Annual Review of Marine Science

Designing More Informative Multiple-Driver Experiments

Mridul K. Thomas[1,*] and Ravi Ranjan[2,3,4,*]

[1]Department F.-A. Forel for Environmental and Aquatic Sciences and Institute for Environmental Sciences, University of Geneva, Geneva, Switzerland; email: mridul.thomas@unige.ch

[2]Helmholtz Institute for Functional Marine Biodiversity at the University of Oldenburg, Oldenburg, Germany; email: ravi.ranjan@hifmb.de

[3]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany

[4]Hanse-Wissenschaftskolleg Institute for Advanced Study, Delmenhorst, Germany

First published as a Review in Advance on August 25, 2023

The *Annual Review of Marine Science* is online at marine.annualreviews.org

https://doi.org/10.1146/annurev-marine-041823-095913

*These authors contributed equally to this article

Keywords

multiple stressors, experimental design, interactions, anthropogenic change, theory–experiment integration, predictive ecology

Abstract

For decades, multiple-driver/stressor research has examined interactions among drivers that will undergo large changes in the future: temperature, pH, nutrients, oxygen, pathogens, and more. However, the most commonly used experimental designs—present-versus-future and ANOVA—fail to contribute to general understanding or predictive power. Linking experimental design to process-based mathematical models would help us predict how ecosystems will behave in novel environmental conditions. We review a range of experimental designs and assess the best experimental path toward a predictive ecology. Full factorial response surface, fractional factorial, quadratic response surface, custom, space-filling, and especially optimal and sequential/adaptive designs can help us achieve more valuable scientific goals. Experiments using these designs are challenging to perform with long-lived organisms or at the community and ecosystem levels. But they remain our most promising path toward linking experiments and theory in multiple-driver research and making accurate, useful predictions.

INTRODUCTION

Why Do We Do Multiple-Driver/Stressor Experiments?

Driver: an environmental dimension of interest (e.g., temperature or pH); here, the term driver is synonymous with the terms stressor and factor (note that factor here does not imply a categorical variable)

The planet is undergoing environmental changes that are unprecedented in recorded history. While we have had periods of extremely large environmental change over longer timescales (the ends of the ice ages, the Paleocene–Eocene Thermal Maximum, etc.), the speed of present historical change and the wide range of dimensions in which change is happening (temperature, CO_2, pH, nitrogen, biodiversity, fishing pressure, etc.) make the present era a human-driven historical anomaly (Halpern et al. 2019, IPCC 2021). Faced with the prospect of some of these changes continuing for centuries and little likelihood of returning to preindustrial conditions for centuries or even millennia, scientists need to understand how our ecosystems are changing and use this understanding to reliably predict what our planet's future holds. If we are to shape policy in a way that mitigates harm (and takes advantage of benefits) to ourselves and to nature, we need quantitative predictions of how populations, communities and ecosystems will respond to—and themselves alter—their future environments (Currie et al. 2004, Dietze 2017, Houlahan et al. 2017, Pennekamp et al. 2017, Pottier et al. 2013, Shaw 2019).

What Forms of Drivers/Stressors and Organisms Do We Focus on Here?

We focus on abiotic drivers such as temperature, nutrients, water availability, light, pH, and CO_2. For simplicity, the experiments and models we discuss are largely at the population level (single species). Our recommendations apply most strongly to short-lived organisms for which experiments measuring vital population rates across environmental gradients are feasible and ethically uncontentious. They can also be applied to longer-lived organisms, but the experiments will be more challenging to perform, especially if they involve trophic interactions and population dynamics. These principles and designs can be used at the community and ecosystem levels as well. For some questions—such as the environment dependence of productivity or respiration—the experiments may be even simpler than those we describe here. But for simple multispecies communities, some questions may be more challenging to answer, because biotic interactions can involve discontinuities that are challenging for statistical and mathematical methods that rely on smooth, differentiable functions. Given that there is plenty of low-hanging fruit, we do not attempt to address these challenges here.

What Do We Need to Predict, at What Scale?

A nonexhaustive list of properties that biologists, biogeochemists, and ecosystem scientists are called upon to predict for the future includes (*a*) elemental cycles of carbon, nitrogen, phosphorus, and more; (*b*) primary productivity at global and regional scales; (*c*) biomass at global and regional scales, as well as its spatial distribution and variation across trophic levels and size categories; (*d*) biodiversity in its various forms (taxonomic, functional, and genetic) at local and regional scales; (*e*) species ranges, population sizes and densities, and rates of reproduction of taxa that are charismatic (e.g., whales) or have large impacts on communities and ecosystems (nitrogen fixers, pathogens, and keystone species such as beavers and sea otters); (*f*) phenological changes, especially in species that are important to other taxa (as flowering plants are to pollinators, and primary producers are to consumers); and (*g*) the vulnerability of different regions to environmental change, to help prioritize conservation efforts. These quantities span spatial scales from local to regional to global and biological scales from intraspecific to whole communities and large clades.

Prediction in ecology is easiest at very small and large scales. Questions concerning single species or genotypes in simple environments are tractable (Keyl & Wolff 2008, Quinn 2003)

because ecologists can identify the most important environmental drivers, quantify their effects, and make reasonable qualitative and occasionally quantitative model predictions (Blasius et al. 2020, Chen & Irvine 2001, Higgins et al. 1997, Tilman 1977, Yoshida et al. 2003). At broader spatial scales such as whole biomes, the large number of organisms, species, and limiting resources involved allows us to average or aggregate these quantities effectively. Modeling techniques such as statistical mechanics and coarse-graining improve predictions over these scales, partly because errors at small spatial scales tend to cancel out at larger ones (Advani et al. 2018, Dewar & Porté 2008, Moran & Tikhonov 2022). Aside from this, elemental cycles (carbon, nitrogen, and phosphorus cycles), primary productivity, and biomass at global scales are heavily constrained by energetics and fundamental chemistry and can likely be approached from basic chemical and physical principles (Brown et al. 2004, Vallino 2010, Yvon-Durocher & Allen 2012, Zakem et al. 2020).

The most challenging scale for predictive science is the intermediate scale, where we run into the middle-number problem (Kay & Schneider 1994, Newman et al. 2019). At the scale of a region and/or a multispecies community, the number of focal quantities is too large to be computed easily but too small for averaging. Processes operating at smaller scales, such as survival, growth, and intraspecific competition, still matter while being joined by processes that emerge due to ecological and spatial complexity (Hollowed et al. 2000). Models at these scales are likely to be sensitive to initial conditions as well (Hastings et al. 1993, Munch et al. 2022). Many of the properties in the list above—about community composition, phenology, traits, and population densities—unfortunately fall into this difficult intermediate-complexity category.

Response: the property that the environmental driver affects that is of primary interest (e.g., population growth rate); here, the term response is synonymous with the term target

Functional form: in this context, a function describing the shape of the relationship between one or more drivers and a response

HOW CAN WE BEST PREDICT THE FUTURE? A FRAMEWORK FOR PROGRESS IN PREDICTIVE ECOLOGY

No feasible experiment can predict how multiple stressors will affect future ecosystems, because any single experiment is too low-dimensional. It cannot capture the complexity of environmental change, let alone the vast physiological, ecological, and evolutionary changes that will accompany it in the future. The only realistic way we can solve this problem is by tightly integrating theory and experiments (Hanson & Walker 2020, Houlahan et al. 2017). Therefore, to predict the consequences of global change, the most valuable and important experiments are ones that can inform or help evaluate theory or models.

What makes an experiment informative to theory or models? This is difficult to provide a comprehensive answer to, but here are a few types of useful experiments:

1. Experiments that characterize the shape of the relationship between drivers and a response (saturating, exponential, etc.) to identify how best to describe this relationship mathematically. The resulting equation—known as the functional form—is unknown for most cases where more than one driver is considered simultaneously (Collins et al. 2022, Thomas et al. 2017). Identifying the functional form is a crucial step in multiple-driver research (see the sidebar titled Why Understanding Functional Forms Is Important for Understanding and Prediction).
2. Experiments that provide reliable parameter estimates with associated uncertainties. The importance of such estimates is likely underappreciated by empiricists and funding agencies—some of our most important ecosystem models rely on a handful of estimates from species that are easy to grow and work with. For example, Ward et al. (2012) used parameter values from Geider et al. (1998) for a size-structured global ecosystem model. These parameters are sometimes appropriate, but a wider range of parameter estimates would provide better estimates of uncertainty, alleviate biases, and enable models to ask a broader range of questions.

WHY UNDERSTANDING FUNCTIONAL FORMS IS IMPORTANT FOR UNDERSTANDING AND PREDICTION

Functional forms are mathematical functions that describe the shape of a response to one or more drivers. In single-driver research, they describe how a biological response changes across an environmental driver such as temperature or pH. Once the functional form is known, we can estimate its parameters for each species (or at any taxonomic level) or community. We can then write dynamic models to predict the species' performance in any environment, including conditions it does not presently experience anywhere.

A helpful feature of these functional forms is their generality. The population growth rate of all organisms is a saturating function of nutrient or food concentration (Holling 1959, 1966; Monod 1949). For all ectotherms, it is a left-skewed unimodal function of temperature (Kingsolver 2009) (but see next paragraph). For all photoautotrophs, it is a right-skewed unimodal function of light intensity (Eilers & Peeters 1988, Platt et al. 1980). In each of these cases, the underlying ecophysiology is common to the entire set of organisms. This means that we have strong a priori constraints on how environmental change will affect even species for which no data are available. It also means that we can intelligently design the experimental treatments to best constrain the parameters and their predictions (see the main-text section titled Optimal Designs, **Figure 1**, and the **Supplemental Material** section titled Leverage).

Over the past century, we have identified these functional forms for most single drivers (for nutrients, see Michaelis & Menten 1913, Monod 1949). For a few drivers, we have multiple mathematical formulations resulting in similar shapes. This is especially true in the case of temperature (e.g., Johnson & Lewin 1946, Norberg 2004, Ratkowsky et al. 1983, Rezende & Bozinovic 2019, Schoolfield et al. 1981; functions are compared in Krenek et al. 2011 and Padfield et al. 2021). These competing formulations provide equivalent shapes and predictions over much of parameter space, so researchers' choice of function is often arbitrary. However, discrepancies among these functional forms reveal important gaps in our understanding. Temperature functions often differ dramatically in their predictions for growth at very high and low temperatures. Some predict growth rate bottoming out at zero, while others predict extreme temperatures driving it negative (mortality), with more extreme temperatures being worse (Krenek et al. 2011). This difference leads to large differences in predicted performance under precisely the conditions we are most concerned about: heatwaves and future climates. Accurately describing these functional forms is therefore an important scientific goal.

Importantly, for no pair of global change drivers do we have an agreed-upon functional form even for population growth rate—one of the most important measures of success. We have promising candidates for cases such as temperature–nutrient interactions (Huey & Kingsolver 2019, Thomas et al. 2017) (**Figure 1**), although they are yet to be thoroughly validated. For most other driver pairs, such as temperature–CO_2 interactions, and for other important biological parameters, we have insufficient data and theory with which to develop the functions, let alone parameterize them (Collins et al. 2022), which massively undermines our ability to predict the biological consequences of global change.

Supplemental Material >

3. Experiments that can rule out possibilities, either of input parameter space (e.g., most vital rates have at least a weak bound) or of biological consequences. An important case is experiments that can provide constraints by linking different parameters mechanistically, such as in the case of trade-offs. These constraints can rule out vast regions of parameter space, making questions lower-dimensional and tractable. For example, grassland plants generally exhibit a trade-off between growth rates and defense (Lind et al. 2013).
4. Experiments that can validate qualitative predictions from ecological models. Ecological models are often developed as toy models to investigate first principles and rules of thumb about ecological processes. Experiments in model systems and settings that do not strictly

adhere to all of the assumptions can help verify qualitative phenomena, such as changes in population cycles (Blasius et al. 2020, Yoshida et al. 2003).

5. Experiments that can validate quantitative predictions from ecological or ecosystem models. These are often the most challenging and are rarely done outside of short-lived microbial taxa, such as bacteria (Friedman et al. 2017), yeast (Letten et al. 2018), and phytoplankton (Burson et al. 2018, Huisman et al. 1999). Unless the model is developed with a focal system in mind, model assumptions are often not met in experimental systems. Quantitative predictions are most likely to be sensitive to some of these assumptions and are therefore hard to validate.

Process-based models are powerful but are not the only tools with which to predict. Theory-free prediction is often necessary for domains where theory is too complex or impossible to parameterize at present. This is especially true in forecasting, where simple methods such as the autoregressive integrated moving average (ARIMA) family often perform best over short timescales. Machine learning methods amplify our ability to make theory-free predictions by taking into account a wider range of data inputs (Ospici et al. 2022, Qin et al. 2017). However, we cannot rely on theory-free prediction when extrapolating into new combinations of environmental conditions for which we have no reliable training data (no-analogue climates; Fitzpatrick & Hargrove 2009). This extrapolation problem is the principal reason why we must take advantage of ecophysiological and evolutionary knowledge to constrain our predictions. A new generation of process-guided or process-constrained models will help address this problem by incorporating biologically realistic constraints (Hanson et al. 2020, Wagner et al. 2023). Working to improve our fundamental understanding will help both the traditional process-based models and this newer generation of process-guided models, because both rely on us having a reasonable understanding of ecophysiological and evolutionary constraints. We therefore believe that improving understanding is our most promising path to generating better ecological predictions.

Level: a value of an individual driver selected for use in the experiment (e.g., 10°C)

Treatment: a combination of one level each from all drivers selected for use in the experiment (e.g., temperature of 10°C and pH of 6); also called treatment combination

Replicate: the number of experimental units/trials/runs at each treatment; note that we are excluding technical replicates here

WHAT SHOULD WE OPTIMIZE MULTIPLE-DRIVER EXPERIMENTS FOR?

We want experimental design to be maximally informative, efficient, simple (logistically), fast, self-contained (not strongly reliant on prior information), and cost-effective. Some of these criteria are self-explanatory, but here we will expand on what we mean by informative, efficient, and self-contained.

We consider an experiment informative if it provides predictive power beyond the specific conditions used in the experiment. If the experiment's results allow us to accurately interpolate and extrapolate, then it is informative. We can quantify how informative a design is using a metric of out-of-sample prediction error, such as root mean square error (RMSE). To do this for any design, however, we need a theoretical/mathematical/statistical framework with which to make predictions. Knowing the relevant functional form provides us a good starting point (see the sidebar titled Why Understanding Functional Forms Is Important for Understanding and Prediction).

Experiments also need to be efficient, in the sense that the use of resources and experimental units must lead to maximal information gain. An efficient experiment is one that maximizes the information gain per unit of experimental effort. Efficiency therefore depends on our ability to quantify how informative the experiment is and on the experimental effort. For simplicity, we treat experimental effort here as equivalent to the number of experimental units, but logistical complexities and equipment needs should figure into these calculations as well. An experiment with the same number of experimental units will often involve far more effort if more levels are involved per driver, for example. Replicates involve less effort and are easier to run but

generally lead to less information gain (for more discussion of this topic, see the section titled Optimal Designs).

Experimental designs that are self-contained are those where choices of where to allocate experimental effort are not strongly reliant on prior knowledge about the functional form and good guesses about at least some of the parameter values for the focal taxon. Some information is always necessary (and always available), but designs differ in how much they depend on previous knowledge.

TYPES OF EXPERIMENTAL DESIGNS

Ordered roughly from simplest to most complex, some experimental designs that can be used in multiple-driver research are present-versus-future designs, ANOVA (analysis of variance) designs, full factorial response surface designs (at least five levels per factor), fractional factorial designs, quadratic response surface designs, custom designs, space-filling designs, optimal designs, and sequential/adaptive designs. The first three are commonly used, but many of the others are relatively unknown. Below, we discuss how each of these experimental designs may be applied to understand a single species' response to interacting drivers. However, these designs can be applied at multiple scales, including individual populations, interacting pairs of taxa such as predator–prey systems, whole communities via mesocosm experiments, and even whole ecosystems in the case of small ecosystems such as shallow lakes (Schindler 1978, Schindler et al. 2008, Matthijs et al. 2012). Although we focus on constant conditions here, variation or fluctuations in each driver can be explored using the same designs by treating them as separate drivers.

We compare how informative and efficient these designs are using simulations from a temperature–nutrient interaction model that has some empirical support (Thomas et al. 2017) (**Figure 1**; for more detail, see the **Supplemental Material**). The model unites a left-skewed

Supplemental Material >

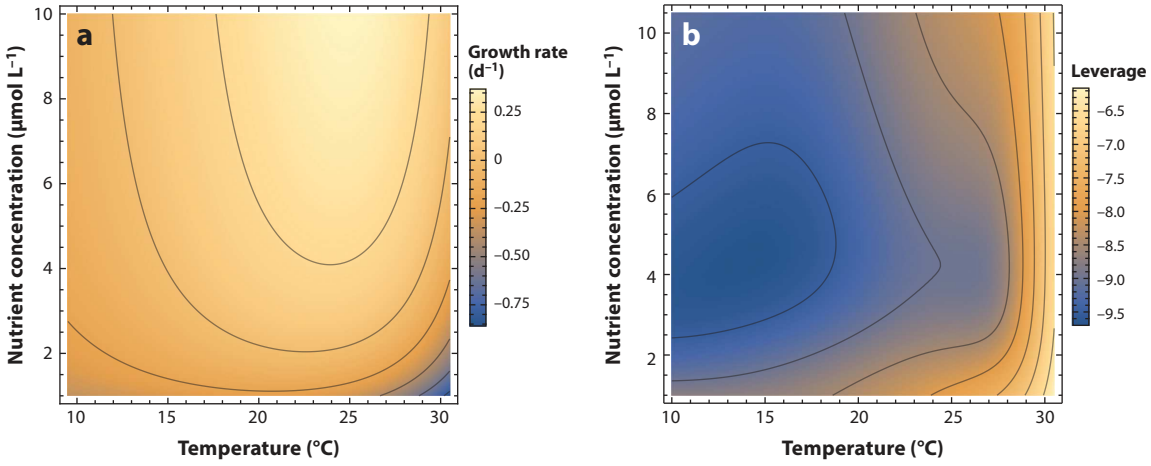

Figure 1

(*a*) An example response surface for our temperature–nutrient interaction model and (*b*) the leverage of points across this surface, which indicates their relative influence on a final model fit (see the **Supplemental Material**). The surface is based on a function and data from Thomas et al. (2017), with parameter values adapted for convenience here. We use this model as the basis of all subsequent simulations. The leverage plot in panel *b* shows the natural logarithm of the tangent plane leverage, which is calculated using a linear approximation of the nonlinear model plotted in panel *a*. High temperatures and low nutrient treatments have high leverage, meaning that experimental points placed in those regions will have a large impact on the fitted surface. Note that this leverage plot is provided mainly for illustration and to guide intuition; an optimal design will not place points solely in high-leverage regions (see the section titled Optimal Designs). We explain how to calculate leverage for nonlinear regressions in the **Supplemental Material**.

Table 1 A qualitative assessment of the different designs' ability to meet our needs

Design	Informative	Efficient	Simple (logistically)	Fast	Self-contained
Present-versus-future	Low	Low	High	High	High
ANOVA	Low	Low	High	High	High
Full factorial response surface	High	Medium/low[a]	Medium/low[a]	High	High
Fractional factorial[b]	Medium	High	Medium	High	Medium
Quadratic response surface	Medium	High	Medium	High	High
Custom	Medium	High	Medium	High	Medium
Space-filling	Medium	Medium	Low	High	High
Optimal	High	High	Low	High	Low
Sequential/adaptive	High	High	Low	Low	Medium

We omit the cost-effective criterion here because it depends strongly on external factors—cost of personnel, equipment, sample analysis, and so on. For all designs except sequential/adaptive, we assume the experiment is conducted in a single run. The simplicity and speed will change for some designs if multiple runs and blocking are necessary.

[a]We judge full factorial response surfaces to be medium in efficiency and simplicity for experiments with two drivers and low for experiments with more than two drivers.

[b]Note that fractional factorial designs apply only to experiments with more than two drivers.

unimodal temperature response with a saturating nutrient response. We use these simulations to quantitatively assess how informative they are, which we follow with a more general qualitative assessment (**Table 1**).

For each design, we simulated 100 experiments (**Figure 2**). Observations have means equal to the "true" values based on the temperature–nutrient response surface with our chosen parameter values (for details, see the **Supplemental Material**). Each point also had noise added that is drawn from a normal distribution with a standard deviation of 0.1. We then used maximum likelihood to fit the known temperature–nutrient function to the simulated data. With the fitted parameter values, we then compared the growth rates estimated across an evenly spaced 100 × 100 grid with the true values at the same points and calculated the RMSE for each simulation. Finally, we compared the RMSE values for all designs to highlight the differences in how informative and efficient they are. We used a 5 × 5 full factorial response surface design as our reference and therefore aimed for a maximum of approximately 25 points. For designs whose main advantage is their efficiency, we used a smaller number of points.

Supplemental Material >

We include replicates where necessary (present-versus-future, ANOVA, and quadratic response surface designs) or required by algorithmic choice (optimal and sequential/adaptive designs), but otherwise we simulate no replication, to highlight the fact that replication is unnecessary in other designs. However, some amount of replication is generally advisable because it is relatively easy and would improve the quality of the fits (decrease RMSE), even though replication is less informative than increasing the number of levels of each driver. We discuss this in more detail in the section titled Optimal Designs.

Present-Versus-Future Designs

Present-versus-future designs are the simplest experiments used to evaluate how systems (populations, communities, or ecosystems) will respond to complex environmental changes. This approach involves setting up two treatments, reflecting present-day and expected future conditions. These experiments are typically analyzed with one-way ANOVAs, which merely establish that there is a difference between the two treatments and their magnitude. The simplicity of this approach makes it easy to integrate many drivers simultaneously and to apply this design to more complex scales (such as mesocosm experiments). The downside is that inferences and predictions

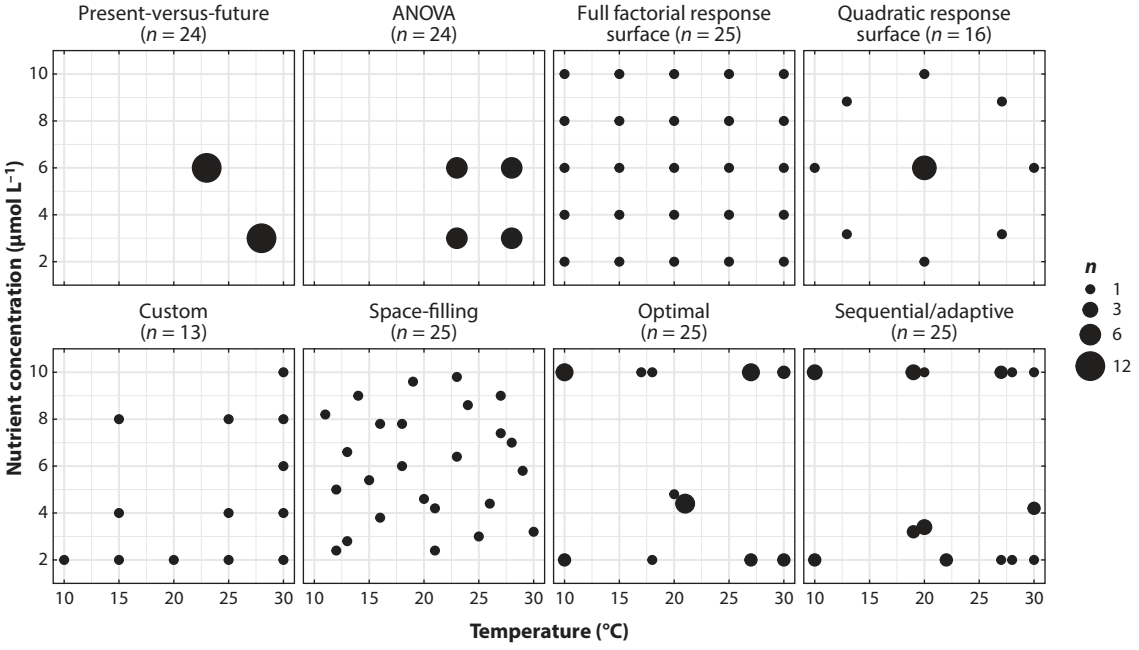

Figure 2

Some of the experimental designs we describe and simulate. Space-filling, optimal, and sequential/adaptive designs differed in each simulation, and we show only one instance here. We omit the fractional factorial design (which is illogical in two dimensions). The sample size *n* used for each experiment is shown, and the point sizes indicate the number of replicates used. The smallest points have no replication.

based on this approach are weak and assumption-heavy. Any deviation in reality from the conditions used in the future treatment makes the results nearly impossible to use for making predictions or inferences.

Recommended use case: None. Although it may be argued that some information is better than none, we believe that this argument neglects the opportunity cost of doing these experiments. Resources directed toward present-versus-future designs are wasted and would be better spent toward more useful, generalizable goals.

ANOVA Designs

Experiments that use a 2 × 2 factorial design, in which two continuous environmental drivers are crossed with each other with two treatment levels each, are the most common type in multiple-driver research (Collins et al. 2022, Jackson et al. 2016, Seifert et al. 2020). These experiments are analyzed with two-way ANOVAs, which establish that the treatment combinations are different. They are easy to run and provide some ability to disentangle the effects of two drivers (or more). These designs can feasibly incorporate a large number of drivers, although not as many as present-versus-future designs.

The existence of an interaction term in the ANOVA conveys the impression of having derived some understanding and predictive power from these experiments, but this is largely illusory. Finding statistically significant differences is largely a question of having enough replicates relative to the process and measurement error, and it is very rare that insights arise from these designs (Cottingham et al. 2005, Kreyling et al. 2018). Where they do, it is either because the experiment

is coupled to a very specific and well-developed hypothesis or question or, more rarely, because the covariation between multiple measured response variables contains useful information. Merely measuring multiple response variables does not make these experiments informative, however.

These 2 × 2 ANOVA results are often used to classify driver interactions into an additive/synergistic/antagonistic framework (Crain et al. 2008, Galic et al. 2018, Piggott et al. 2015). We believe that these classifications and their extensions carry extremely limited and low-value information for population-level experiments (Collins et al. 2022, Orr et al. 2020), though they may be more useful for ecosystem-level measurements. Because of the complex shapes of population response surfaces, sampling the same surface at different locations (i.e., using slightly different treatments for an ANOVA experiment) can lead to any of the three possibilities (additive, synergistic, or antagonistic) (Collins et al. 2022). These and other such classification frameworks rely heavily on unjustified assumptions to make predictions that are qualitative at best and contribute little to predictive power (Orr et al. 2020).

On rare occasions, ANOVA designs with more levels per driver are used, such as 3 × 3 designs. These are curious beasts, in between the 2 × 2 ANOVAs we have criticized here and the substantially more useful response surface designs we discuss in later sections. Though 3 × 3 designs do contain more information, they still reflect the largely pointless ANOVA focus on establishing the existence of nonzero differences between treatments. With the same level of experimental effort, many of the subsequent designs will provide substantially more information.

Recommended use case: Pilot experiments, though even for these they are not superior to alternatives. In rare cases, 2 × 2 ANOVA designs may suffice with a well-developed hypothesis or question and when monotonic responses are strongly expected, as in the case of food or nutrients (e.g., Frisch et al. 2014). And, of course, ANOVA designs are entirely appropriate for categorical factors, which are not the focus of this review.

Full Factorial Response Surface Designs (at Least Five Levels per Factor)

The experimental designs we discuss from here on provide a large increase in information (**Figure 3**).

A full factorial response surface (or regression) design is in principle just an extension of the 2 × 2 ANOVA design. It differs in using more levels per factor and in its lack of requirement for replication (which can be helpful but is not necessary). We define a full factorial response surface design in ecology as at minimum a 5 × 5 experiment. This is because ecophysiological and ecological responses to single environmental drivers are nonlinear and generally need five or six levels to capture them well. The threshold is arbitrary; a 4 × 4 with well-chosen levels may do nearly as well, though this relies on prior knowledge and careful choices to be successful.

Despite the small conceptual difference between a 5 × 5 design and a 2 × 2 ANOVA design, response surface designs enable the prediction of ecological responses at environmental conditions that were not measured during the experiment (Boyd et al. 2018, Collins et al. 2022). That is, they enable meaningful interpolation and extrapolation (though extrapolation is more challenging, especially when not constrained by theory). We can rely on an experiment measuring an ecological response at 10°C, 15°C, 20°C, 25°C, and 30°C to provide good estimates of that response at any specific value between 10°C and 30°C, and to some limited extent below 10°C and above 30°C as well. This would not be true if only 10°C and 30°C (or any pair of temperatures) were measured, as in the first two designs.

Full factorial response surface experiments are inefficient, but a large advantage is that they do not require much prior information to be designed well. If prior information is available, it can be used to space their levels well across the environmental gradients, but simulations suggest that an unevenly spaced full factorial experiment does not meaningfully improve on a well-designed

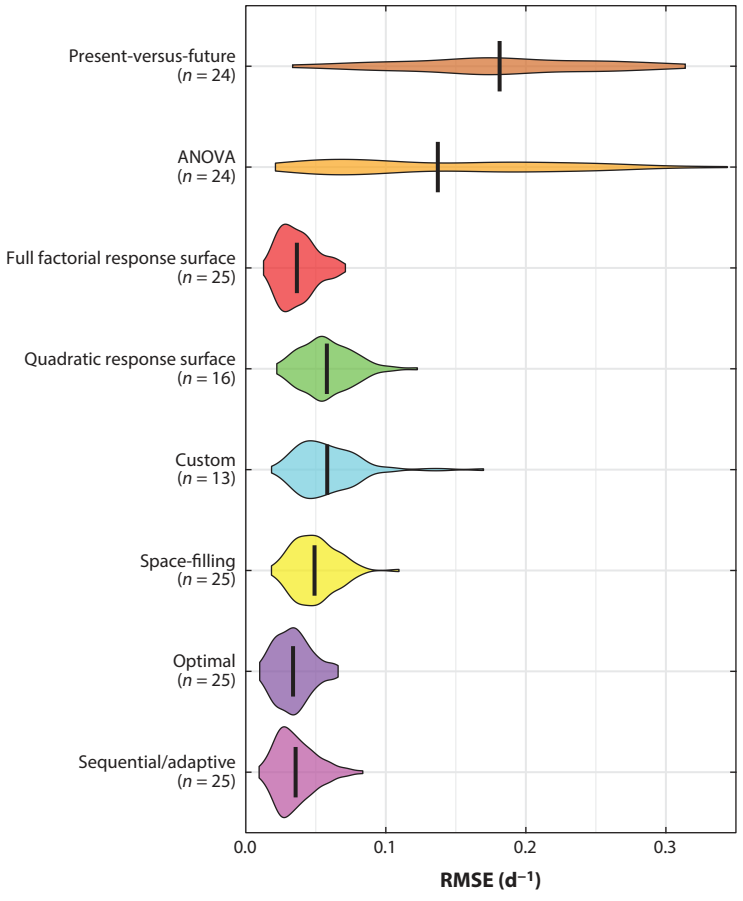

Figure 3

The out-of-sample prediction error from 100 simulated experiments using each of the different designs we discuss (except for fractional factorial, which requires at least three dimensions). Lower RMSE is better. Optimal designs are best, with full factorial response surface and sequential/adaptive designs very close behind. Quadratic response surface, custom, and space-filling designs are approximately twice as bad, although note that the first two were simulated with a smaller sample size. Present-versus-future and ANOVA designs are three to four times worse on average even though we stacked the deck in their favor (see the section titled Comparison of Designs). For a comparison of RMSE across the full surface for all designs, see **Supplemental Figure 4**. Abbreviation: RMSE, root mean square error.

Supplemental Material >

evenly spaced one (**Supplemental Figure 1**). In the temperature–nutrient interaction, a good design includes treatments at low nutrients and at high temperatures (above the optimum temperature for growth; **Figure 1**).

Recommended use case: Quantifying the effects of two and possibly three interacting factors when limited prior information is available. Full factorial response surface designs should be the standard approach until the functional form of an interaction is reasonably well understood and making reasonable parameter guesses is possible (see the sidebar titled Why Understanding Functional Forms Is Important for Understanding and Prediction). These designs help to define the functional forms and are a stepping stone toward more efficient designs. With more than two or three drivers, however, they become unmanageably complex and inefficient.

Fractional Factorial Designs

Fractional factorial designs (Box & Hunter 1961a,b; Box et al. 2005; Gunst & Mason 2009) are a family of methods that address the principal problems with full factorial designs: scale and cost (and therefore efficiency). The central insights are that not all points in a full factorial design are equally informative (**Figure 1*b***; **Supplemental Material**) and that higher-order interactions are typically less important than lower-order ones and main effects (the sparsity-of-effects principle). Experimental designs can therefore be scaled down by neglecting the higher-order interactions.

Supplemental Material >

The more drivers of interest there are, the more useful these methods are. For the two-factor examples we have discussed, they are no different from ANOVA experiments. However, there are often many environmental drivers to consider, such as nutrients, temperature, pH, light availability, predation pressure, and pathogen load. A full factorial experiment with five levels for each of these factors would need $5^6 = 15{,}625$ treatments, which would be far too expensive even if it were logistically possible. Even a full factorial experiment with just two levels for each of these factors (an ANOVA design) would be $2^6 = 64$ treatments. In contrast, a fractional factorial experiment with k factors, I levels per factor, and a fraction of the full factorial experimental effort p would use I^{k-p} treatment combinations. If we use two levels per factor and a half fraction, this amounts to just $2^{6-1} = 32$ treatment combinations, and even lower fractions are feasible. This reduction in size is achieved by intentionally confounding (aliasing) specific combinations of factors relative to the full factorial experiment. As a result, we cannot estimate each separate main effect and interaction term, but the much smaller number of treatments does allow us to rapidly explore many drivers. This is a notable advantage for multiple-driver research, especially if we consider environmental variation for each driver to be an additional dimension.

Fractional factorial experiments therefore enable the extraction of useful information with comparatively little effort. They are used in fields that prioritize predictive power, efficiency, and cost, such as engineering and chemistry. They have seen rare use in ecology (Porter & Busch 1978, Warner et al. 1993) but ought to be a more prominent part of the methodological tool kit in global change biology (Boyd et al. 2018). The main downside to these designs is that they do not characterize the response in a mechanistic way and do not capture nonlinear ecophysiological responses well. The results will be sensitive to the choice of driver levels, especially when the response is nonmonotonic, as most global change drivers are. They therefore do not provide sufficient information to develop mathematical models of response surfaces, but they may suffice to prioritize specific drivers for further investigation with other experimental designs (especially sequential designs) and to provide some predictive power.

Recommended use case: Quickly assessing responses when there are more than five drivers of interest, to prioritize specific drivers (or sets of drivers) to investigate further. Fractional factorial designs are especially useful when exploring the impact of changes in variance (or other higher moments) in addition to changes in mean, by treating each driver's variance as an independent driver itself.

Quadratic Response Surface Designs

Quadratic response surface designs are extensions or modifications of factorial and fractional factorial designs that are commonly used in so-called response surface methodology (RSM) (Box et al. 2005) (see the section titled Sequential/Adaptive Designs). In RSM, first-order methods such as ANOVA designs with an added point at the center enable the estimation of a plane. We focus here on second-order methods, which are sometimes called response surface designs, and emphasize the quadratic aspect to avoid ambiguity, because unlike first-order methods, these approaches are designed to capture curvature in a response surface. We briefly describe two of the most common

methods: central composite designs and Box–Behnken designs. We refer readers to relevant textbooks for details for these and other related designs, such as Doehlert designs (Myers et al. 2009, Wu & Hamada 2009).

Central composite designs begin with an embedded 2×2 factorial. In the simplest case, points are added at the center and extending out from the center through the midpoints of each of the four faces (star or axial points; see **Figure 2**). The projection of this design onto a single dimension captures five distinct levels per driver, enabling the characterization of nonlinear responses. This design scales up well from two dimensions to several more and is most commonly used to identify regions of maxima or minima for industrial production as part of RSM. It is also entering use in ecotoxicology and the natural sciences (Hashemi et al. 2019, Mehdizadeh Allaf & Trick 2019).

Box–Behnken designs include a central point but no embedded factorial design. Instead, points are placed at the midpoints of all edges of a cube or hypercube defined by two levels per driver. This approach therefore requires applying at least three drivers (which is why we do not evaluate it in our simulations here) and uses three levels per factor instead of five. It uses fewer experimental units than central composite designs but at the cost of flexibility. It also omits the corners of the cube or hypercube, and these regions of combined extremes may be important in multiple-driver research.

Broadly, quadratic response surface designs offer efficient ways of capturing high-dimensional interactions with some limited nonlinearity, making them more informative than fractional factorial experiments.

Recommended use case: Quantifying responses when there are more than three drivers of interest, to capture important nonlinearities and interactions when functional forms are unknown. Quadratic response surface designs are especially useful when experimental units are at a premium (such as in mesocosms) and as part of sequential/adaptive designs.

Custom Designs

Fractional factorial and quadratic response surface designs are two ways in which a full factorial design can be reduced to a smaller, more manageable experiment. However, designs can be more flexible when functional forms are known (see the sidebar titled Why Understanding Functional Forms Is Important for Understanding and Prediction) and when there are clear scientific goals (Boyd et al. 2015). These custom designs can downscale factorial experimental designs in a targeted manner, saving experimental effort without much loss of information. They therefore improve on full factorial response surface designs in terms of their efficiency.

When estimating the parameters describing a response surface, there is no strong need for the design to be factorial or even balanced. For example, we do not need to have the same number of nutrient levels at each temperature level in a temperature–nutrient experiment (**Figure 2**). Most experimentalists are aware of the relative value of different points on a response curve or surface, and there is strong statistical justification for this intuition (see the **Supplemental Material**). We can show with simulations and with leverage calculations that some points in the factorial design are less important than others (**Figure 1**). Excluding them only slightly reduces the uncertainty in the parameter estimates and the out-of-sample prediction error when compared with other, more important points. Therefore, when the functional form of the response surface is known (from first principles or previous experiments) and when rough guesses of the parameter estimates (or surface shape) are possible, custom designs can improve experimental efficiency. They allow us to redirect effort toward parts of the surface that are more informative and more likely to occur in nature. Note, however, that custom designs are an inferior form of optimal design, guided by intuition rather than mathematical optimization.

Supplemental Material >

Recommended use case: Parameterizing response surfaces when only a small number of experimental units is possible and the functional form is known, along with rough parameter guesses. Custom designs are also sometimes useful for answering targeted questions that do not require a full response surface.

Space-Filling Designs

Factorial designs fill the two-dimensional (or higher-dimensional) space of the environmental drivers in a simple grid design. Space-filling designs aim to cover this parameter space more efficiently when experimental effort is a limiting factor. Space-filling designs make intelligent use of limited experimental units to capture information about the shape of the response surface by maintaining a reasonably constant density across the surface. A variety of decision criteria are used for the positioning of these units, including maximin, minimax, and the more commonly used Latin hypercube sampling and its variants. For details, we direct interested readers to reviews of these methods (Crombecq et al. 2011, Joseph 2016, Joseph et al. 2019).

Space-filling designs are common in computer simulations, but we have not encountered them in empirical work. However, there is no conceptual reason why they should not be used for multiple-driver research. In principle, they allow one to quantify response surfaces adequately with limited experimental effort. Their major downside is that they achieve this efficiency by avoiding reuse of the same factor levels; for example, in a temperature–pH experiment, no temperature or pH levels will be used more than once. They therefore require the precise manipulation of treatment conditions in individual experimental units, which is not easy to achieve for some drivers. These logistical challenges probably outweigh the benefits of space-filling designs, except perhaps in rare cases, such as mesocosm experiments.

Recommended use case: Quantifying response surfaces when the number of experimental units is a major constraint but manipulating individual treatments is not difficult, such as with mesocosm experiments.

Optimal Designs

Making experiments efficient and informative is an optimization problem. Optimal designs can be thought of as algorithmically optimized versions of custom designs. How should we allocate limited experimental resources to learn the most from them? This question is extremely important since collecting better data is often more useful than collecting more data. The field of optimal design focuses on maximizing the information contained within the data, using optimization theory and linear algebra (Smith 2005, Steinberg & Hunter 1984). Developments in this field have come from attempts to improve yields in chemical and industrial experiments (Box 1954, Box & Wilson 1951, Box & Youle 1955), and despite their success, the resulting ideas have yet to influence the oceanography and global change biology communities. We believe that incorporating ideas from optimal design could lead to more informative and efficient global change experiments. We provide a brief overview here and direct interested readers to introductory books in the field (Berger & Wong 2009, Smith 2005).

Experimental designs can be optimized for different goals, including improving predictions, obtaining the best parameter values, distinguishing among functional forms, and more. These goals lead to different criteria against which designs can be evaluated and the optimal design identified. We focus here on criteria related to minimizing prediction error, though in practice different optimization criteria often lead to similar designs. Prediction-focused optimal designs seek to minimize the variance of the predicted values (i.e., the prediction error). G-optimal designs aim to improve the worst-case scenario and minimize the *maximum* variance across all predicted

values. I-optimal designs focus instead on minimizing the *average* variance across the experimental region (for a brief mathematical explanation of these design criteria, see the sidebar titled Optimal Design Criteria).

OPTIMAL DESIGN CRITERIA

Experimenters can customize experiments for different goals using a range of optimal design criteria. Here, we briefly explain some that might be useful for global change experiments. We consider a regression experiment where

$$y_i = \eta(x_i;\boldsymbol{\theta}) + \varepsilon_i.$$

Here, $y_i (i = 1 \ldots n)$ represents the n observed values of the response, x_i represents the values of the driver, and ε_i represents the error. The function $\eta(x_i;\boldsymbol{\theta})$ relates the driver x to the response y. The vector of parameters to be estimated is represented by $\boldsymbol{\theta}$ with length p. Next, we can define the design matrix, $\mathbf{X}(n \times p)$, such that

$$X_{ij} = \frac{\partial \eta}{\partial \theta_j}(x_i),\ i = 1 \ldots n, j = 1 \ldots p.$$

The ith row of the design matrix $\mathbf{X}$ contains the sensitivities $\partial\eta/\partial\theta_j$ of the function $\eta(x_i;\boldsymbol{\theta})$ to each of the p parameters at the ith observation x_i.

If there is only one parameter ($p = 1$) to be estimated in the regression, a good measure of the quality of fit is the variance of the parameter estimate. The reciprocal of the variance in this case gives us the Fisher information, which is a measure of the information contained in the parameter estimate. Since the variance of the parameter estimate is inversely related to the Fisher information, minimizing the variance means maximizing the information.

However, models typically contain multiple parameters ($p > 1$). In this case, the variances of the parameter estimates form a matrix. Since minimizing a matrix can mean multiple things mathematically, different optimality criteria optimize different quantities. The most popular criterion is known as the D-optimality criterion, and it minimizes the determinant of the information matrix $\mathbf{X^T X}$. Minimizing the determinant of the information matrix is equivalent to minimizing the variance in the parameter estimates. We focus instead on criteria related to minimizing prediction error, as we believe these are more relevant to oceanographers.

Optimality criteria for prediction seek to minimize the variance in the predicted values. We can define the variance $d(x_i;\boldsymbol{\theta})$ of the predicted value at any observation x_i as

$$d(x_i;\boldsymbol{\theta}) = \frac{\partial \eta}{\partial \boldsymbol{\theta}^{\mathbf{T}}}(x_i)(\mathbf{X^T X})^{-1}\frac{\partial \eta}{\partial \boldsymbol{\theta}}(x_i).$$

The G-criterion focuses on the worst-case prediction error, minimizing the maximum variance across all sampling points ($x_i = 1 \ldots n$) in a design. If the designs are mathematically continuous—meaning that the design is specified in terms of the proportion of total experimental units at a point in the design space—then G-optimal designs are equivalent to D-optimal designs (Kiefer & Wolfowitz 1960). The most useful alternative is probably the I-optimality criterion, which minimizes the average prediction variance across the entire range of the predictor variables (χ):

$$\frac{\int_\chi d(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x}}{\int_\chi \mathrm{d}\mathbf{x}}.$$

Other optimality criteria of interest to oceanographers focus on extrapolation. These criteria usually require the point at which extrapolation is desired to be specified in advance, and the design often depends on this point's location. The mathematics involved in deducing extrapolation-focused optimal designs is complex, however, and I-optimality is more useful for predictive purposes.

Prediction error also results from model misspecification. This bias occurs when the fitted function is different from the true function. Scientists often do not know the true function and instead use approximations based on graphical shape (see the sidebar titled Why Understanding Functional Forms Is Important for Understanding and Prediction). The difference between the true function and the approximation can cause prediction error despite an experimental design minimizing prediction variance. Bias-focused optimal designs are model-robust—that is, they minimize bias in experimental designs by accounting for the difference between the true and fitted functions (Box & Draper 1959, 1963; Läuter 1974). Computing bias-focused designs often requires large assumptions and is an active area of research. Most research on these designs has been done on simple polynomial functions, and even this is complex to execute. In our opinion, bias-focused designs require substantial research before they will be of use to oceanographers, and we do not evaluate them here. In contrast to model-robust designs, model-sensitive designs instead allow experimenters to distinguish among candidate functions (Atkinson 1981; Atkinson & Fedorov 1975a,b; Box & Hill 1967; Hunter & Reiner 1965). This may be particularly useful in the context of drivers like temperature, where the shape of the relationship is well known but numerous functions are used (see the sidebar titled Why Understanding Functional Forms Is Important for Understanding and Prediction).

Design space: the set of all possible experimental designs for a given set of drivers and a response

Calculating optimal designs is straightforward in the case of linear regression, but in nonlinear regressions, the optimal design relies on the function's parameter values. Since the true parameter values are unknown before the experiment is performed, this results in a catch-22 where the optimal design cannot be derived without the parameter values and the parameter values cannot be derived without the optimal design. This problem can be solved by using a range of parameter guesses, calculating the optimal design for each, and integrating them into a single final design. There appears to be no prescribed way to select this single final design, which poses a problem. In our simulations (**Supplemental Figure 2**), however, we found reasonable consistency in the optimal designs across a wide range of parameter values. We chose a random subset of these designs to simulate experiments, and most (though not all) of the designs chosen performed very well. One alternative strategy for integrating designs could involve selecting points based on their frequency of occurrence across the optimal designs for different parameter guesses. A different solution to this problem is sequential design (Fedorov 1972), which we discuss in the next section, titled Sequential/Adaptive Designs.

Supplemental Material >

Whichever approach is used, calculating the optimal design is a challenge and is done using one of a variety of exchange algorithms, such as the KL exchange algorithm (Atkinson et al. 2007, Wheeler 2022). These algorithms all start with an initial design and then iteratively perturb it, comparing the optimality criteria at each iteration to decide whether the new design represents an improvement. This is continued until some time- or iteration-based stopping criterion is reached. As it is impossible to explore all possible designs for most reasonable design spaces, there is no guarantee that these algorithms will find a global optimum, but our simulations suggest that this is not a substantial problem.

A design optimal for one functional form is not necessarily optimal for another and may even be worse than a simpler design like a full factorial response surface, highlighting the importance of identifying the functional form (see the sidebar titled Why Understanding Functional Forms Is Important for Understanding and Prediction). Given this function and a range of reasonable parameter guesses, optimal designs tend to outperform even full factorial experiments, though not by much in our simulations (**Figure 3**). When the number of possible experimental units is small, their advantage over full factorial designs will be much larger—our simulations suggest that the optimal design for 15 points performs nearly as well as the full factorial response surface with 25 points (**Figure 3**; **Supplemental Figure 3**).

Optimal designs also offer a more nuanced picture of the oft-debated topic of replication (versus using more treatment levels) in experimental design (Chalcraft 2019, Kreyling et al. 2018). In our simulations, the optimal designs always involved replication at a few treatment combinations but not all (**Figure 2**). This is not to say that the truth is exactly in between the opposing camps, however: In our simulations, optimal designs were only marginally better than full factorial designs with no replication, while replication-focused designs (present-versus-future and ANOVA) were substantially worse than both (**Figure 3**).

Even when the information needed for their calculation is unavailable, optimal designs offer useful insights. At least for our temperature–nutrient function, we found that the borders of the design space (the maximum and minimum values of temperature and nutrients that we were willing to entertain) were of the highest importance, especially the corners. A few points were useful inside those borders, mostly at low to medium nutrients and medium to high temperatures (**Supplemental Figure 2**). The locations of greatest importance will change based on the driver pair (or set), but similar insights are likely possible in each case, once the functional forms are identified.

Supplemental Material >

Recommended use case: Quantifying response surfaces when a reasonable amount of prior information is available (in the form of parameter guesses) and the functional form is known, or to distinguish among possible functional forms. Sufficient prior information is often available for single-driver relationships such as the Holling type II functional response, but at present, such information is harder to come by for multiple-driver response surfaces. Optimal design is especially valuable when the number of possible experimental units is small and in conjunction with sequential design.

Sequential/Adaptive Designs

All of the experimental designs discussed above implicitly involved a single experimental step. However, experiments are often done in multiple rounds because of logistical constraints. This introduces some complexities that need to be accounted for by blocking (which we do not address in this review), but it also creates opportunities because we gain information at each stage of the experiment. Sequential or adaptive experimental design takes advantage of the information gained to improve the subsequent stage of experimentation by allocating experimental treatments efficiently. We have already briefly mentioned one type of sequential design, RSM, in the section titled Quadratic Response Surface Designs. RSM usually focuses on identifying a peak or maximum on the surface, which is not an important goal to us. Instead, we focus on a design strategy to maximize predictive power in a scenario more relevant to global change biologists.

These designs begin with a first experimental round based on any of the previous designs; full factorial response surface, fractional factorial, quadratic response surface, custom, and optimal designs are all reasonable choices. After this first round, we fit the functional form to the data and estimate its parameters. We use these parameter estimates to calculate a new optimal design that adds treatments and/or replicates to the design in the first round. Each step in sequential design improves the parameter estimates, and the process can be continued until the experimenter is satisfied with the estimates or adding more sampling points is not feasible (**Figure 4**). The degree of improvement will decrease in each round, and so a few rounds is likely to be enough to achieve a good fit. While many advanced algorithms are available to implement sequential design efficiently (Ryan et al. 2016), we implemented it relatively simply here for pedagogical purposes. We started with an optimal experimental design in the first step based on reasonable parameter guesses (randomly chosen), with 15 sampling points; we then added 5 points each in the second and third steps, for a total of 25 points (identical to the sample size in the full factorial response

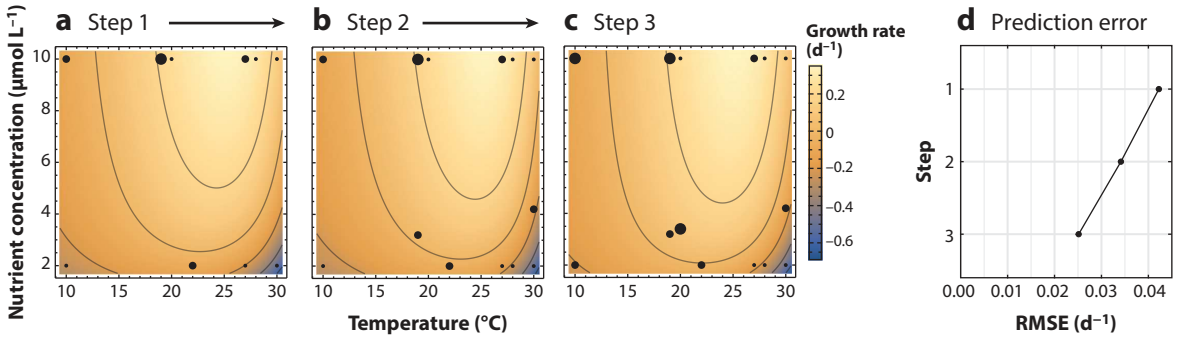

Figure 4

A simulated sequential/adaptive design experiment. We used 100 sets of parameter values as guesses for the initial step, but we are showing only one of them here for illustration. (*a*) We started with an I-optimal design of 15 points generated using the KL algorithm. The KL exchange algorithm iteratively identifies the best design by exchanging the K least promising points at each iteration with the L most promising candidate design points. The values of K and L can be specified by the experimenter. The dots correspond to the position of the experimental points, and the fitted surface is shown beneath. (*b*) In the second step, we added 5 more experimental points. The locations of these points were determined using a modified KL exchange algorithm (see the **Supplemental Material**). (*c*) In the final step, we added another 5 points using the same methodology, bringing the total to 25. While we stopped our sequential design process here, this process could be continued further if needed. (*d*) The effectiveness of the sequential design can be seen via the decrease in prediction error from each fit. As we go from step 1 to step 3 (*top to bottom*), the mean RMSE decreases steeply. Note that even the initial RMSE is good, indicative of the value of the optimal design even with 15 points. Abbreviation: RMSE, root mean square error.

surface simulations). A detailed description of the process we followed is in the **Supplemental Material**. As expected, the mean RMSE decreased with each step as we moved forward in the sequential design process (**Figure 4*d***; **Supplemental Figure 3**).

Supplemental Material >

Sequential design is most useful when there is insufficient information available to do a one-shot optimal design. Designing the first step remains a challenge, and the simplest solution is using a full factorial response surface or optimal design with reasonable parameter guesses (described in the section titled Optimal Designs). Pseudo-Bayesian sequential designs define prior distributions for all model parameters and average design criteria over these distributions to achieve robustness (Pronzato & Walter 1985, D'Arzenio 1990). Bayesian sequential designs or Bayesian adaptive experimental designs go one step further and explicitly define the design criteria to be dependent on the posterior distribution (Ryan et al. 2016). This can be computationally complex and often requires numerical approximations or stochastic solution methods. Applications of sequential design in ecology are therefore scarce (however, see Moffat et al. 2020).

Recommended use case: Quantifying response surfaces when the functional form is known, there is limited prior knowledge to guess parameter values, and time is not a strong constraint. Sequential/adaptive designs are especially helpful when equipment or logistics allows for only a limited set of treatments at one time.

COMPARISON OF DESIGNS

We simulated each experimental design 100 times with random noise added to each measurement and then calculated the prediction error (RMSE) between the fitted surfaces and the true surface. This resulted in a distribution of RMSEs for each design (**Figure 3**). The designs fell into three broad groups. First, full factorial response surface, optimal, and sequential/adaptive designs performed best (i.e., had the lowest prediction error) (**Figure 3**). These models tended to also do best in high-temperature and low-nutrient conditions even when extrapolating slightly, where

Supplemental Material >

other models struggled (**Supplemental Figure 4**). Second, quadratic response surface, custom, and space-filling designs had prediction errors that were approximately twice as high as those in the first group. Note that the principal advantage of these three designs and fractional factorial designs is efficiency (we even used small sample sizes for the quadratic and custom designs), and so these are reasonable designs to use. This is especially true when studying more than two environmental drivers, where efficiency is a much more important criterion. The first group of designs will still outperform the second group in more dimensions but will need substantially more experimental effort (full factorial response surface) or more prior knowledge (optimal) or time (sequential/adaptive designs) (**Table 1**). Note that fractional factorial and quadratic response surface designs are most often used as part of RSM, a sequential/adaptive design framework.

By far the worst designs were the two that are most commonly used: present-versus-future and ANOVA designs. These designs had prediction errors that were three to four times as large as those of the best designs on average, despite using the same number of experimental units (**Figure 3**). While it may be argued that they were disadvantaged by their narrower range of temperature and nutrient values, these values were chosen to be reflective of commonly published experiments. We also made choices that strongly favored these designs by using excellent starting guesses close to the true values, using relatively informative treatment levels, constraining the parameter optimization algorithm to stay within a reasonable range of parameter values, and evaluating performance across the entire surface instead of focusing on regions that had strong interactions. Therefore, this comparison arguably understates the advantage of the better designs.

A PROPOSAL FOR A RESEARCH PLAN FOR THE MULTIPLE-DRIVER COMMUNITY

These simulations (**Figure 3**) illustrate the value of knowing functional forms and choosing appropriate experimental designs based on the objectives and constraints (**Table 1**). The challenge we face as a community is integrating these ideas into a larger project to predict how multiple drivers will reshape ocean ecology. To that end, we propose a general outline of what a community-wide effort to achieve this goal may look like, taking advantage of these designs. Not every question needs species-level response surfaces to answer, and so some of these steps will be unnecessary for some goals.

For any species, a small set of relevant drivers can be identified using fractional factorial or quadratic response surface experiments. Note that variation, fluctuations, and extreme events can be thought of as separate dimensions in this context and that the relevant drivers may already be known in well-studied taxa. Once these drivers are known, we need to generate candidate models for how they shape the response of individual species (or whole communities) to multiple interacting drivers. These models can be generated from first principles or based on full factorial response surface experiments. If there are multiple candidate models, model-sensitive designs (see the section titled Optimal Designs) can be used to identify the best functional form. Experiments with custom, optimal, or sequential/adaptive designs can then be used to estimate the parameters for these functional forms for additional species.

With the functional form and these parameter values, marine scientists can take advantage of theory–experiment feedback. Mathematical models can be used to generate expectations for single-species dynamics in constant and varying environments. Theoretical expectations can be validated and refined from experiments with single species (or communities) in these environments (Gerhard et al. 2023). This approach can then be scaled up to models with multiple species and multiple trophic levels. Mesocosm experiments will be particularly advantageous for validation at these intermediate scales. For longer-lived organisms or organisms otherwise unsuited to

experiments, models can be used to predict patterns in nature and validated with observational data. After these rounds of validation, we can scale up further with Earth system models and other global models (whose predictions will be largely untestable except over long timescales) to develop projections with associated uncertainties based on different climate scenarios.

Optimizing our experimental designs at every stage is key to linking theory and experiments, and therefore to building a more predictive and mechanistic science. This will be a challenging community-wide undertaking that requires a wide range of expertise. However, we believe it is the most promising path toward accurately predicting the future of our ecosystems.

APPENDIX: SOFTWARE TOOLS USED

To perform the simulations and analyses in this review, as well as generate the figures, we used Mathematica 12.3 (Wolfram Res. 2021) and R 4.2.2 (R Core Team 2022) along with the following R packages: AlgDesign 1.2.1 (Wheeler 2022), bbmle 1.0.25 (Bolker & R Dev. Core Team 2022), cowplot 1.1.1 (Wilke 2020), gridExtra 2.3 (Auguie & Antonov 2017), lhs 1.1.6 (Carnell 2022), OptimalDesign 1.0.1 (Harman & Filova 2019), optimx 2022.4.30 (Nash & Varadhan 2011, Nash 2014), rootSolve 1.8.2.3 (Soetaert & Herman 2009, Soetaert et al. 2021), rsm 2.10.3 (Lenth 2009), and tidyverse 2.0.0 (Wickham et al. 2019).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The work presented in this article results in part from funding provided by national committees of the Scientific Committee on Oceanic Research (SCOR) and from a grant to SCOR from the US National Science Foundation (OCE-1840868) to the Changing Oceans Biological Systems project.

LITERATURE CITED

Advani M, Bunin G, Mehta P. 2018. Statistical physics of community ecology: a cavity solution to MacArthur's consumer resource model. *J. Stat. Mech. Theory Exp.* 2018(3):033406

Atkinson AC. 1981. A comparison of two criteria for the design of experiments for discriminating between models. *Technometrics* 23(3):301–5

Atkinson AC, Donev AN, Tobias RD. 2007. *Optimum Experimental Designs, with SAS*. Oxford, UK: Oxford Univ. Press

Atkinson AC, Fedorov VV. 1975a. Optimal design: experiments for discriminating between several models. *Biometrika* 62(2):289–303

Atkinson AC, Fedorov VV. 1975b. The design of experiments for discriminating between two rival models. *Biometrika* 62(1):57–70

Auguie B, Antonov A. 2017. gridExtra: miscellaneous functions for "grid" graphics. *Comprehensive R Archive Network.* **https://CRAN.R-project.org/package=gridExtra**

Berger MPF, Wong WK. 2009. *An Introduction to Optimal Designs for Social and Biomedical Research*. Chichester, UK: Wiley & Sons

Blasius B, Rudolf L, Weithoff G, Gaedke U, Fussmann GF. 2020. Long-term cyclic persistence in an experimental predator-prey system. *Nature* 577(7789):226–30

Bolker B, R Dev. Core Team. 2022. Bbmle: tools for general maximum likelihood estimation. *Comprehensive R Archive Network.* **https://CRAN.R-project.org/package=bbmle**

Box GEP. 1954. The exploration and exploitation of response surfaces: some general considerations and examples. *Biometrics* 10(1):16–60

Box GEP, Draper NR. 1959. A basis for the selection of a response surface design. *J. Am. Stat. Assoc.* 54(287):622–54

Box GEP, Draper NR. 1963. The choice of a second order rotatable design. *Biometrika* 50(3–4):335–52

Box GEP, Hill WJ. 1967. Discrimination among mechanistic models. *Technometrics* 9(1):57–71

Box GEP, Hunter JS. 1961a. The 2^{k-p} fractional factorial designs part I. *Technometrics* 3(3):311–51

Box GEP, Hunter JS. 1961b. The 2^{k-p} fractional factorial designs part II. *Technometrics* 3(4):449–58

Box GEP, Hunter JS, Hunter WG. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Hoboken, NJ: Wiley-Intersci. 2nd ed.

Box GEP, Wilson K. 1951. On the experimental designs for exploring response surfaces. *Ann. Math. Stat.* 13:1–45

Box GEP, Youle PV. 1955. The exploration and exploitation of response surfaces: an example of the link between the fitted surface and the basic mechanism of the system. *Biometrics* 11(3):287–323

Boyd PW, Collins S, Dupont S, Fabricius K, Gattuso J-P, et al. 2018. Experimental strategies to assess the biological ramifications of multiple drivers of global ocean change—a review. *Glob. Change Biol.* 24(6):2239–61

Boyd PW, Lennartz ST, Glover DM, Doney SC. 2015. Biological ramifications of climate-change-mediated oceanic multi-stressors. *Nat. Clim. Change* 5(1):71–79

Brown JH, Gillooly JF, Allen AP, Savage VM, West GB. 2004. Toward a metabolic theory of ecology. *Ecology* 85(7):1771–89

Burson A, Stomp M, Greenwell E, Grosse J, Huisman J. 2018. Competition for nutrients and light: testing advances in resource competition with a natural phytoplankton community. *Ecology* 99(5):1108–18

Carnell R. 2022. lhs: Latin hypercube samples. *Comprehensive R Archive Network*. **https://CRAN.R-project.org/package=lhs**

Chalcraft DR. 2019. To replicate, or not to replicate – that should not be a question. *Ecol. Lett.* 22(7):1174–75

Chen DG, Irvine JR. 2001. A semiparametric model to examine stock recruitment relationships incorporating environmental data. *Can. J. Fish Aquat. Sci.* 58(6):1178–86

Collins S, Whittaker H, Thomas MK. 2022. The need for unrealistic experiments in global change biology. *Curr. Opin. Microbiol.* 68:102151

Cottingham KL, Lennon JT, Brown BL. 2005. Knowing when to draw the line: designing more informative ecological experiments. *Front. Ecol. Environ.* 3(3):145–52

Crain CM, Kroeker K, Halpern BS. 2008. Interactive and cumulative effects of multiple human stressors in marine systems. *Ecol. Lett.* 11(12):1304–15

Crombecq K, Laermans E, Dhaene T. 2011. Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling. *Eur. J. Oper. Res.* 214(3):683–96

Currie DJ, Mittelbach GG, Cornell HV, Field R, Guegan J-F, et al. 2004. Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecol. Lett.* 7(12):1121–34

D'Arzenio DZ. 1990. Incorporating prior parameter uncertainty in the design of sampling schedules for pharmacokinetic parameter estimation experiments. *Math. Biosci.* 99:105–18

Dewar RC, Porté A. 2008. Statistical mechanics unifies different ecological patterns. *J. Theor. Biol.* 251(3):389–403

Dietze MC. 2017. Prediction in ecology: a first-principles framework. *Ecol. Appl.* 27(7):2048–60

Eilers PHC, Peeters JCH. 1988. A model for the relationship between light intensity and the rate of photosynthesis in phytoplankton. *Ecol. Model.* 42(3–4):199–215

Fedorov V. 1972. *Theory of Optimal Experiments*. New York: Academic

Fitzpatrick MC, Hargrove WW. 2009. The projection of species distribution models and the problem of non-analog climate. *Biodivers. Conserv.* 18(8):2255–61

Friedman J, Higgins LM, Gore J. 2017. Community structure follows simple assembly rules in microbial microcosms. *Nat. Ecol. Evol.* 1(5):0109

Frisch D, Morton PK, Chowdhury PR, Culver BW, Colbourne JK, et al. 2014. A millennial-scale chronicle of evolutionary responses to cultural eutrophication in *Daphnia*. *Ecol. Lett.* 17(3):360–68

Galic N, Sullivan LL, Grimm V, Forbes VE. 2018. When things don't add up: quantifying impacts of multiple stressors from individual metabolism to ecosystem processing. *Ecol. Lett.* 21(4):568–77

Geider RJ, MacIntyre HL, Kana TM. 1998. A dynamic regulatory model of phytoplanktonic acclimation to light, nutrients, and temperature. *Limnol. Oceanogr.* 43(4):679–94

Gerhard M, Koussoroplis A-M, Raatz M, Pansch C, Fey SB, et al. 2023. Environmental variability in aquatic ecosystems: avenues for future multifactorial experiments. *Limnol. Oceanogr. Lett.* 8(2):247–66

Gunst RF, Mason RL. 2009. Fractional factorial design. *Wiley Interdiscip. Rev. Comput. Stat.* 1(2):234–44

Halpern BS, Frazier M, Afflerbach J, Lowndes JS, Micheli F, et al. 2019. Recent pace of change in human impact on the world's ocean. *Sci. Rep.* 9(1):11609

Hanson PC, Stillman AB, Jia X, Karpatne A, Dugan HA, et al. 2020. Predicting lake surface water phosphorus dynamics using process-guided machine learning. *Ecol. Model.* 430:109136

Hanson PJ, Walker AP. 2020. Advancing global change biology through experimental manipulations: Where have we been and where might we go? *Glob. Change Biol.* 26(1):287–99

Harman R, Filova L. 2019. OptimalDesign: a toolbox for computing efficient designs of experiments. *Comprehensive R Archive Network*. **https://CRAN.R-project.org/package=OptimalDesign**

Hashemi SH, Kaykhaii M, Jamali Keikha A, Sajjadi Z, Mirmoghaddam M. 2019. Application of response surface methodology for silver nanoparticle stir bar sorptive extraction of heavy metals from drinking water samples: a Box-Behnken design. *Analyst* 144(11):3525–32

Hastings A, Hom CL, Ellner S, Turchin P, Godfray HCJ. 1993. Chaos in ecology: Is Mother Nature a strange attractor? *Annu. Rev. Ecol. Syst.* 24:1–33

Higgins K, Hastings A, Sarvela JN, Botsford LW. 1997. Stochastic dynamics and deterministic skeletons: population behavior of Dungeness crab. *Science* 276(5317):1431–35

Holling CS. 1959. The components of predation as revealed by a study of small-mammal predation of the European pine sawfly. *Can. Entomol.* 91(5):293–320

Holling CS. 1966. The functional response of invertebrate predators to prey density. *Mem. Entomol. Soc. Can.* 98:5–86

Hollowed AB, Bax N, Beamish R, Collie J, Fogarty M, et al. 2000. Are multispecies models an improvement on single-species models for measuring fishing impacts on marine ecosystems? *ICES J. Mar. Sci.* 57(3):707–19

Houlahan JE, McKinney ST, Anderson TM, McGill BJ. 2017. The priority of prediction in ecological understanding. *Oikos* 126(1):1–7

Huey RB, Kingsolver JG. 2019. Climate warming, resource availability, and the metabolic meltdown of ectotherms. *Am. Nat.* 194(6):E140–50

Huisman J, Jonker RR, Zonneveld C, Weissing FJ. 1999. Competition for light between phytoplankton species: experimental tests of mechanistic theory. *Ecology* 80(1):211–22

Hunter WG, Reiner AM. 1965. Designs for discriminating between two rival models. *Technometrics* 7(3):307–23

IPCC (Intergov. Panel Clim. Change). 2021. *Climate Change 2021: The Physical Science Basis; Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. V Masson-Delmotte, P Zhai, A Pirani, SL Connors, C Péan, et al. Cambridge UK: Cambridge Univ. Press

Jackson MC, Loewen CJG, Vinebrooke RD, Chimimba CT. 2016. Net effects of multiple stressors in freshwater ecosystems: a meta-analysis. *Glob. Change Biol.* 22(1):180–89

Johnson FH, Lewin I. 1946. The growth rate of E. coli in relation to temperature, quinine and coenzyme. *J. Cell Comp. Physiol.* 28(1):47–75

Joseph VR. 2016. Space-filling designs for computer experiments: a review. *Qual. Eng.* 28(1):28–35

Joseph VR, Gu L, Ba S, Myers WR. 2019. Space-filling designs for robustness experiments. *Technometrics* 61(1):24–37

Kay JJ, Schneider E. 1994. Embracing complexity: the challenge of the ecosystem approach. In *Perspectives on Ecological Integrity*, Vol. 5, ed. L Westra, J Lemons, pp. 49–59. Dordrecht, Neth.: Springer

Keyl F, Wolff M. 2008. Environmental variability and fisheries: What can models do? *Rev. Fish Biol. Fish.* 18(3):273–99

Kiefer J, Wolfowitz J. 1960. The equivalence of two extremum problems. *Can. J. Math.* 12:363–66
Kingsolver JG. 2009. The well-temperatured biologist (American Society of Naturalists Presidential Address). *Am. Nat.* 174(6):755–68
Krenek S, Berendonk TU, Petzoldt T. 2011. Thermal performance curves of *Paramecium caudatum*: a model selection approach. *Eur. J. Protistol.* 47(2):124–37
Kreyling J, Schweiger AH, Bahn M, Ineson P, Migliavacca M, et al. 2018. To replicate, or not to replicate – that is the question: how to tackle nonlinear responses in ecological experiments. *Ecol. Lett.* 21(11):1629–38
Läuter E. 1974. Experimental design in a class of models. *Math. Oper. Stat.* 5(4–5):379–98
Lenth RV. 2009. Response-surface methods in R, using rsm. *J. Stat. Softw.* 32(7):1–17
Letten AD, Dhami MK, Ke P-J, Fukami T. 2018. Species coexistence through simultaneous fluctuation-dependent mechanisms. *PNAS* 115(26):6745–50
Lind EM, Borer E, Seabloom E, Adler P, Bakker JD, et al. 2013. Life-history constraints in grassland plant species: a growth-defence trade-off is the norm. *Ecol. Lett.* 16(4):513–21
Matthijs HCP, Visser PM, Reeze B, Meeuse J, Slot PC, et al. 2012. Selective suppression of harmful cyanobacteria in an entire lake with hydrogen peroxide. *Water Res.* 46(5):1460–72
Mehdizadeh Allaf M, Trick CG. 2019. Multiple-stressor design-of-experiment (DOE) and one-factor-at-a-time (OFAT) observations defining *Heterosigma akashiwo* growth and cell permeability. *J. Appl. Phycol.* 31(6):3515–26
Michaelis L, Menten ML. 1913. Die kinetik der invertinwirkung. *Biochem. Z.* 49:333–69
Moffat H, Hainy M, Papanikolaou NE, Drovandi C. 2020. Sequential experimental design for predator-prey functional response experiments. *J. R. Soc. Interface* 17(166):20200156
Monod J. 1949. The growth of bacterial cultures. *Annu. Rev. Microbiol.* 3:371–94
Moran J, Tikhonov M. 2022. Defining coarse-grainability in a model of structured microbial ecosystems. *Phys. Rev. X* 12(2):021038
Munch SB, Rogers TL, Johnson BJ, Bhat U, Tsai C-H. 2022. Rethinking the prevalence and relevance of chaos in ecology. *Annu. Rev. Ecol. Evol. Syst.* 53:227–49
Myers RH, Montgomery DC, Anderson-Cook CM. 2009. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Hoboken, NJ: Wiley. 3rd ed.
Nash JC. 2014. On best practice optimization methods in R. *J. Stat. Softw.* 60(2):1–14
Nash JC, Varadhan R. 2011. Unifying optimization algorithms to aid software system users: optimx for R. *J. Stat. Softw.* 43(9):1–14
Newman EA, Kennedy MC, Falk DA, McKenzie D. 2019. Scaling and complexity in landscape ecology. *Front. Ecol. Evol.* 7:293
Norberg J. 2004. Biodiversity and ecosystem functioning: a complex adaptive systems approach. *Limnol. Oceanogr.* 49(4):1269–77
Orr JA, Vinebrooke RD, Jackson MC, Kroeker KJ, Kordas RL, et al. 2020. Towards a unified study of multiple stressors: divisions and common goals across research disciplines. *Proc. R. Soc. Edinb. B* 287(1926):20200421
Ospici M, Sys K, Guegan-Marat S. 2022. Prediction of fish location by combining fisheries data and sea bottom temperature forecasting. In *Image Analysis and Processing – ICIAP 2022*, pp. 437–48. Cham, Switz.: Springer
Padfield D, O'Sullivan H, Pawar S. 2021. *rTPC* and *nls.multstart*: a new pipeline to fit thermal performance curves in R. *Methods Ecol. Evol.* 12(6):1138–43
Pennekamp F, Adamson MW, Petchey OL, Poggiale J-C, Aguiar M, et al. 2017. The practice of prediction: What can ecologists learn from applied, ecology-related fields? *Ecol. Complex.* 32:156–67
Piggott JJ, Townsend CR, Matthaei CD. 2015. Reconceptualizing synergism and antagonism among multiple stressors. *Ecol. Evol.* 5(7):1538–47
Platt T, Gallegos CL, Harrison WG. 1980. Photoinhibition of photosynthesis in natural assemblages of marine phytoplankton. *J. Mar. Res.* 38:103–11
Porter WP, Busch RL. 1978. Fractional factorial analysis of growth and weaning success in *Peromyscus maniculatus*. *Science* 202(4370):907–10

Pottier J, Dubuis A, Pellissier L, Maiorano L, Rossier L, et al. 2013. The accuracy of plant assemblage prediction from species distribution models varies along environmental gradients: climate and species assembly predictions. *Glob. Ecol. Biogeogr.* 22(1):52–63

Pronzato L, Walter E. 1985. Robust experiment design via stochastic approximation. *Math. Biosci.* 75(1):103–20

Qin M, Li Z, Du Z. 2017. Red tide time series forecasting by combining ARIMA and deep belief network. *Knowl.-Based Syst.* 125:39–52

Quinn TJ. 2003. Ruminations on the development and future of population dynamics models in fisheries. *Nat. Resour. Model.* 16(4):341–92

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna: R Found. Stat. Comput. **https://www.R-project.org**

Ratkowsky DA, Lowry RK, McMeekin TA, Stokes AN, Chandler RE. 1983. Model for bacterial culture growth rate throughout the entire biokinetic temperature range. *J. Bacteriol.* 154(3):1222–26

Rezende EL, Bozinovic F. 2019. Thermal performance across levels of biological organization. *Philos. Trans. R. Soc. Lond. B* 374(1778):20180549

Ryan EG, Drovandi CC, McGree JM, Pettitt AN. 2016. A review of modern computational algorithms for Bayesian optimal design. *Int. Stat. Rev.* 84(1):128–54

Schindler DW. 1978. Factors regulating phytoplankton production and standing crop in the world's freshwaters. *Limnol. Oceanogr.* 23(3):478–86

Schindler DW, Hecky RE, Findlay DL, Stainton MP, Parker BR, et al. 2008. Eutrophication of lakes cannot be controlled by reducing nitrogen input: results of a 37-year whole-ecosystem experiment. *PNAS* 105(32):11254–58

Schoolfield RM, Sharpe PJH, Magnuson CE. 1981. Non-linear regression of biological temperature-dependent rate models based on absolute reaction-rate theory. *J. Theor. Biol.* 88(4):719–31

Seifert M, Rost B, Trimborn S, Hauck J. 2020. Meta-analysis of multiple driver effects on marine phytoplankton highlights modulating role of pCO_2. *Glob. Change Biol.* 26(12):6787–6804

Shaw RG. 2019. From the past to the future: considering the value and limits of evolutionary prediction. *Am. Nat.* 193(1):1–10

Smith WF. 2005. *Experimental Design for Formulation*. Philadelphia: Soc. Ind. Appl. Math.

Soetaert K, Herman PMJ, eds. 2009. *A Practical Guide to Ecological Modelling: Using R as a Simulation Platform*. Dordrecht, Neth.: Springer

Soetaert K, Hindmarsh AC, Eisenstat SC, Moler C, Dongarra J, Saad Y. 2021. rootSolve: nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations. *Comprehensive R Archive Network*. **https://CRAN.R-project.org/package=rootSolve**

Steinberg DM, Hunter WG. 1984. Experimental design: review and comment. *Technometrics* 26(2):71–97

Thomas MK, Aranguren-Gassis M, Kremer CT, Gould MR, Anderson K, et al. 2017. Temperature-nutrient interactions exacerbate sensitivity to warming in phytoplankton. *Glob. Change Biol.* 23(8):3269–80

Tilman D. 1977. Resource competition between plankton algae: an experimental and theoretical approach. *Ecology* 58(2):338–48

Vallino JJ. 2010. Ecosystem biogeochemistry considered as a distributed metabolic network ordered by maximum entropy production. *Philos. Trans. R. Soc. Lond. B* 365(1545):1417–27

Wagner T, Schliep EM, North JS, Kundel H, Custer CA, et al. 2023. Predicting climate change impacts on poikilotherms using physiologically guided species abundance models. *PNAS* 120(15):e2214199120

Ward BA, Dutkiewicz S, Jahn O, Follows MJ. 2012. A size-structured food-web model for the global ocean. *Limnol. Oceanogr.* 57(6):1877–91

Warner SC, Travis J, Dunson WA. 1993. Effect of pH variation of interspecific competition between two species of hylid tadpoles. *Ecology* 74(1):183–94

Wheeler B. 2022. AlgDesign: algorithmic experimental design. *Comprehensive R Archive Network*. **https://CRAN.R-project.org/package=AlgDesign**

Wickham H, Averick M, Bryan J, Chang W, D'Agostino McGowan L, et al. 2019. Welcome to the tidyverse. *J. Open Source Softw.* 4(43):1686

Wilke CO. 2020. Cowplot: streamlined plot theme and plot annotations for "ggplot2." *Comprehensive R Archive Network*. **https://CRAN.R-project.org/package=cowplot**

Wolfram Res. 2021. *Mathematica*. Champaign, IL: Wolfram Res. **https://www.wolfram.com/mathematica**
Wu C-F, Hamada M. 2009. *Experiments: Planning, Analysis, and Optimization*. Hoboken, NJ: Wiley. 2nd ed.
Yoshida T, Jones LE, Ellner SP, Fussmann GF, Hairston NG. 2003. Rapid evolution drives ecological dynamics in a predator-prey system. *Nature* 424(6946):303–6
Yvon-Durocher G, Allen AP. 2012. Linking community size structure and ecosystem functioning using metabolic theory. *Philos. Trans. R. Soc. Lond. B* 367(1605):2998–3007
Zakem EJ, Polz MF, Follows MJ. 2020. Redox-informed models of global biogeochemical cycles. *Nat. Commun.* 11(1):5680