, which we illustrate in the following examples.

Baseline: pre/post sampling at impact area only

We define a baseline design as one that compares pre- and post-impact conditions with data from the impacted area only. Under our definition, pre-impact data are opportune and limited in time series, and are unavailable for the reference areas. Further, the investigator cannot assume that the resource varies around some grand mean, especially over pre- and post-sampling periods. Under these conditions, steady-state equilibrium is not a tenable assumption and the investigator needs to be aware of the confounding effect temporal variation has on the recovery process. For planned impacts with long time series of pre and post data, Stewart-Oaten and Bence (2001) show how baseline studies using impact data only can employ covariates and time-series analysis in order to reduce and obviate, respectively, the need to assume a steady-state condition.

Baseline itself, of course, has a broad range of definitions in environmental science. By CERCLA's definition (see *Defining recovery*), baseline is actually a future condition. Green (1979) and Stewart-Oaten and Bence (2001) use the term baseline to describe studies where data may or may not be available from reference areas for the pre-impact condition. In the following example, we compare baseline results (data from impact area only) with those from BACI (before–after control–impact; Stewart-Oaten et al. 1986). BACI appears in this section only to illustrate difficulties encountered with interpreting results from baseline studies. By our definition, BACI is not a baseline design that relies on steady-state equilibrium (in contrast to Green [1979] and Stewart-Oaten and Bence [2001]); rather, BACI is a multiyear design that relies on the assumption of dynamic equilibrium.

Murphy et al. (1997) used a baseline study for assessing the recovery of 12 species of seabirds from the *Exxon Valdez* oil spill. For impacted sites, densities for post-spill years (1989, 1990, 1991) were each compared to pre-spill densities in 1984. To illustrate the difficulties with using baseline studies to assess recovery, we compare the baseline results with those of a BACI analysis (Murphy et al. 1997:Tables 1 and 2). BACI adopts features of multiyear designs and assumes dynamic equilibrium between impact and reference sites. Using BACI, the absence of significant differences between post-spill impact and reference means relative to the pre-spill difference in means signals recovery (or no impact). Fig. 2 (based on Murphy et al. 1997: Tables 1 and 2) shows changes in density for Harlequin Ducks (*Histrionicus histrionicus*), Black Oystercatchers (*Haematopus bachmani*), and Pigeon Guillemots (*Cepphus columba*) and compares baseline and BACI results, thus contrasting the inappropriate assumption of steady-state equilibrium against the more robust assumption of dynamic equilibrium. For Harlequin Ducks, the baseline analysis showed nominal negative changes, whereas BACI showed a relative nominal increase of impact over reference sites. For Black Oystercatchers, the baseline results showed the absence of significant differences for all years, whereas BACI showed a significant relative decrease in 1989 and subsequent recovery. Baseline and BACI results for Pigeon Guillemots were similar, showing negative

TABLE 1. Three design strategies for assessing recovery from environmental impacts on biological resources in temporally and spatially varying environments.

Attributes	Baseline	Single year	Multiyear	
			No reason to reject/ suspect assumptions†	Reason to reject/ suspect assumptions†
When to use	temporally invariant taxa	spatial equilibrium achievable, short recovery period	temporally variant taxa, long recovery period, taxa on multiple recovery periods, information on recovery process desired	
Data needs	pre- and post-impact only	impact and reference sites, covariates	time series for impact and reference areas or for gradient‖	
Comparison	pre- vs. post-impact	impact vs. reference, matched pairs, gradient‖	impact vs. reference and gradient over time‖	
Equilibrium assumption	steady-state	spatial	dynamic	reject or suspect assumptions
Breakdown in assumptions	temporal variation confounds with recovery	spatial variation confounds with recovery	temporal variation differs for impact and reference categories	NA
Statistical methods‡	*t* test: Student's, paired; BACI§	ANCOVA, paired *t* test, gradient‖	level-by-time, trend-by-time, repeated measures	gradient‖ (with or without covariates), impact/ref, others
Conditions needed for recovery	equal pre- and post-means	impact and reference means equal, no impact on gradient‖	difference in means constant, gradients‖ constant	failure to reject multiple assessments of impact effect
Advantages	reference sites not required (though useful)	single year of data, extrapolation reasonable	nonrandom site selection	
Disadvantages	equilibrium assumption not reasonable, pre-impact data required	recovery snapshot, covariables needed, matched sites for matched pairs	multiyear data required, difficult to extrapolate from nonrandom samples	
Comments	use with multi- or single-year studies, provides partial information on recovery process	corroborate with contamination and toxicity (triad approach)	use pre-impact data, validate assumption	verify with habitat changes, use α level > 0.05

Note: NA, not applicable.
† Reasons may include zero means. Entries that span the last two columns pertain to both situations.
‡ Methods addressed in Wiens and Parker (1995).
§ BACI uses pre-spill data at impact and reference sites and relies on the assumption of dynamic equilibrium.
‖ Gradients are dose–response regressions of biological resources vs. gradients (i.e., continuous measures) of exposure.

impacts through 1990. The absence of significant differences in results were shown for BACI in 1991, suggesting recovery, even though baseline showed continuing negative effects. Using a longer timeframe of data (1984–2001), Wiens et al. (2004) showed that the apparent temporal invariance in abundances of Harlequin Ducks and oystercatchers is short-lived, and therefore misleading in assessing recovery: Harlequin Ducks were not impacted by the spill; oystercatchers recovered from a negative impact by 1991.

Baseline studies would appear to work well only for assessing recovery for temporally invariant taxa (i.e., those in steady-state equilibrium), but they do have the advantage of requiring data only from the impact area. In addition, baseline studies provide an absolute measure of temporal variation, which can be useful in interpreting BACI results (Murphy et al. 1997).

Single-year sampling following impact: impact/reference

Because single-year studies typically compare impact and reference areas, the investigator needs to employ design and analytical strategies to control for spatial variation among the areas being compared (i.e., meet the assumption of spatial equilibrium). Following the *Exxon Valdez* spill, preliminary assessments indicated that shoreline biota were recovering faster than expected, so a study was initiated to provide a snapshot

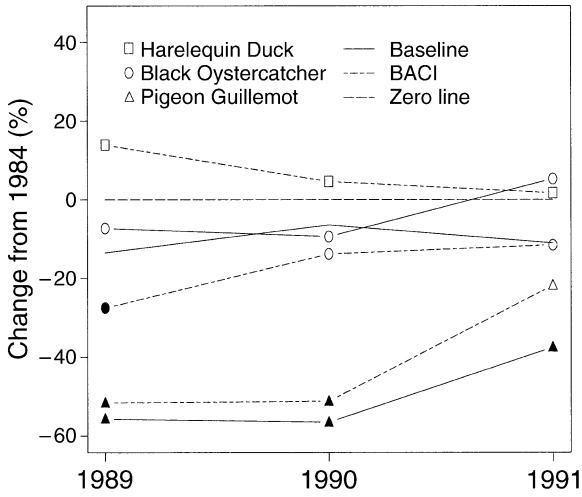

FIG. 2. A comparison of percentage changes in density from 1984 under steady-state (baseline study) and dynamic (BACI) equilibrium assumptions, based on Murphy et al. (1997: Tables 1 and 2). For Harlequin Ducks, baseline and BACI showed nominal decreases and increases, respectively. Black Oystercatchers showed a negative impact in 1989 and subsequent recovery using BACI; baseline showed a nominal decrease, but no statistically significant impact. BACI showed recovery after 1990 for Pigeon Guillemots, but continuing impact as of 1991 for baseline. Solid symbols show significant effects.

of the degree of recovery 1 year after the spill. Gilfillan et al. (1995) used an impact/reference design (Wiens and Parker 1995) to estimate the percent recovery of shoreline biota by 1990, comparing mean levels of biota at three categories of shoreline oiling (heavy, moderate, and light) with a reference (no oiling) category. Gilfillan et al. used random sampling and covariance analysis to help reduce confounding effects of spatial factors. Because oil was not randomized spatially across these factors, random sampling alone would only reduce the confounding effects of spatial variation, not eliminate them entirely as in a designed experiment (Wiens and Parker 1995). ANCOVA was used to reduce further the confounding effects of three important natural factors governing distribution and abundances of intertidal organisms. Where E(y) is the expected response (e.g., abundance, diversity, species richness),

$$E(y) = \mu + \text{wave} + \text{TOC} + \text{grain size} + \text{oiling} + \text{error}. \quad (2)$$

The grand mean is μ. Effects of wave energy (wave), total organic carbon (TOC), and sediment grain size (grain size) were removed before testing for an oiling effect, reducing the confounding effects of natural spatial variation. Gilfillan et al. found significant differences between mean species diversity in reference samples vs. moderately and heavily oiled samples from lower intertidal and −3 m zones (suggesting no recovery), but no differences for middle and upper intertidal zones (suggesting recovery or no impact) (Fig. 3). They concluded that, overall, 73% to 91% of the shoreline biota had recovered by 1990.

Wiens and Parker (1995) describe two other designs for assessing impact and recovery for data collected at a single time after an impact event: matched pairs and gradient (dose–response regressions). McDonald et al. (1995) used matched pairs to assess the impact of the *Exxon Valdez* oil spill on shorelines in PWS in 1990, using environmental factors to match impact sites with reference sites (thus implicitly accounting for covariates). Recovery was not explicitly defined by McDonald et al., but in the matched pairs design, recovery occurs when differences between impact and reference means are no longer significant. Gradient designs use dose–response regressions over a gradient of exposure, where ''dose'' is a measure of exposure and ''response'' is a measure of the biological system. For example, Wiens et al. (2004) regressed seabird density on measures of initial shoreline oiling magnitude, with and without covariates. Significant regressions infer impact and subsequent absence of significant regressions infer recovery.

The strength of single-year assessments depends on the effectiveness of the sampling and analytical strategies used to reduce the confounding effects of spatial variation, and thus approximate the assumption of spa-

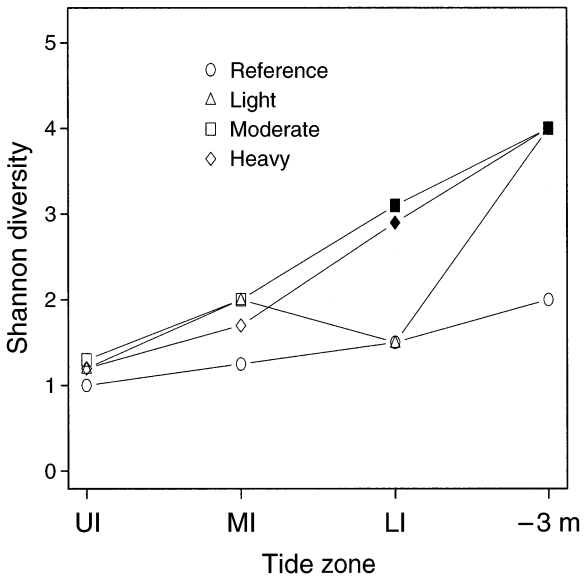

FIG. 3. Mean Shannon diversity as a function of tide zone and oiling level for pebble/gravel shorelines in Prince William Sound, Alaska, USA, in 1990. UI, MI, and LI indicate upper, middle, and lower intertidal zones; −3 m indicates 3 m below mean low tide. Solid symbols show significant differences from reference means ($\alpha = 0.05$) using ANCOVA, which reduced spatial effects of wave energy, grain size, and total organic carbon (based on Gilfillan et al. [1995: Fig. 7]). As of 1990, one year following the spill, diversity had not recovered at moderately and heavily oiled sites at lower intertidal and −3 m zones.

tial equilibrium. Using habitat criteria (strata and covariates [Gilfillan et al. 1995], habitat criteria for matched pairs [McDonald et al. 1995], randomly sampling across factors for gradients [Wiens et al. 2004]) presumes that the investigator has specific knowledge about which factors govern the distribution and abundance of the taxa being studied. In the case of shoreline biota, these factors are many and of different levels of influence (some known and measurable, others not). In such cases, measurable overarching factors such as wave energy work best, especially when the investigator is concerned with impacts on many different organisms and communities. Hence, for single-year designs, knowledge of governing spatial factors is very important. Failing to account for important spatial factors will lead one to confound spatial variation with recovery and weaken results: the investigator will be uncertain if recovery has or has not occurred.

Single-year studies, however, have an advantage over baseline and multiyear studies in that only one year of post-impact data is required. Further, depending on the sampling design, single-year study results based on random samples of impact and reference areas can be extrapolated to the entire spill area (Page et al. 1995). Because they provide only a snapshot of recovery, however, single-year studies will not reveal when full recovery occurs, unless the study is delayed for a sufficient (and, in the absence of any sampling, unknown) length of time. In the case of the *Exxon Valdez* spill this would have been approximately five years after the spill for shoreline biota (Skalski et al. 2001).

Multiyear sampling following impact: level-by-time interactions

Multiyear designs address the assumption of dynamic equilibrium and are useful for temporally variant taxa, especially for those recovering at different rates. Skalski et al. (2001) used a multiyear design to assess recovery of oiled shorelines by testing for equal trends in means over years at impact and reference sites using a ''parallelism'' design (level-by-time [Wiens and Parker 1995]; analogous to optimal design in Green [1979]). Recovery in abundance at impact sites was evidenced when annual changes in abundances at impact and reference locations paralleled one another (i.e., lines connecting annual means were parallel). For example, means of *Littorina sitkana* abundance for washed (initially oiled and then cleaned with pressurized warm water) and reference sites in 1989 and eight additional post-spill years showed a general decline (Fig. 4), presumably due to natural temporal variation or regional environmental changes. Snail abundances were disproportionately reduced at washed sites in 1989–1991, indicating impacts; after 1991, annual changes in abundances in the washed area tracked those at the reference area, suggesting recovery.

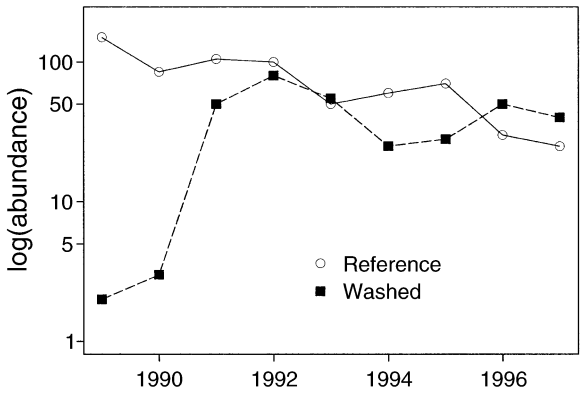

FIG. 4. Recovery of *Littorina sitkana* from the *Exxon Valdez* spill occurred after 1991. Recovery was measured as number of *Littorina*/m^2. Following recovery, parallel profiles in abundance at reference and oiled and washed sites showed evidence for dynamic equilibrium between these two classes of sites (based on Skalski et al. [2001: Fig. 7B]). Using similar analyses on other species and taxonomic groups, Skalski et al. concluded that oiled shorelines recovered between 1992 and 1994.

Skalski et al. (2001) tested for level-by-time interaction between reference and washed means (type) with an ANOVA:

$$E(y) = \mu + \text{type} + \text{year} + \text{type} \times \text{year} + \text{error} \quad (3)$$

where E(y) is the expected abundance of, say, *Littorina* snails, type is either oiled/washed or reference, and μ is the grand mean over type and year. The interaction term of type $\times$ year tests for a constant difference between washed and reference means over years, i.e., for parallel trends in means over time. Skalski et al. used a backward sequential procedure to identify periods of recovery and impact. Starting with the last year sampled (1997), the interaction term was tested with a 3-yr time window. If the null hypothesis was not rejected, the window was moved back one year and the test for interaction was repeated. This back-step procedure was repeated until the interaction term was significant, indicating the final year of impact. Tests on main effects were not useful for assessing recovery because significant effects of year and type would show differences only in mean abundance over time and between impact/washed means, respectively; both could occur due to natural spatial and temporal variation.

Trend-by-time designs (Wiens and Parker 1995) also rely on the assumption of dynamic equilibrium. In a trend-by-time design, dose–response regressions of biota on a continuous measure of impact are compared over years following the impact event; the disappearance of a dose–response relationship over time signals recovery. For example, Day et al. (1995) regressed seabird density on an index of initial oiling to assess impact and recovery of marine-oriented birds following the spill. In another study, Lance et al. (2001) regressed logs of seabird density against year for six years

between 1989 and 1998 and compared slopes of the regressions using a homogeneity of slopes test (i.e., parallelism).

In both parallelism and trend-by-time designs, the assumption of dynamic equilibrium requires that the effects of natural temporal variation be similar among the site categories being compared. Sampling sites therefore need to be geographically close enough to experience the same changes in short- and long-term climactic and oceanographic variability. Even for sites that seemingly experience similar temporal variations, local environmental changes may affect sites differently due to complex interactions of spatial and environmental factors. For example, in a 10-yr study of shoreline organisms, Gilfillan and Parker (2003) showed that dynamic equilibrium broke down for some shoreline organisms. Between 1998 and 1999, changes in water temperature differed between impact and reference areas in ways not seen for other years, differentially affecting shoreline organisms at oiled and reference areas and upsetting the common pattern of dynamic equilibrium. When the assumption of dynamic equilibrium fails, there is no reason to expect yearly trends in means to track (parallel) each other. Over the long term, the sites may fall back into dynamic equilibrium, but during the intervening interval, recovery may be falsely concluded to have or have not occurred. Sampling after recovery occurred was necessary for the analytical strategy employed by Skalski et al. (2001). In addition, sampling for several years after recovery helped to assess the assumption of equal temporal dynamics at impact and reference areas (Skalski et al. 2001, Wiens et al. 2004).

Multiyear designs have several advantages over baseline and single-year studies. Most importantly, the assumption of dynamic equilibrium is more realistic than either temporal (baseline) or spatial (single-year) equilibrium. In addition, sites do not need to be randomly selected, although this means that it will be difficult to extrapolate from the nonrandomly selected sites to a larger area.

Uncertain equilibrium assumptions: weight-of-evidence approaches

Sometimes equilibrium assumptions for multiyear data may be uncertain. Erratic non-parallel trends in impact and reference means would trigger concerns for uncertainty. Dramatic region-wide declines in abundance or shifts in geographic centers of abundance would also call into question equilibrium assumptions, as would zero mean abundances for some years (discussed later in *Guidance: Multiyear studies*). Wiens et al. (2004) encountered all these elements of uncertainty and used a weight-of-evidence approach to assess recovery of 25 seabird species over a 12-yr period from 1989 to 2001. Wiens et al. synthesized results from four separate analyses, taking into account effects of potentially confounding habitat variables. The four analyses treated ecological assumptions differently. (1) For each year, densities were regressed on an index of shoreline oiling for 10 bays. It was assumed that the 10 bays were in spatial equilibrium for each year, although the use of quadratic regression relaxed this assumption somewhat. (2) Covariates of important habitat variables, selected by AIC (Akaike Information Criterion, Burnham and Anderson 2002), were used in the same regression analyses as in (1). Covariates reduced the potentially confounding effects of spatial nonequilibrium due to environmental differences among the bays by subtracting out covariate effects before testing for continuing effects of oiling. (3) Densities from 1984 were used in a repeated-measures design, which assumed dynamic equilibrium in natural factors among categories of moderately-to-heavily and none-to-lightly oiled bays. (4) Plots of trends in density by oiling category were visually examined. In interpreting plots, either steady-state or dynamic equilibrium was assumed, depending on whether one considered oiling categories separately over time or relative to each other. Using a weight-of-evidence approach based on all of these analyses strengthened the possibility of finding a continuing effect of oiling under different assumptions.

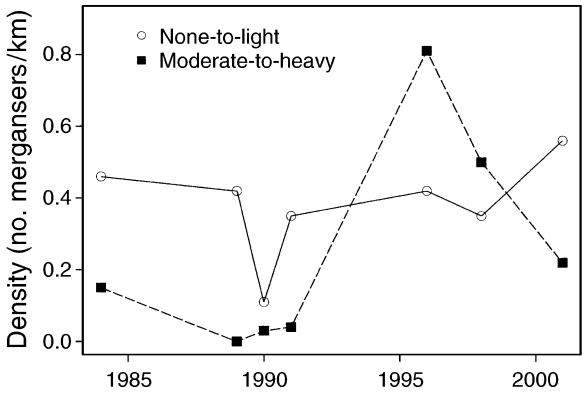

FIG. 5. Mean density of Common Mergansers for none-to-lightly oiled and moderate-to-heavily oiled bays in Prince William Sound. Using a weight of evidence approach consisting of four analyses, Wiens et al. (2004) concluded that mergansers recovered by 1996.

Common Mergansers (*Mergus merganser*) provide an example of how weight of evidence was used to detect initial spill-caused impacts, followed by recovery. Regressions indicated negative relationships with oiling level that remained after habitat factors were included as covariates in the analysis (Table 1 in Wiens et al. 2004). BACI comparisons with the 1984 baseline yielded weak, nonsignificant indications of reduced abundance in oiled bays in 1989 and 1991 (−47.4% and −28.9%, respectively; Wiens et al. 2004:Table 2). The oiled-unoiled abundance plot (Fig. 5) showed that numbers were low in oiled bays in 1984–1991 and that many more mergansers were seen in previously oiled

bays in 1996–2001. The oiling-gradient plots (Wiens et al. 2004:Fig. 4), which include density values for individual bays, clearly illustrated this shift in distribution over time. Wiens et al. concluded that merganser habitat use was negatively affected by the spill but had recovered by 1996.

Guidance

Given the various ways in which assumptions (or their violations) can affect assessments of recovery, it is important to consider how to design studies and conduct analyses to reduce the confounding effects of temporal and spatial variation (Table 1).

Baseline

Designs that rely solely on an assumption of steady-state equilibrium (e.g., baseline and pre/post pairs at impact sites) are inadequate for assessing recovery of biological resources in temporally varying environments. Recovery may be confused with temporal variations, leading to false conclusions. Used in conjunction with other multi- and single-year designs, however, baseline data may provide some insight into recovery processes. Murphy et al. (1997) used a baseline design for assessing recovery of seabirds, but only in consort with BACI, a multiyear design that relies on the more realistic assumption of dynamic equilibrium. In this case, the baseline studies were used more to gain insight into long-term trends in populations than to determine whether or not recovery occurred. Because of the inadequacy of baseline designs to assess impact and recovery, however, additional data beyond the pre-impact data will be needed. Moreover, baseline studies that rely on fortuitously available data from nonrandomly selected sites can be used to extrapolate results to the total impacted area only with great care. For populations that have been severely impacted beyond a normal range of variation (i.e., reduced to zero by obvious impacts), baseline studies may be useful in identifying the initial stages of recovery. Supplementary data on long-term trends in abundance from locations outside of the impact zone, as well as on climate and oceanographic conditions, may be helpful in estimating when recovery occurs in baseline studies.

Single-year studies

An impact/reference design is valid when sampling efforts are spread over a broad geographic area and conducting multi-year sampling is not feasible. Recovery is identified by the absence of significant differences between impact and reference areas; if sampling sites are randomly selected, results of impact/reference studies can be extrapolated to the total area impacted. Single-year studies require acceptance of the assumption of spatial equilibrium, either through covariance analysis (Gilfillan et al. 1995) or by pairing sites by shared environmental factors (McDonald et al. 1995). Regressing measures of biota on a continuous measure of impact (gradient analysis; Wiens and Parker 1995) can also be used. Associated environmental variables should be sampled concurrently with the biological resources of interest; however, some of the ecologically important variables may be unknown and/or unmeasurable, and different variables may be needed when multiple species are assessed. Concomitant variables also consume degrees of freedom, resulting in a loss in statistical power.

When designing single-year studies, it is also prudent to consider the effects of spatial scale. Contaminants (e.g., spilled oil) do not distribute randomly across habitats, and reference samples are often drawn from a larger area than are the impact samples. Differing spatial scales can result in differing scales of environmental variation, further confounding tests of equality in resource levels between the contaminated and reference areas. Although the spatial scale of the contaminated area in PWS was less than that of the reference area, there was sufficient length of oiled and reference shorelines to randomly sample similar habitats in both areas (Gilfillan et al. 1995, Page et al. 1995). Hence, an equal number of samples at environmentally similar sites were taken in oiled and reference areas, reducing the potentially confounding influences of differing scales. Covariates (Gilfillan et al. 1995) also helped reduce potential differences in variation between the two areas.

Of course, a single-year study provides only a snapshot of recovery and, in the absence of corroborative studies, conclusions are inevitably speculative. The choice of when to conduct a single-year study affects conclusions since different species will have different times to recovery. The single-year study of Gilfillan et al. (1995) on shoreline biota was part of a sediment-quality triad approach (Long and Chapman 1985) where conclusions on recovery were based on concurrent sampling of biota, sediment chemistry and toxicity (Page et al. 1995). Boehm et al. (1995) concluded that by 1990–1991, sediment oil concentrations were low and acute sediment toxicity was virtually absent except at a small number of isolated locations, corroborating the conclusions of Gilfillan et al. In addition, the single-year study of Gilfillan et al. was part of 10-yr study to assess spill impacts at heavily oiled sites. (Gilfillan and Parker 2003).

Multiyear studies

Multiyear studies are superior to single-year and baseline studies because they can reduce the confounding effects of natural variation and provide an opportunity to observe the recovery process and evaluate the assumption of dynamic equilibrium. Multiyear studies are necessary for taxa that exhibit large temporal variations or that have long recovery periods or for multiple taxa that differ in their temporal dynamics. If environmental factors change similarly over time at impact and reference areas (i.e., dynamic equilibrium

holds), assessing parallel trends in means has the advantage of requiring sampling data only on the resource being assessed; no measures of concomitant environmental variables are needed. BACI (Stewart-Oaten et al. 1986) is a good example of a multi-year analysis that relies on the assumption of dynamic equilibrium. Green (1979) also presents material on a multi-year BACI-type design.

The assumption of dynamic equilibrium in factors at impact and reference areas should be verified by assessing the degree to which environmental factors really are similar in the areas being compared. For example, embayed rocky shorelines may be comparable with each other, but not with exposed sandy habitats. Impact and reference areas should also experience the same regional changes in climate and oceanographic conditions; similarity in temporal dynamics can be evaluated with regression analyses and tests of additivity, given sufficient data. The disadvantage of a multivariate approach is that samples need to be collected during time-limited sampling seasons over fairly long time periods (e.g., Skalski et al. [2001], nine years; Gilfillan and Parker [2003], 10 years; Wiens et al. [2004], 12 years).

Both level-by-time (parallelism) and trend-by-time designs are useful statistical methods for assessing multi-year data on recovery (Wiens and Parker 1995). The level-by-time design compares year-to-year changes in means (e.g., abundance) at impact and reference areas. After recovery occurs, the difference in means remains constant (Skalski et al. 2001). The trend-by-time design compares dose–response regressions between measures of biota and a continuous measure of impact. After recovery, dose–response regressions remain constant. Level-by-time and trend-by-time designs work well for abundant species or for multi-species metrics (e.g., species richness, indexes of diversity) where few zeros occur. Level-by-time breaks down when zero means occur. Trend-by-time breaks down for messy dose–response relationships, usually due either to zero levels of the resource or to complex nonlinear relationships between dose and response under the null hypothesis of no impact. Both designs can be extended beyond apparent recovery to assess the assumption of dynamic equilibrium.

There are disadvantages to using multi-year studies to assess recovery. Because multi-year data are needed, consistent sampling methods must be used over time and the studies will be more costly and difficult to mange than will be single-year or baseline studies. For multi-year studies some redundancy in sampling stations should be considered because stations may be lost or affected by unexpected events over a long time period. If sampling sites are not randomly selected, results from multiyear studies may be difficult to extrapolate to an impact-wide area.

When available, pre-impact (i.e., baseline) data are useful in understanding the pre-impact relationships between impact and reference areas (Murphy et al. 1997, Wiens et al. 2004). BACI is a special case of level-by-time that incorporates pre-impact data and is useful for year-to-year comparisons. Repeated-measures analysis (Skalski and Robson 1992, Wiens et al. 2004) accounts for temporal dependence in observations and is the preferred way to analyze time-series data at impact and reference areas, given sufficient data (Skalski and Robson 1992, Skalski et al. 2001). However, caution should be exercised when interpreting results from multivariate repeated-measures analyses, since repeated-measures tests are based on randomization, and randomization can only be approximated with random sampling in observational studies.

Finally, a weight-of-evidence approach is useful for assessing multi-year data for biological resources that do not meet the assumption of dynamic equilibrium. In this case, analytical strategies that are robust to the natural variability of the data should be used. For example, Wiens et al. (2004) used analytical strategies that relied on dynamic and spatial equilibrium, covariance analysis to remove the effects of spatial variation among impact categories, and examination of figures and tables of abundance over time. They used a high α level (0.20) to reduce the decision error of falsely declaring recovery and to favor evidence of ongoing impacts (Wiens et al. 2004).

Other designs

Our examples and citations provide sampling designs and analytical strategies that have demonstrated their usefulness for assessing recovery from environmental accidents in temporally and spatially varying environments. Which design to apply in order to reduce the confounding effects of varying natural factors depends on several issues, including the availability of pre-impact data, the spatial scale of the accident, and the financial and personnel resources needed to address the requirements of CERCLA. In addition to variations on these designs, other designs (e.g., time-series analysis, reference as covariate, intervention analysis; Stewart-Oaten and Bence 2001) may prove equally useful or better depending on the nature of the impact and the reasonableness of the ecological assumptions one makes. Stewart-Oaten and Bence provide a useful statistical and environmental framework for choosing sampling designs in order to assess impact and recovery.

In addition to standard tests of hypothesis, assessments of recovery can be performed with at least three other statistical methods: confidence intervals, decision theory, and Bayesian methods. Confidence intervals provide useful information on how confident one can be of the true state of nature being close to the hypothesized null state. Thus, confidence intervals incorporate information on variability which can be useful for judging the strength of evidence for making decisions on recovery. More formally, if a confidence

interval does not cover the hypothesized parameter values then the value is refuted by the observed data, much in the same way as in a test of hypothesis. Hoenig and Heisey (2001) conclude that confidence intervals have no advantage over traditional methods for testing hypotheses and making decisions based on type I error. Decision theory (DeGroot 2004) allows the investigator to adjust type I and II error based on penalties for wrong decisions. Fisheries and wildlife managers use decision theory, where associated economic and social penalties (e.g., suboptimal fishery yields) are quantifiable. Decision theory would be practical for assessing recovery given penalties for making the wrong decision were quantifiable. In the absence of known or estimable economic, social, and/or ecological value, it is difficult to objectively quantify penalties. Bayesian statistics (Box and Tiao 1973) rely on prior probabilities (for the truth of the null hypothesis). Each dependent variable would likely need its own prior, making it difficult to use Bayesian methods for multiple species, e.g., 25 species of seabirds. Such priors and their influence on results would be difficult to justify, especially for such high-profile assessments as the *Exxon Valdez* spill. In addition to these three methods, nonparametric, computer-intensive methods can also be used to test hypotheses. Bootstrap, jackknife, permutation tests, and other simulation-based tests of hypothesis provide ways to avoid making distributional assumptions (typically for normality) on the residual error. In the examples cited previously, Gilfillan et al. (1995, shoreline ecology) and Wiens et al. (2004, seabirds) found generalized linear models (McCullagh and Nelder 1989) on members of the exponential family (e.g., negative binomial and Poisson) to be more realistic of natural conditions than computer-intensive nonparametric methods.

Conclusions

Recovery from an environmental accident involves re-establishment of the physical environment to a nonimpacted state and restoration of complex biological systems. Recovery takes time. The greater the injury and complexity of the system, the longer the time to recovery. Natural environmental variation and the passage of time enlarge the probability field of what a nonimpacted resource would have been had the injury not occurred. Consequently, the longer the time to recovery, the less likely the recovered state will be what it was before the injury occurred. Moreover, as Pimm (1991) has noted, population variability increases with the length of time considered, increasing the likelihood that one or another of the equilibrium assumptions will be violated. Defining the recovered state of a biological resource and inferring when that state has been reached depend on what one can reasonably assume about natural variation of *both* populations and the environments they occupy.

Studies conducted after the *Exxon Valdez* oil spill provide practical examples of the appropriateness (or inappropriateness) of assumptions about equilibrium and natural variation for assessing impact and recovery of biological resources. Steady-state equilibrium is not applicable to PWS, at least for resources that require a year or more to recover. High-latitude extremes in climate as well as energetics of wind, waves, floating ice and debris can annually reset the dynamics of shoreline and nearshore habitats and the fish, birds and mammals that use those habitats. On the scales of measurement needed to assess the relationships between organisms and declining oil contamination over time, variability in natural factors governing abundance is likely to overshadow long-term recovery under the assumption of steady-state equilibrium.

On its own, the assumption of spatial equilibrium does not apply in PWS either. Physical factors and the biological resources they affect are heterogeneously distributed over multiple scales, and even superficially similar areas are likely to differ in details that are important to the resources being assessed. The assumption of dynamic equilibrium, however, may be more realistic, at least at some scales. Regional changes in climate and oceanographic conditions appear to affect many areas of the Sound similarly, perhaps because it is a semi-enclosed body of water that remains ice-free in the winter. At the finer spatial scales of individual embayments or different habitats such as soft sediment vs. exposed bedrock shorelines, the assumption that variations among locations are temporally concordant would be more likely to be violated.

The appropriateness or inappropriateness of the different assumptions about equilibrium is scale-dependent. This means that there is no single ''best'' study design or statistical analysis that will fit all situations. The risks of reaching conclusions based on incorrect assumptions about equilibrium are likely to be different at different scales of investigation, and the scales in turn depend on the species, communities, or environments being considered, the temporal and spatial scales of their natural dynamics, and the scale of the environmental accident itself. The scale of the impact is usually well-defined, whereas scales of interdependencies among biological resources and their environments and of natural dynamics are generally unknown, at least with any degree of precision.

The burden is therefore on the investigator to use sound judgment in assessing the state of recovery of an injured biological resource. The underlying assumptions about equilibrium must be explicitly recognized and addressed. And, just as in restoration efforts, there should be a clear a priori statement of what represents an acceptable level or end point for recovery. Given natural variation in the environment and the resources of interest, this means that one must consider the width of the envelope of natural variation, which

in turn helps to define the appropriate ''recovery zone'' for the system.

Assessments of recovery from environmental perturbations are often carried out in a contentious atmosphere where the assessor's decisions will come under intense scrutiny. Decisions must be made about how to design a study and analysis to minimize the confounding influences of sampling and natural variation, and to do this with regard to the natural history of the resources being assessed, physical features of the environment, short- and long-term climactic conditions, habitat use, and the level of certainty required. These decisions will have direct consequences on effort, cost, and the persuasiveness and certainty of the conclusions. It is critically important that one give honest and explicit consideration to the equilibrium assumptions and which (if any) are likely to apply, and to recognize the consequences of violating these assumptions in designing and carrying out studies, analyzing data, and interpreting results.

Acknowledgments

Evaluating the biological consequences of the *Exxon Valdez* oil spill prompted our thinking on issues of assessing recovery from environmental accidents. We thank R. Day and S. Murphy (ABR, Inc.), and P. Kareiva (The Nature Conservancy). ExxonMobil provided funding for our activities. The views expressed here are our own.

Literature Cited

Bestelmeyer, B. T., J. R. Brown, K. M. Havstad, R. Alexander, G. Chavez, and J. E. Herrick. 2003. Development and use of state-and-transition models for rangelands. Journal of Range Management **56**:114–126.

Bestelmeyer, B. T., J. E. Herrick, J. R. Brown, D. A. Trujillo, and K. M. Havstad. 2004. Land management in the American Southwest: a state-and-transition approach to ecosystem complexity. Environmental Management **34**:38–51.

Boehm, P. D., D. S. Page, J. S. Brown, J. S. Neff, and W. A. Burns. 2004. Polycyclic aromatic hydrocarbon levels in mussels from Prince William Sound, Alaska, U. S. A. Environmental Toxicology and Chemistry **23**:2916–2929.

Boehm, P. D., D. S. Page, E. S. Gilfillan, W. A. Stubblefield, and E. J. Harner. 1995. Shoreline ecology program for the *Exxon Valdez* oil spill. Part 2—chemistry and toxicology. Pages 347–397 *in* P. G. Wells, J. N. Butler, and J. S. Hughes, editors. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219, American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Box, G. E. P., and G. C. Tiao. 1973. Bayesian inference in statistical analysis. John Wiley & Sons, New York, USA.

Bradshaw, A. D. 2002. Introduction and philosophy. Pages 3–9 *in* M. R. Perrow and A. J. Davy, editors. Handbook of ecological restoration. Volume 1. Principles of restoration. Cambridge University Press, Cambridge, UK.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Chesson, P. L., and T. J. Case. 1986. Overview: nonequilibrium community theories: chance, variability, history, and coexistence. Pages 229–239 *in* J. Diamond and T. J. Case, editors. Community ecology. Harper and Row, New York, New York, USA.

Day, R. H., S. M. Murphy, J. A. Wiens, G. D. Hayward, E. J. Harner, and L. N. Smith. 1995. Use of oil-affected habitats by birds after the *Exxon Valdez* oil spill. Pages 727–761 *in* P. G. Wells, J. N. Butler, and J. S. Hughes, editors. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

DeGroot, M. H. 2004. Optimal statistical decisions. John Wiley and Sons, New York, New York, USA.

Exxon Valdez Oil Spill Trustee Council. 2002. 2002 status report. *Exxon Valdez* Oil Spill Trustee Council, Anchorage, Alaska, USA.

Gilfillan, E. S., D. S. Page, E. J. Harner, and P. D. Boehm. 1995. Shoreline ecology program for the *Exxon Valdez* oil spill. Part 3—biology. Pages 398–443 *in* P. G. Wells, J. N. Butler, and J. S. Hughes, editors. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Gilfillan, E. S., and K. R. Parker. 2003. Multivariate analysis of community structure over ten years following the *Exxon Valdez* oil spill. Pages 559–567 *in* Proceedings of the 2003 Oil Spill Conference. American Petroleum Institute, Washington, D.C., USA.

Giller, P. S., and J. R. Gee. 1987. The analysis of community organization: the influence of equilibrium, scale and terminology. Pages 519–542 *in* J. H. R. Gee and P. S. Giller, editors. Organization of communities past and present. Blackwell Scientific Publications, Oxford, UK.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. Wiley-Interscience, New York, New York, USA.

Gunderson, L. H., and C. S. Holling, editors. 2002. Panarchy: understanding transformations in human and natural systems. Island Press, Washington, D.C., USA. H., and C. S. Holling, editors. 2002. Panarchy: understanding transformations in human and natural systems. Island Press, Washington, D.C., USA.

Hare, S. R., and N. J. Mantua. 2000. Empirical evidence for North Pacific regime shifts in 1977 and 1989. Progress in Oceanography **47**:103–146.

Hobbs, R. J. 2004. Restoration ecology: the challenge of social values and expectations. Frontiers in Ecology **2**:43–44.

Hoenig, J. M., and D. M. Heisey. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. American Statistician **55**(1):19–24.

Irons, D. B., S. J. Kendall, W. P. Erickson, L. L. McDonald, and B. K. Lance. 2000. Nine years after the *Exxon Valdez* oil spill: effects on marine bird populations in Prince William Sound, Alaska. Condor **102**:723–737.

Kingsland, S. E. 1985. Modeling nature. University of Chicago Press, Chicago, Illinois, USA.

Lance, B. K., D. B. Irons, S. J. Kendall, and L. L. McDonald. 2001. An evaluation of marine bird population trends following the *Exxon* Valdez oil spill, Prince William Sound, Alaska. Marine Pollution Bulletin **42**:298–308.

Long, E. R., and P. M. Chapman. 1985. A sediment quality triad: measures of sediment contamination, toxicity, and infaunal community composition in Puget Sound. Marine Pollution Bulletin **16**:405–415.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Second edition. Chapman and Hall, New York, New York, USA.

McDonald, L. L., W. P. Erickson, and M. D. Strickland. 1995. Survey design, statistical analysis, and basis for inferences in coastal habitat injury assessment: *Exxon Valdez* Oil Spill. Pages 296–311 *in* P. G. Wells, J. N. Butler, and J. S. Hughes, editors. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Murphy, S. M., R. H. Day, J. A. Wiens, and K. R. Parker. 1997. Effects of the *Exxon Valdez* oil spill on birds: comparisons of pre- and post-spill surveys in Prince William Sound, Alaska. Condor **99**:299–313.

National Research Council. 1992. Restoration of aquatic ecosystems: science, technology and public policy. National Academy Press, Washington, D.C., USA.

Neff, J. M., E. H. Owens, S. W. Stoker, and D. M. McCormick. 1995. Shoreline oiling conditions in Prince William Sound follow ing the *Exxon Valdez* oil spill. Pages 312–346 *in* P. G. Wells, J. N. Butler, and J. S. Hughes, editors. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Page, D. S., E. S. Gilfillan, P. D. Boehm, and E. J. Harner. 1995. Shoreline ecology program for the *Exxon Valdez* oil spill. Part 1—study design. Pages 347–397 *in* P. G. Wells, J. N. Butler, and J. S. Hughes, editors. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Peters, D. P. C., R. A. Pielke, Sr., B. T. Bestelmeyer, C. D. Allen, S. Munson-McGee, and K. M. Havstad. 2004. Cross-scale interactions, nonlinearities, and forecasting catastrophic events. Proceedings of the National Academy of Sciences (USA) **101**:15130–15135.

Peterson, C. H. 2001. The ''*Exxon Valdez*'' oil spill in Alaska: acute, indirect and chronic effects on the ecosystem. Advances in Marine Biology **39**:1–103.

Peterson, C. H., S. D. Rice, J. W. Short, D. Esler, J. L. Bodkin, B. E. Ballachey, and D. B. Irons. 2004. Long-term ecosystem response to the Exxon Valdez oil spill. Science **302**: 2082–2086.

Peterson, W. T., and F. B. Schwing. 2003. A new climate regime in northeast pacific ecosystems. Geophysical Research Letters 30(17), 1896, doi: 10.1029/2003GL017528, 2003.

Pimm, S. L. 1991. The balance of nature? Ecological issues in the conservation of species and communities. University of Chicago Press, Chicago, Illinois, USA.

Rice, S. D., R. B. Spies, D. A. Wolfe, and B. A. Wright. 1996. Proceedings of the *Exxon Valdez* oil spill symposium. Symposium No. 18. American Fisheries Society, Bethesda, Maryland, USA.

Short, J. W., M. R. Linderberg, P. M. Harris, J. M. Maselko, J. J. Pella, and S. D. Rice. 2004. Estimate of oil persisting on the beaches of Prince William Sound 12 years after the *Exxon Valdez* oil spill. Environmental Science Technology **38**:19–25.

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.

Skalski, J. R., and D. S. Robson. 1992. Techniques for wildlife investigations. Academic Press, San Diego, California, USA.

Society for Ecological Restoration. 1996. Ecological restoration: definition. ⟨http://www.ser.org⟩

Stewart-Oaten, A., and J. R. Bence. 2001. Temporal and spatial variation in environmental impact assessment. Ecological Monographs **71**:305–339.

Stewart-Oaten, A., W. W. Murdoch, and K. R. Parker. 1986. Environmental impact assessment: ''pseudoreplication'' in time? Ecology **67**:929–940.

U.S. Code of Federal Regulations. 2001. Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), revised 2001. U.S. Code of Federal Regulations. Volume 43. Sections 11:14 and 11:72.

Wells, P. G., J. N. Butler, and J. S. Hughes, editors. 1995. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Wheelwright, J. 1994. Degrees of disaster. Prince William Sound: how nature reels and rebounds. Simon and Schuster, New York, New York, USA.

Wiens, J. A. 1977. On competition and variable environments. American Scientist **65**:590–597.

Wiens, J. A. 1984. On understanding a non-equilibrium world: myth and reality in community patterns and processes. Pages 439–457 *in* D. R. Strong, D. Simberloff, L. G. Abele, and A. B. Thistle, editors. Ecological communities: conceptual issues and the evidence. Princeton University Press, Princeton, New Jersey, USA.

Wiens, J. A. 1995. Recovery of seabirds following the *Exxon Valdez* oil spill: an overview. Pages 854–893 *in* P. G. Wells, J. N. Butler, and J. S. Hughes, editors. *Exxon Valdez* oil spill: fate and effects in Alaskan waters. STP 1219. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Wiens, J. A., R. H. Day, S. M. Murphy, and K. R. Parker. 2001. On drawing conclusions nine years after the *Exxon Valdez* oil spill. Condor **103**:886–892.

Wiens, J. A., R. H. Day, S. M. Murphy, and K. R. Parker. 2004. Changing habitat and habitat use by birds after the *Exxon Valdez* oil spill, 1989–2001. Ecological Applications **14**:1806–1825.

Wiens, J. A., and K. R. Parker. 1995. Analyzing the effects of accidental environmental impacts: approaches and assumptions. Ecological Applications **5**:1069–1083.

Worster, D. 1977. Nature's economy: the roots of ecology. Sierra Club Books, San Francisco, California, USA.