Causal inference with observational data and unobserved confounding variables

Jarrett E. K. Byrnes[1*] and Laura E. Dee[2*]

1 - Department of Biology, University of Massachusetts Boston, Boston, MA 02125
2 - Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO 80308-0334
* - Co-first authors

Correspondence: jarrett.byrnes@umb.edu and laura.dee@colorado.edu

Code Repository: https://github.com/jebyrnes/ovb_yeah_you_know_me

App for one simulated dataset: https://shiny.umb.edu/shiny/users/jarrett.byrnes/shiny_ovb/

App for replicate simulations: https://shiny.umb.edu/shiny/users/jarrett.byrnes/ovb_sims/

Keywords: omitted variable bias, causal inference, endogeneity, structural causal model, observational data, correlation, panel regression, correlated random effects

Abstract

As ecology tackles progressively larger problems, we have begun to move beyond the scale at which we can conduct experiments to derive causal inferences. Randomized controlled experiments have long been seen as the gold standard for quantifying causal effects in ecological systems. In contrast, observational data, though available at larger scales, has primarily been used to either explore ideas derived from experiments or to generate patterns to inspire experiments – not for causal inference. This avoidance of using observational data for causal inference arises from the valid fear of confounding variables – variables that influence both the causal variable of interest and the studied effect that can lead to spurious correlations. Unmeasured confounders can lead to incorrect conclusions – a problem known as Omitted Variable Bias – that leads to the common saying, "Correlation is not causation." However, many other scientific disciplines that cannot do experiments for reasons of ethics or feasibility have developed rigorous approaches for causal inference from observational data. Here we show how Ecologists can harness these approaches, starting by using causal diagrams to identify potential known and unknown sources of confounding. We use a motivating example of assessing the

effects of warming on intertidal snails to discuss how ecologists currently handle observational survey data and inference - often incorrectly with mixed models that produce biased coefficient estimates. We present alternative sampling designs and the statistical model designs that make use of them, discuss how they work using the language of causal path diagrams, demonstrate how easily they can be applied to common ecological datasets, and finally how well they are able to overcome problems of unmeasured confounding variables. We show how all of these techniques out-perform common approaches via simulation with respect to both bias and power. Our goal is to enable researchers to advance the field of Ecology at scale using observational data both on its own and as an important complement to experiments.

Introduction

As Ecology advances to address problems at scales from the continental to global, we are putting our theories to the test like never before – working at larger scales in space and time and with unprecedented streams of data. To address fundamental questions in Ecology with these data, we desire to answer questions about causal relationships. These answers allow us to either test basic theory at scale or inform conservation and ecosystem management. Classically in Ecology, understanding causal relationships has been the domain of experiments. Experiments, however, have limitations for generalizing to large scales and contexts beyond study conditions. As Ecology seeks to address theory and application at scale, we must rapidly move beyond a scale where ideal randomized experiments are possible (reviewed in Kimmel *et al.* 2021), and seize – responsibly – the opportunity of new large-scale sources of observational data.

Our ability to test hypotheses about causal relationships in observational data is limited by two fundamental challenges. First, nature is complex! When we use observational data to attempt to answer causal questions, we face numerous **confounding variables** – variables correlated with the cause and the outcome of interest – that can lead to incorrect estimates of causal effects (Fig. 1). Failing to control for these confounding variables leads to **bias** in our estimate of the relationship between a predictor and its response; the estimate will not be equal to its true value. A simple solution is to statistically control for confounding variables; this requires knowing what they are and measuring then. Even when we know what confounders to account

for, collecting the data needed to account for each and every one is likely impossible. Further, measuring them with error introduces other sources of bias. The second major challenge is that, as humans, we are limited in our ability to imagine how the different elements of complex ecological systems are related. Thinking through the entire natural history of a system to design an analysis of observational data enabling credible causal inferences is really hard. As a result, causal inference from observational data is often dismissed as impossible due to the potential for spurious correlations, prompting the common saying "correlation is not causation." The correlation between number of pirates and global average temperature (Henderson 2006) stands as a cautionary tale when teaching about the dangers of inferring causation from correlation. At its core, inferring causation from correlations centers on dealing with the problem of unmeasured confounding variables – those that influence both a causal variable and the response of interest and can lead to spurious correlations or mask true causal relationships (Fig. 1).

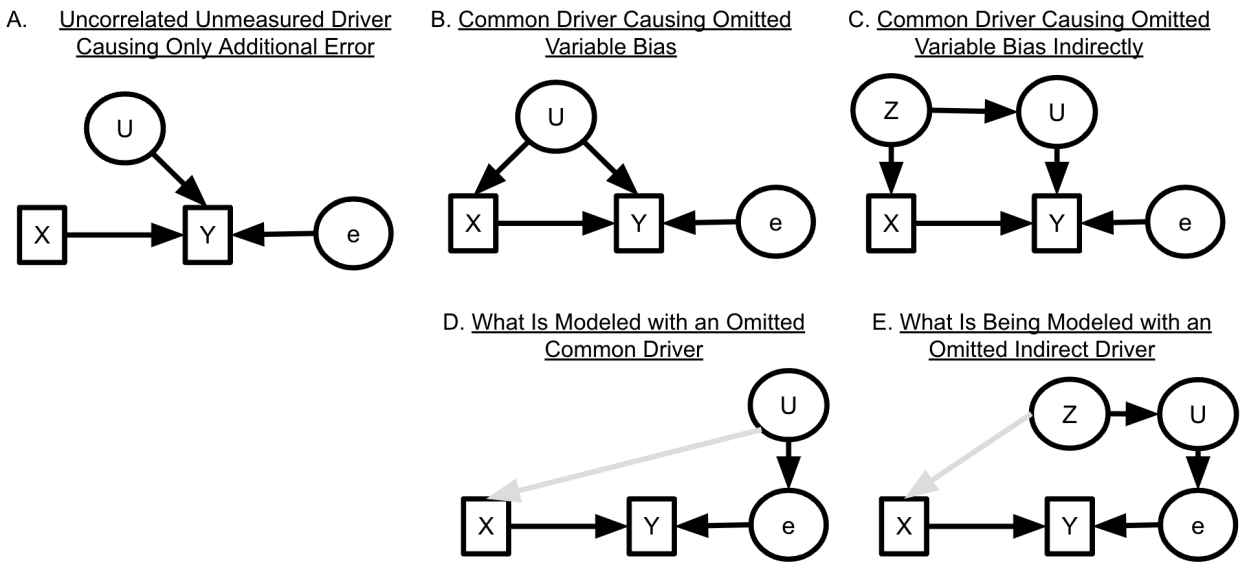

Figure 1. Illustrating Omitted Variable Bias from Confounding Variables and How it Affects Statistical Models. A response variable of interest (*Y*) is caused by both a measured variable (*X*) and an unmeasured variable (*U*). The random error term (*e*) represents other unmodeled influences in the system that do not influence X. In **(A)**, *X* and *U* are uncorrelated, and thus the lack of inclusion of *U* in a statistical model increases the standard error of the estimate (decreases model precision) but does not lead to bias in the effect of *X* on *Y*. However, if *U* also drives *X* as in **(B)** or if *U* and *X* are driven by a common driver *Z* as in **(C),** then omitting *U* from a statistical model causes omitted variable bias in the estimate of the effect of *X* on *Y*. The direction of the bias in the estimator depends on the effect of *U* on *Y*, with the result that the causal effects are of *X* on *Y* are either over- or under-estimated, masking or mimicking a causal effect. If we modeled the system in **(B)** without including *U*, then we would be modeling a system as seen in (**D**), where *U* is left as part of the error term thereby inducing a correlation between *X* and *e*, creating an endogeneity problem, because the gray path from *U* to *X* is not controlled for in the model. Similarly, for a common driver having an indirect effect on *Y* as in **(C),** if we were to model that, we would be modeling a system as seen in (**E**). Again, the path from *Z* to *X* is not included in the model (gray path), inducing a correlation between *e* and *X* and creating an endogeneity problem.

Excluding known but unmeasured, or unknown and unmeasured, confounding variables from a statistical analysis creates what is known as **omitted variable bias** (Rinella *et al.* 2020; Wooldridge 2015). Omitted variable bias (OVB) could be positive or negative and results in estimators yielding incorrect the magnitudes – and even signs – of estimates (i.e., biased estimators). Omitted confounding variables could occur because of missing measurements or due to failures of imagination, because we do not yet know all of the confounding variables. For example, one might measure plant communities to study competition, but not measure all the soil properties that drive all species, due to financial or time constraints. Similarly, working with long-term survey data or in human impacted systems, missing data on confounding variables is common. Consider using historical measures of fish abundance to study the impacts of changes in biogenic habitat availability without measurements of fishing pressure during the same time-period. We have no way of knowing the direction or magnitude of the bias, because knowing all possible confounding variables and their relationships in a system is hard, if not impossible. Measuring, controlling for, and even knowing all potential confounding variables is nearly impossible in complex ecological systems (*reviewed in* Dee *et al.* 2023). In short, in observational data collection and analysis, we are always going to miss something, threatening the validity of our causal inferences.

Do these challenges mean that we should not try to use observational data for causal inference? No. Rather than throwing up our hands, discounting and abandoning the use of observational data for causal inference, we suggest that ecologists consider adopting techniques from other disciplines that cannot do experiments – often for logistical or ethical reasons. For instance, it is not ethical to make a person smoke cigarettes daily to test the causal effect of smoking on dementia (Hernan & Robins 2023); one can only manipulate curricula so far in an effort to understand educational outcomes. Disciplines such as psychology, economics, education, epidemiology, sociology, computer science, and more have been building tools to handle OVB in the causal analysis from observational data for decades (Angrist & Pischke 2008; Heckman 2000; Hernan & Robins 2023; Holland 1986; Imbens & Rubin 2015; Morgan & Winship 2015; Pearl 2009; Robins 1989; Rubin 1974, 2005). Indeed, these advances received the 2021 Nobel prize in Economics.

As ecologists, we have a decades-long tradition of considering experiments as a gold standard for causal inference (Benedetti-Cecchi & Cinelli 1997; Carpenter *et al.* 1985; Gotelli & Ellison 2012; Kimmel *et al.* 2021; Lubchenco 1980; Paine 1966; Power 1990; Reichman 1979; Silvertown *et al.* 2006; Underwood *et al.* 1997). However, experiments also rely on assumptions for causal interpretation of effects (Kimmel *et al.* 2021), which can be hard to meet in the field. Experiments can also induce artefacts or rely on conditions that make them hard to generalize to natural ecosystems (e.g., see discussions in Ruesink 2000; Stachowicz *et al.* 2008; Wolkovich *et al.* 2012). Further, this reliance on the primacy of experiments has meant that the tools of other disciplines have been largely absent from the ecologist's toolbox (*but see* Butsic *et al.* 2017; Grace 2021; Kendall 2015; Larsen *et al.* 2019; Rinella *et al.* 2020; Shipley 2016 for example). Recently, though, interest in causal inference in ecology for observational data has grown, drawing on a diverse suite of methods from other fields focused on casual inference (Arif & MacNeil 2022, 2023; Dee *et al.* 2023; Dudney *et al.* 2021; Grace & Irvine 2020; Larsen 2013; Larsen *et al.* 2019; MacDonald & Mordecai 2019; Simler-Williamson & Germino 2022). If we are, as a discipline, to move to more widespread use of observational data for causal inference, we need to carefully consider the problems of such approaches and the techniques we can use to mitigate them, and their assumptions.

Here, we aim to provide a guide to readily available ways to cope with Omitted Variable Bias (OVB) for Ecologists. We begin by describing the status quo for how ecologists most often deal with omitted variable bias. We then review tools for identifying potential sources of omitted variable bias before conducting a study or analysis, building on the foundation of using directed acyclic graphs that has become increasingly common in ecology (Arif & MacNeil 2023). To illustrate problems with OVB once identified and different ways to address it, we present a motivating example aiming to study the causal effect of temperature on marine snail abundances. With this example, we then outline sampling and statistical designs for dealing with omitted variable bias. We show the conclusions that would be drawn from the typical approaches an ecologist might take with data that might contain omitted confounders (e.g., random effects in a mixed model, *see* Bolker et al. 2009) and discuss why they fall short of dealing with OVB (i.e., have statistical bias). We then present several other statistical designs that can more adequately control for omitted variables. Most of these statistical designs are underutilized, if not novel, for ecology. We demonstrate their utility with results from simulation analyses showing that these

designs are more robust to OVB; they successfully eliminate more sources of bias from confounding variables. We provide guidance for choosing among these designs for different data contexts and questions, and hands-on tutorials with R code for prospective users. Our goal is to enable researchers to advance the field of Ecology at scale using observational data.

How are ecologists coping with Omitted Variables Bias?

Omitted variable bias is commonly dealt with in one of five ways in Ecology. First, Ecologists use randomized controlled experiments. In an ideal, randomized controlled experiment, the effect of confounding variables is eliminated when their assumptions are met (but see Kimmel *et al.* 2021 on why this can be difficult in practice), by the random assignment of treatments (e.g., Nitrogen addition) to units (plots), so that the treatment and control groups have the same level of any confounders on average. However, randomized controlled experiments, particularly at scale, are not always feasible and have limitations in terms of their ability to generalize beyond the experimental conditions. Second, in observational studies, ecologists attempt to deal with confounding variables by measuring the confounders and controlling for them in a multivariate statistical model. As described above, measuring all confounders, however, is often impossible – particularly for retrospective analyses of existing data. Further, all potential confounders in the system might not be known. Third, Ecologists fold unmeasured cluster-level variables into random effects in mixed models (Bolker *et al.* 2009; Harrison *et al.* 2018; Schielzeth & Nakagawa 2012). As we will discuss extensively below in our section on statistical model designs, if random effects are correlated with causal drivers of interest, this will lead biased estimates. Fourth, Ecologists sometimes make causal claims rooted in their knowledge of the natural history of a system. These claims can be problematic due to a lack of transparency and potential for incorrect statements of effect size; even the knowledge of the most accomplished naturalist can have gaps in their understanding of a system. Fifth, Ecologists often qualify their results verbally to avoid making causal claims – even when their research focus is causal understanding, rather than description (*but see* Laubach *et al.* 2021 on causal aims and claims). This practice muddies the waters and can create confusion over whether an author claiming an association or implying causation while allowing themselves plausible deniability. We feel that given our current need to understand causal relationships from large-

scale observational data sets, these solutions are inadequate and can even lead to misleading inferences.

Ecologists have the opportunity, nay, obligation, to leverage the solutions to Omitted Variable Bias in causal data analysis that other disciplines have been building for decades. This paper provides an entry point into several approaches and complements recent reviews of what are commonly referred to quasi-experimental methods (e.g., Antonakis *et al.* 2010; Arif & MacNeil 2022; Bell *et al.* 2018; Bellemare *et al.* 2024; Butsic *et al.* 2017; the appendices of Dee *et al.* 2023; Ferraro & Miranda 2017; Grace & Irvine 2020; Kendall 2015) by expanding on panel regression designs for accounting for OVB.

Using causal diagrams to clarify causal assumptions and ferret out Omitted Variables Bias

Causal diagrams (a.k.a. Structural Causal Models from Pearl 1995; see Grace & Irvine 2020; Arif & MacNeil 2023 for in depth introductions for Ecologists) are one of the first tools for identifying and addressing omitted variable bias (Arif & MacNeil 2023; Pearl 1995; Pearl *et al.* 2016). Causal diagrams in the form of Directed Acyclic Graphs (DAGs, see Box 1) allow us to visualize our understanding of causal relationships and confounding variables within a system. In doing so, DAGs transparently clarify many assumptions on which one relies for making causal claims about relationships inferred from observed data, and potential sources of bias from confounding variables. Critically, a causal diagram needs to include both measured and *unmeasured* confounding variables. Finally, causal diagrams can also show what variables should <u>not</u> be included in an analysis, such as those that cause collider bias. Collider bias occurs when evaluating a relationship between two variables, but conditioning on a variable they both cause For example, conditioning on plant abundance when examining the relationship between disturbance intensity and herbivory intensity when both are causes of plant abundance (for an excellent discussion of this topic beyond the scope of this manuscript, see McElreath McElreath 2020 Chapter 6; Laubach *et al.* 2021; Griffith *et al.* 2020). Thus, we suggest drawing a DAGs before conducting an analysis from which one wants to make any causal conclusions.

If possible, we recommend making a DAG *before* data collection to inform which covariates might be confounding and should be measured if possible. However, due to feasibility

constraints or if one is analyzing pre-existing data, measuring all potential confounders might not be possible. Further, the data could have been collected for another purpose or question, so a set of confounders were deemed unimportant.

Box 1: A Brief Overview of the Elements of Directed Acyclic Graphs for Causal Analysis

We briefly review the uses and the elements of causal diagrams (e.g., Fig. 1), called Directed Acyclic Graphs (DAGs). For the variables and implied causal relationships (as paths), we adopt symbology to differentiate between observed and unobserved variables to reveal where confounding variables might lurk. In our examples of DAGs, first, observed variables that can be or have been measured are represented as terms within boxes, as for *X* and *Y* in Figure 1. Second, our DAG in Figure 1 shows *unobserved* (i.e., unmeasured) variables contained in ellipses, such as the variable *U.* The error term is shown as *e*. This term is a stand-in for all other possible influences on *Y*. While *e* is often omitted by convention from DAGs, we find it helpful to include to force researchers to consider whether unobserved sources of variation affect other variables in the DAG that ultimately *X*. Variables are connected by paths, i.e., arrows. The direction of these arrows represents a direct causal connection going in the direction the arrow is pointed. These arrows are non-parametric; we make no assumptions of functional form of the relationship. If the value of a causal variable of interest changes (e.g., via manipulation, exposure, or a natural process), there will be a concomitant change in the response variable(s) or "outcomes" it affects. If a response variable changes, say via direct manipulation, there will be no associated change in the causal variable of interest.

A common critique is that DAGs do not include feedbacks, to which we respond by asking the reader to think of their definition of causality. Here, we adopt the Neyman-Rubin counterfactual causality framework (Holland 1986; Rubin 1974, 2005) where we recognize that cause temporarily precedes effect. Therefore, feedbacks can be handled by thinking about a system with a temporal lag (e.g., Larson *et al.* 2008). If an instantaneous feedback is truly present, or if a time-series of both the driver and response variable is not available, one will likely require other tools such as instrumental variables - something beyond the scope of this manuscript (for a comprehensive review see Imbens 2014; for an ecological perspective see Grace 2021; for examples, MacDonald & Mordecai 2019; Dee *et al.* 2023).

In this article, we emphasize how thinking in terms of DAGs helps to determine where confounding variables might cause problems with omitted variables bias and, in turn, helps identify solutions in terms of sampling and statistical designs. As applied researchers, we have found that DAGs paired with robust statistical approaches for causal inferences have often clarified our own thinking and communication about ecological systems.

After building a DAG, as described in Box 1, one can determine potential sources of omitted variable bias from variables influencing both the cause of interest and outcome that have not been observed in the system (e.g., confounding unobservable variables, *U* in Fig. 1B). Not controlling for this confounding variable opens a "back-door" for information to flow between your causal variable of interest and its response variable via an unassessed pathway (Pearl 2009). In the case of Figure 1B, *U* would be folded into a statistical model's error term, along with random sources of error, as seen in Figure 1D. We have the same problem if the unobserved variable has an indirect effect (see Fig. 1C and E). In all cases, the model's error term and causal variable of interest would be correlated, leading to **endogeneity** – a violation of a core assumption of the Gauss-Markov theorem (Abdallah *et al.* 2015; Antonakis *et al.* 2010).

What is endogeneity and why is it problem? Consider an example of evaluating the relationship between nitrogen availability and plant biomass across a series of fields. If nitrogen availability depends on field soil characteristics, and field soil characteristics also drive plant biomass, but field soil characteristics were omitted from an analysis, then 1) the effects of soil characteristics would be included in the error term of that model so that 2) nitrogen is no longer **exogenous** (external to the system of interactions). Rather, it is **endogenous** – it is affected by elements that are part of the error term, e.g., in a causal diagram an arrow would go *from* error *to* nitrogen. This endogeneity creates a correlation between the error and nitrogen in a naïve statistical model, misattributing the effect of soil characteristics to nitrogen and leading to incorrect estimates of nitrogen alone on plant biomass. Said another way, we are no longer only estimating the effect of nitrogen controlling for soil characteristics; as a consequence, the estimate of the nitrogen effect will be wrong – different from the true effect in magnitude or even sign. As discussed below, making field a random effect does not resolve this problem. If we

draw a DAG, we will be able to see where these types of endogeneity problems are likely to occur.

In the absence of a randomized experiment, one way to eliminate confounding from a measured variable (i.e., *U* in Fig. 2) is to include a variable in a statistical model. Including the variable controls for its confounding effect, thereby blocking all paths between *X* and *Y* via *U*. This variable to control for could be the confounder itself or another variable that is a "child node" or downstream of U that blocks the path affecting *X* or *Y* (e.g., *W* in Fig. 2 – if there was more than one path and *W* didn't block both, we'd need to control for a variable on the second path as well). Controlling for such a confounding variable means that the ensuing analysis will satisfy the **back-door criterion** (Pearl 1995, Fig. 2A); when all back doors are "shut", the influence of confounders are removed. Thus, the causal effect of our predictor of interest will be identified, as we have established conditional independence between the causal variable of interest and the error term. Without including and thus controlling for these confounding variables or others that block their influence (see Fig 2. and its caption for several examples) in a statistical model, omitted confounders will cause Omitted Variable Bias.

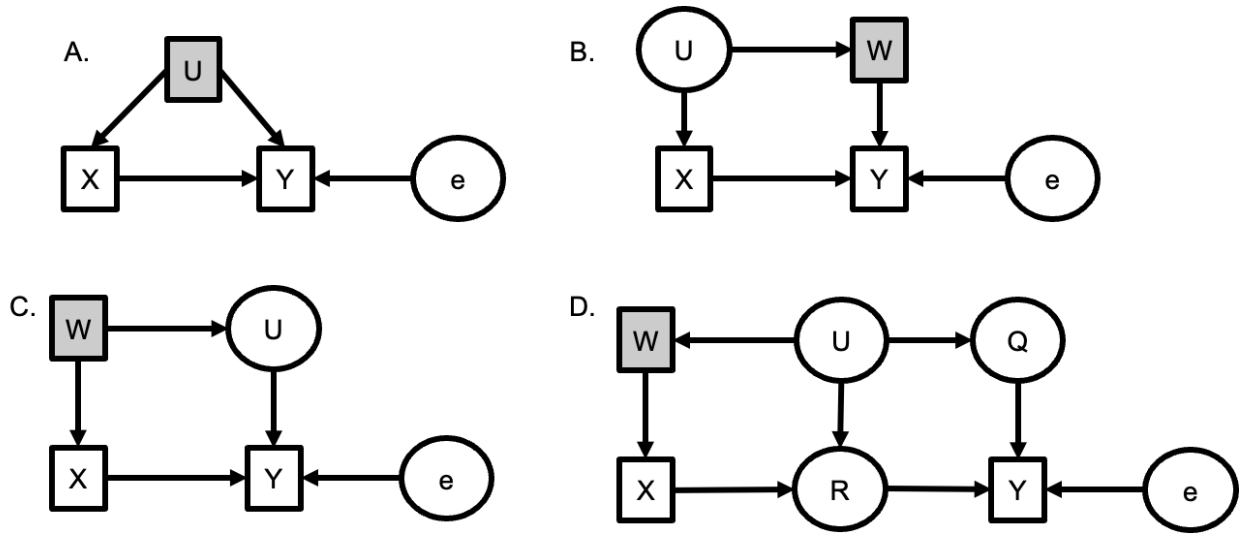

Figure 2. Examples of statistical control for confounding variables informed by causal graphs. By including shaded observed variables, either *U* or *W*, in a statistical analysis of the effects of *X* on *Y*, omitted variable bias is controlled for the results have a causal interpretation. The relationship between the control variable and *Y* might (as in panel **A** and **B**) or might (as in panel **C** and **D**) not have a causal interpretation, depending on the structure of the system. Note, in (**D**), *Q* would have also served as an adequate control instead of *W*. *R* would have been a bad control.

DAGs can help identify how and when to control for confounding variables. With a diagram in hand, confounding variables can either be visually obvious or one can utilize a variety of

software for analyzing DAGs and conditional independence among variables (e.g., Textor *et al.* 2016). With this knowledge, we can find "open back-doors" (confounding pathways between the causal variable of interest and response variable) that need to be controlled for to eliminate omitted variable bias. Perhaps, most importantly, one can justify their choice of control variables with a DAG, making their assumptions about how a system works before an analysis transparent to readers of their work in the literature. Even further, a DAG can show us where it is not possible to "shut the back door" for causal identification of effects, requiring us to use other techniques like instrumental variables or Pearl's "front door criterion" (Pearl 2009; Bellemare *et al.* 2024).

A causal diagram is, therefore, the first step on the way for identifying potential omitted variable bias. On their own, however, they do not in and of themselves provide a means for statistically controlling for OVB, particularly if we have not measured the confounding variable. Nor does a causal diagram help us in the face of *unknown* confounding variables that we have failed to imagine as part of our system. Indeed, we may also not know the true DAG that represents the data generating process and causal and confounding relationships in the system. A DAG only represents a researcher's current understanding and own assumptions about the causal relationships within a system. However, they provide a useful tool to begin to think about where spatial and temporal confounders that we have failed to think about might be lurking and begin to consider choices to address them. Even if we do not have the correct DAG or have not thought of all possible confounders, recognizing the places where spatial and temporal confounders might lurk allow us to leverage complementary approaches that lessen our reliance on a perfectly correct DAG to eliminate effects of confounding variables. We next review these approaches that combine observational sampling designs with statistical designs to also account for *unobserved* and potentially *unknown* confounding variables to eliminate more sources of omitted variables bias.

A Problem of Omitted Snails

To illustrate these empirical challenges and suite of potential solutions, we consider a marine benthic ecosystem, modeled after the Gulf of Maine, USA, where a researcher aims to study the causal effect of temperature on snail abundance. They hypothesize that temperature influences snail metabolic and mortality rates and wish to estimate its effect on snail population

abundance. Snail population abundance is also driven by recruitment, in part influenced by regional oceanography (i.e., the flow of major currents that differ in a myriad of properties) that drives both water temperature and recruitment patterns (Broitman *et al.* 2005; Yund *et al.* 2015). We assume that the researcher measured snail abundance and temperature at several sites but not recruitment or any measurement of oceanography. Thus, recruitment and oceanography are unmeasured, or so-called "unobserved" confounding variables. Estimates produced from an analysis of just the temperature-snail relationship will almost certainly be or different from the true value. Even if the researcher had measured recruitment, though, what if there are other lurking confounding variables? Even if oceanography or recruitment were accounted for, omitted variable bias remains a real possibility – and the estimated effect of temperature on snails will be incorrect, and could even differ in not only the magnitude but also the sign of the true effect. Fortunately, our researcher drew out a causal diagram of the system as a DAG (Fig. 3) and recognized that temperature at the scale of measurement was also influenced by local variation (e.g., small-scale oceanographic features, weather, or other sources of local or microclimatic variability). With this causal diagram in hand, they realized they could control for both observed and unobserved confounding variables with appropriate sampling and statistical model designs.

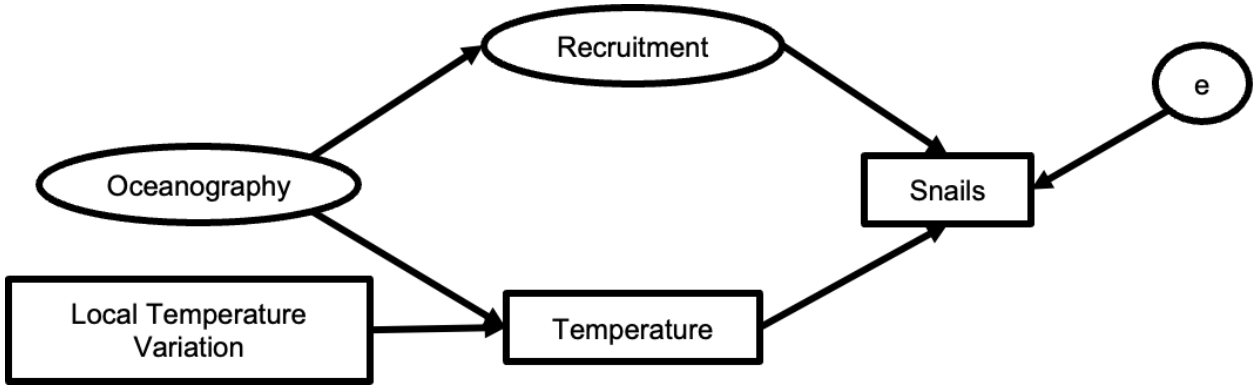

Figure 3. A causal diagram describing the controls of snail abundance in an intertidal ecosystem. Oceanography drives both temperature and recruitment, both of which drive snail abundance. Temperature, however, is also driven by local influences as well. This could be variability in plot-level temperature within a site – i.e., sources of variation in microclimate - or site-level temperature variability over space or time uncorrelated with local oceanography, recruitment, or other site- or plot-level confounders.

Sampling Designs that enable statistical methods to cope with omitted variable bias

Multiple sampling designs for data collection enable the use of statistical model designs that can address omitted variable bias from confounding variables that vary across space, time, or both. A key feature in these sampling designs is that there is some hierarchical or clustered

structure to the data: the **nesting** of multiple observations within a cluster or group (e.g. site) can allow the causal variable of interest to vary across replicates while the confounder varies at the cluster level (Fig. 4) which we will see becomes very important to aid in deconfounding our results. Clustered data is often also referred to as a hierarchical or nested sampling design (Gelman & Hill 2006). We use these terms interchangeably. Using our snail and temperature example, we outline different nested sampling designs and discuss how they generate different source of variation in space and time that enable the use of statistical model designs that deal with confounders.

Nested sampling designs can take several forms and generate difference types of variation to study. First, a sampling design could include multiple plots sampled within sites at a single point in time (Fig. 4A) – a **cross-sectional design**. When sites span environmental gradients with variation in a causal variable of interest (i.e., temperature differences), confounding variables also vary across these spatial gradients. In our example (Fig. 3), a spatial gradient in temperature across sites also reflects the spatial gradient in oceanography that affects both temperature and recruitment, thus confounding this causal relationship of interest between temperature and snails across sites. However, with data collected from a cross-sectional sampling design with plots within sites, we can use variation in plot temperature *within*-sites to isolate its effect on snails rather than the variation *between* sites, which contains sources of confounding variation (e.g., rockpools of different size that warm to different degrees during low tide for within versus site-level oceanographic features that drive temperature and recruitment for between).

Second, one could sample the same plots (or sites) repeatedly through time (Fig. 4B) in a **longitudinal** or **panel data design**. This data structure enables using approaches that can leverage variation *within-sites through time*. As such, longitudinal data can enable many approaches to remove the effects of confounding variables that vary across sites, switching the variation we study to variation within-site (or plot) through time as opposed to between sites. Developing an understanding of how cross-sectional and panel data structures, along with variations and extensions (Box 2 and 3), can be used in conjunction with statistical designs to remove variation from confounders is key to confronting OVB.

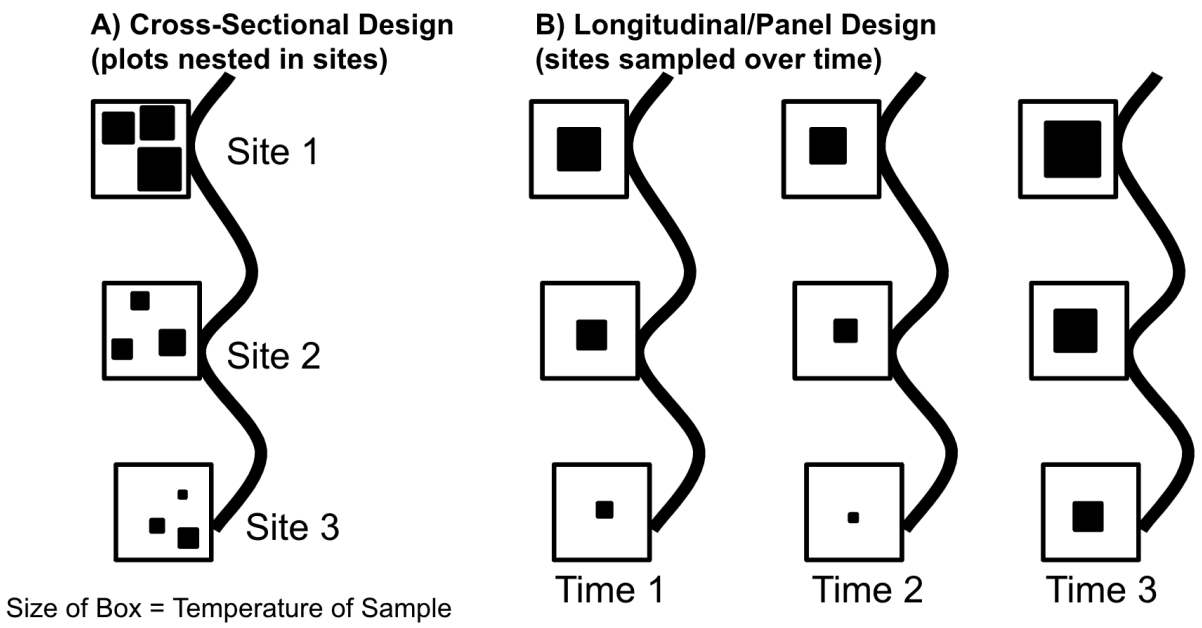

Figure 4. Visual examples of hierarchical study designs with plots nested within sites sampled at one point in time in A and through time in B. This figure shows sites distributed along a coastline with a corresponding thermal gradient, with one or more plots sampled within each site, depending on the design. Open squares are sites. Closed squares are plots within sites. Size of square is proportional to temperature. These sampling designs therefore have variation across space, as in the cross-sectional sampling design in A, or in both space and time as in B. which shows longitudinal or panel data, where the same plots within sites are observed through time. The sampling design in (A) can allow researchers to study temperature variation within sites as well as between sites. The design in (B) enables a researcher to leverage variation in space and time, including examining variation within sites through time.

Statistical Model Designs to Coping with Omitted Variables

With hierarchical/clustered (hereafter clustered) data and a DAG in hand, there are well-established statistical designs to handle omitted confounders for causal analysis. We emphasize the term *'designs'* over *'methods'* because one could implement these statistical designs using a variety of estimation approaches (e.g., linear regression, Generalized Linear Models, as a part of Structural Equation Models, or with Bayesian techniques). These statistical designs have different costs and benefits, and they differ in their assumptions required for interpreting an estimate as a causal effect. Yet, most of the following designs – with the exception of random effects models as shown below – allows us to flexibly control for confounding variables that are both known and unknown (see Angrist & Pischke 2008; Dudney *et al.* 2021; Ferraro & Miranda 2017) – something many Ecologists worry about. Thus, we believe these statistical designs are a key advance worth considering for ecologists.

We illustrate how the different designs work using a common set of terms for causal variables of interest *(x*; e.g. local temperature), the outcome or response of interest (*y*; e.g. snail counts), and confounding variables (*w*; e.g. recruitment) in a regression, applied to our example of the snail system in Figure 3. Our example includes data from different sites (*i*) sampled either at multiple time points in panel data design or in multiple plots (*j*) in the case of a cross-sectional data design as above. For the sake of simplicity, we assume a linear model form with normally distributed error (ϵ), although the framework applies for generalized linear models as well. The model takes the form of

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma w_i + \epsilon_{ij} \quad (1)$$

Here, y_{ij} is the abundance of snails at site *i* in year or plot *j*, β_0 is the intercept – the abundance of snails if the temperature and recruitment were 0, β_1 is the effect of temperature x_{ij} at site *i* in year or plot *j* on snails, γ is the effect of recruitment w_i at site *i* on snail abundance, Our goal is to estimate β_1 (the effect of temperature on snail abundance) and eliminate the effect of confounding variables (and resulting bias). Due to shared oceanographic influences, x_{ij} and w_i are correlated. If we had measured w_i, then we could include it in our model, and by conditioning on observables with γas the effect of *w* on *y*, produce a causally identified estimate of β_1, assuming no other confounders. Without measuring and including the confounder, *w*, in the design above, and then fitting the equation of

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij} \quad (2)$$

our causal inference about β_1 would be incorrect: different from the true causal effect because γw_i would be included in the error term, inducing a correlation between our error and causal variable of interest (i.e., "endogenous"). This endogeneity problem violates the assumptions of the Gauss-Markov theorem and its extensions (Wooldridge 2015) and is what underlies the problem of omitted variable bias (see simulations below to see this consequences of this bias in action).

What Ecologists Typically Do: Random or Mixed Effects Models That Fail to Solve OVB

Mixed effects models have been popular in ecology for the past two decades (for useful reviews, see Bolker et al. 2009, Schielzeth and Nakagawa 2012, Harrison et al. 2018). Originally used to partition variation in heritability between different relatives (Fisher 1919), **random effects –** the effects of clusters in data assumed to come from a random distribution (but see Gelman & Hill 2006 on the linguistic ambiguities surrounding fixed and random effects) – quickly became a mainstay in the partitioning of variation in randomized experiments with subsamples taken within clusters (Cochran 1937; Eisenhart 1947). They have become a standard part of the toolbox for analyzing ecological experiments (Schielzeth & Nakagawa 2012) and are frequently used when analyzing observational data in ecology.

Random effects account for clustering in data via the error structure of the model (Bolker *et al.* 2009; Gelman & Hill 2006), rather than estimating cluster means as part of the data generating process of a model (i.e., via fixed effect for each cluster's mean, using the terminology of the mixed models literature). This results in gains in efficiency (i.e., costing fewer degrees of freedom). Further, as random effects are assumed to be drawn from a common distribution, they have benefits for analyses of unbalanced samples as well as regularizing of cluster means (i.e., shrinkage, drawing them towards the grand mean, see Efron & Morris 1975).

For these reasons, Ecologists conducting a study akin to our snail-temperature example would likely gravitate towards a mixed effects model to account for variation between sites in snail abundances, using a mixed effects model design such as:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \delta_i + \epsilon_{ij}$$
$$\delta_i \sim \mathcal{N}(0, \sigma^2_{site})$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (3)$$

What is new here relative to eqn. 2 is δ_i, the site-specific deviation at site *i* from our common intercept, β_0, due to random variation, assumed to follow a normal distribution. As we will see, because this is a random effect, if site is correlated with temperature, we cannot resolve the problem of OVB with this model.

What assumptions is a random effects design making when it comes to omitted variables bias?

Why does the above model not control for omitted confounders via its site effect? Why do mixed effects designs produce incorrect results in the face of omitted confounders (i.e., a statistically biased estimate of the causal effect)? To understand this problem, remember that, when we model random effects, we are not modeling group means *per se* (Robinson 1991). Rather, we are modeling correlation in our error structure due to clustering in our data (Bolker *et al.* 2009; Schielzeth & Nakagawa 2012; Wooldridge 2010). The coefficient estimates of the causal variable of interest are unaffected by including or not including a random effect (we recommend you try this with any demo data set you have lying around). This difference – modeling error instead of modeling means *per se* – results in many of the above benefits, but also introduces one new assumption not often considered – a variation on the assumption of endogeneity we call the **Random Effects Assumption**. This assumption states that the random effects, which are part of the error term, do not correlate with any covariates in the regression (Antonakis *et al.* 2021; Wooldridge 2010).

We can see how this plays out with our earlier example of endogeneity in a system where we wished to know the effect of nitrogen on plant biomass, but nitrogen itself was driven by soil characteristics in different fields which themselves also affected plant biomass. We could have created a model with nitrogen as our causal variable of interest influencing biomass and made field a random effect. This model would be result in an endogeneity problem, however. When we estimate the nitrogen effect, the effect of soils characteristics differing by fields would not be accounted for, as in this model nitrogen is assumed to be exogenous despite it actually being endogenous, affected by fields. Nitrogen is correlated with the random effect, but this is not controlled for in the model. Any time a predictor is correlated with random effects, a statistical model will have an endogeneity problem. It is a violation of the Random Effects Assumption.

Bringing this back to our snail and temperature example, in a mixed model in equation 3, the random effects of site are part of the error term and assumed to be uncorrelated with temperature for the random effects estimator to be unbiased (Schielzeth & Nakagawa 2012; Wooldridge 2010). In equation 3, while site is incorporated into δ_i the effect of temperature on snails is not causally identified and this estimator is biased due to the violation of the Random Effects Assumption; in short, estimates of β_1 will be wrong.

The case we describe above will be common in many ecological analyses when a causal variable of interest varies at the site level in a way that is confounded with other drivers occurring at the site level, i.e., x_{ij} and u_i are correlated. This correlation violates of the Random Effects Assumption such that a random effects estimator will be biased.

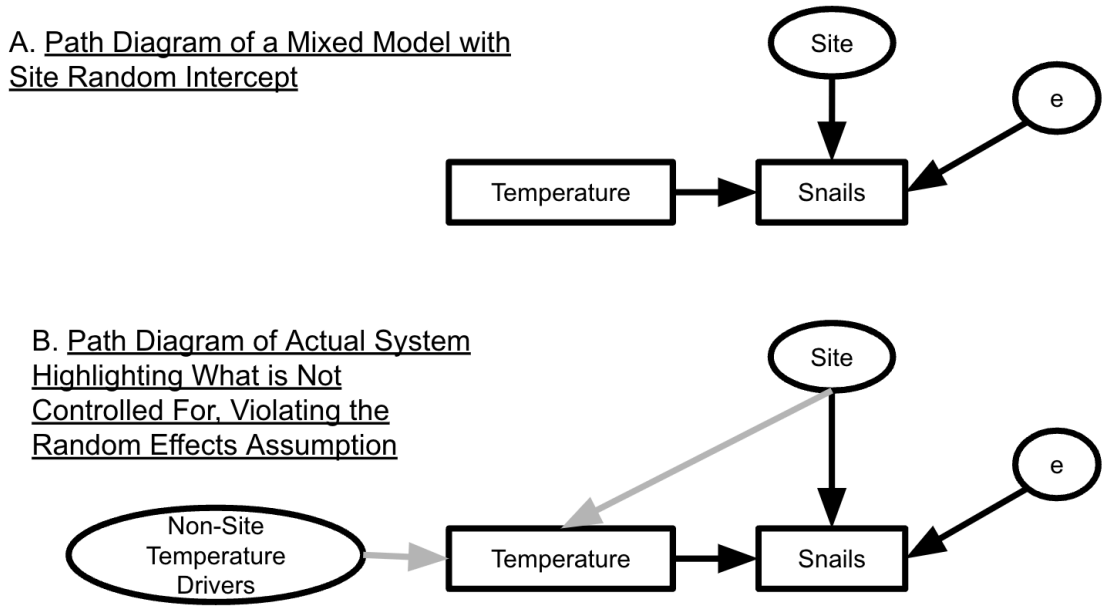

Figure 5. The system assumed to be underlying a mixed model (A) versus the true system in (B). The mixed model in **(A)** assumes that there are no site-level drivers of temperature. Thus, it does not account for the correlation between site and temperature. Thus, the effect of temperature on snails is confounded by any correlated site-level drivers that correlate with temperature at the site level. This contrasts with the actual system, shown in **(B)** in which temperature is determined by both site-level drivers and non-site-level temperature drivers. This difference between the diagrams in panels (**A)** and (**B)** is what leads to bias in coefficients from a mixed effects model. When we fit a mixed model, the gray path in (**B**) is not included, creating an endogeneity problem and violating the Random Effects assumption.

We can see more clearly how a mixed effects model would violate the Random Effects Assumption using a causal diagram in Figure 5a. In essence, site effects here are site-level residuals drawn from a normal distribution. They represent all other abiotic and biotic forces happening at the site level, but they also are assumed to all be uncorrelated with temperature at the site level. However, given the information in Figure 3, we know that this is not accurate, so the key assumption for an unbiased estimator is violated. If we were to take a step back and think about our analysis goals and our causal understanding, again representing unmeasured quantities in ellipses, what we have is more like Figure 5b. Here, while a random site effect would be

wonderful in terms of all the benefits discussed above, we would need to remove the effects of site-level confounders to use it – which is not done with the mixed effects model design above, as shown in Figure 5a. This example illustrates the difficulty in satisfying the Random Effects Assumption. More generally, we posit that satisfying this assumption is often quite difficult in Ecology – particular in observational data that spans environmental gradients – yet how badly this assumption is violated the is not well explored or acknowledged widely enough. We need solutions that do not produce biased results due to easily violated assumptions.

Enter the Econometric Fixed Effects Design

The Econometric Fixed Effects Design represents a familiar starting point for many ecologists who are used to using categorical variables in ANOVA and ANCOVA (e.g., Gotelli & Ellison 2012). Before getting into some admittedly confusing language, we note that, for Ecologists, this design is just using a categorical variable for cluster. The approach is that simple. To get further into the weeds, here we use Fixed Effect in two senses of the phrase to describe this model. The first is the use of the term "fixed effect" is drawn from the econometrics literature, where it refers to attributes of a system (e.g., site, plot, or year) that vary by cluster (i.e., a within cluster intercept) that are encoded in models as dummy variables. In Ecology, this as a categorical predictor representing site, block, or other descriptor of how our data is clustered. In our snail example, would be a site-level time-invariant categorical variable acting as a stand-in for recruitment. We also use "fixed effect" in the language of the mixed model literature – i.e., that the cluster means are estimated as part of the data generating process of the model, not as part of the random error component. Unfortunately, there are many uses and definitions of "fixed effect," leading to a wealth of confusion with different uses of the term across fields (Gelman & Hill 2006). We hope to not add to the confusion.

Recognizing that confounding variables vary at the cluster-level (e.g. site), and thus by removing the effects of clusters we remove the effects of our confounding variables, we have two options to control for confounding and OVB. First, we can use a bit of algebra known as the **within transformation** or **fixed effects estimator** (Bell *et al.* 2018; Wooldridge 2010) and has some similarities to within-subjects centering in Ecology (van de Pol & Wright 2009). To illustrate, we manipulate the following equation:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij} + u_i$$

(4)

where x_{ij} is our casual variable of interest, and the error term is composed of idiosyncratic (random error), ϵ_{ij}, and u_i , which represent differences across sites *i* including unmeasured confounding variables. To remove the effect of site-level confounding drivers, we transform the data by subtracting this average value $\bar{y}_i$ from both sides across all years. On the right-hand side, we can expand this to subtract $\beta_0 + \beta_1 \bar{x}_i + \bar{\epsilon}_i + \bar{u}_i$ which leads to a transformed model.

$$y_{ij} - \bar{y}_i = \beta_1 (x_{ij} - \bar{x}_i) + (\epsilon_{ij} - \bar{\epsilon}_i)$$

(5)

Using simple algebra, we have removed the confounding influence of time invariant, confounding variables for each site, whether or not they were observed. To achieve the same effect as this group means transformation (see Fig. 6A for a causal diagram), we could instead use a design with a categorical or so-called dummy variables for each cluster (i.e., a 0/1 encoding for each cluster, known as an econometric fixed effect). We can represent this as a site effect in a causal diagram (Fig. 6B). This design will control for omitted variable bias from site-level observed and unobserved confounding variables and produce identical results to the preceding model for β_1 (Angrist & Pischke 2008; Wooldridge 2010). This model can be implemented either by incorporating the dummy variables (x_{2i} coded as 0/1 for each site) and site effects (λ_i) or with just the site effect alone – i.e., means model notation (Gelman & Hill 2006). We present this without using treatment contrasts (i.e., with β_0 as a reference level and $\lambda_i$as the deviation from the reference level) for clarity.

$$y_{ij} = \beta_1 x_{1ij} + \sum \lambda_i x_{2i} + \epsilon_{ij}$$

$$= \beta_1 x_{1ij} + \lambda_i + \epsilon_{ij}$$

(6)

Note that unlike random effects in a mixed model design, λ_i is not constrained to be drawn from any predefined distribution.

562 Returning to our snail example, with site as an econometric fixed effect as in equation 6,
563 we can control for different sites having differences in their levels of recruitment or other
564 omitted variables that are also correlated with temperature – whether those confounding
565 variables were measured or not (see Fig. 6b). Thus, this design allows us to relax the strong
566 assumption that all confounding variables are observed and measured so that we can interpret
567 β_1 as causal, provided other assumptions are met (see Discussion). For ecological examples using
568 this design, see Larsen (2013), Dee et al. (2016), Dudney et al. (2021), Ratcliffe et al. (2023),
569 and Dee et al. (2023). We note that accounting for serial correlation, heteroskedasticity, and
570 clustering of the error, such as through cluster robust standard errors, are likely important for
571 both approaches for inference (Abadie *et al.* 2017; see Box 4 and Cameron & Miller 2015). We
572 return to a discussion of standard errors used for inference in Box 4.

573 The fixed effect design has some drawbacks, despite its simplicity and its strength in
574 controlling for both observed and unobserved confounding variables. First, while fixed effect
575 estimators make much weaker assumptions about confounding variables, these estimators are
576 inefficient compared to random effects. For each fixed effect (each site in our example), we
577 estimate a separate coefficient and thus are estimating more parameters and eating up degrees of
578 freedom. We therefore need a larger sample size to achieve the same level of precision for our
579 estimate using fixed effects versus random effects, presenting a bias-variance trade-off (Bell *et*
580 *al.* 2018). In comparison, random effects are more efficient, costing fewer degrees of freedom to
581 estimate as we assume cluster means follow from a distribution (i.e., estimating a grand mean
582 and variance), rather than directly estimating a separate coefficient for each cluster mean with no
583 relationship to any other cluster mean. With this efficiency can come an improvement in the
584 estimates of *precision* for coefficient estimates for our causal variable of interest (Gelman & Hill
585 2006) relative to fixed effects cluster means; however, this efficiency does not guard against or
586 reduce *bias*. Thus, in the case of omitted variable bias with the goal of causal inference, the fixed
587 effects framework is still preferable over a mixed model. Fixed effects make weaker assumptions
588 about our ability to observe, measure, and control for confounding variables compared to random
589 or effects.

590 Finally, with the fixed effects approach, we lose information about between-site
591 variation, including gradients between sites that may be of interest. This variation is absorbed

into the fixed effects. These gradients, while confounded with other variables, could be the focus of some research questions which cannot be easily addressed using fixed effect model designs.

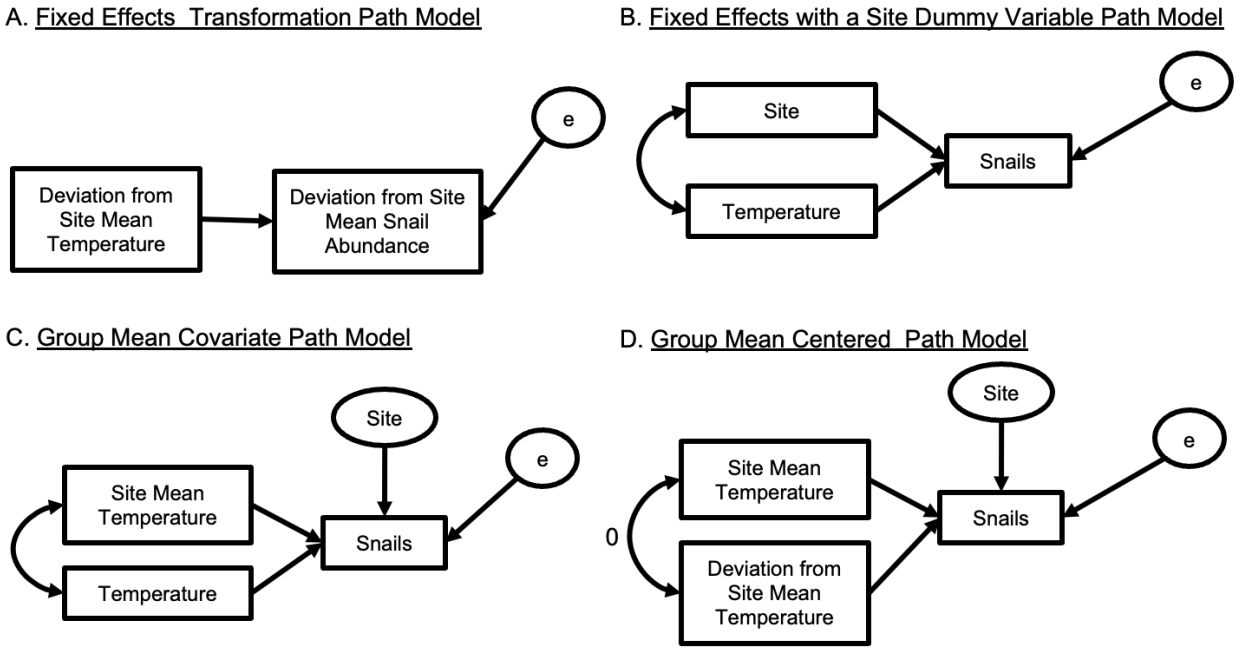

Figure 6. Directed Acyclic Graph representations of different statistical models handling omitted variables in the text. (**A**) and (**B**) show two variations on the econometric fixed effect model design corresponding to equations [5] and [6] respectively. (**C**) represents the group mean covariate design in equation [7] and (**D**) represents the group mean centered design from equation [8].

Group Means for Efficiency, Inference, Fun, and Profit

What if we care about the between-site variation and comparisons across site are central to our question? To study between-site variation and mitigate the loss of efficiency from the fixed effect design, we can instead use **correlated random effects designs** (using terminology of Antonakis *et al.* 2021)**.** Correlated random effects use group means of our causal variable of interest to control for confounding variables. For every cluster (e.g., each site, year, or region), researchers calculate a group mean of the causal variable of interest (e.g. average temperature of a site) and include it as a group-level predictor. These group means of our causal variable of interest control for the effects of confounders at the cluster (e.g., site) level by acting as a proxy for confounders. Using group means of our causal variable also enables us to estimate a coefficient for between-cluster effects (e.g., between site) in our causal variable of interest, although these coefficients contain a combination of causal and confounded effects.

In Econometrics, one CRE model design is the **Mundlak Device** (Mundlak 1978) and has many extensions (e.g., Wooldridge 2021). For clarity, we term it a **Group Mean Covariate** design. For the group mean covariate model design, we use the following equation:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 \overline{x_i} + \delta_i + \epsilon_{ij}$$
$$\delta_i \sim \mathcal{N}(0, \sigma^2_{site})$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \quad (7)$$

where $\beta_2 \overline{x_i}$ accounts for the effect of cluster-level confounders and δ_i is a random effect of that cluster (i.e., site). We can see what this looks like as a DAG in Figure 6c. From this diagram, we see that the site mean temperature is controlled for in estimating the temperature effect. The mean temperature of a site is estimated while controlling for each measured temperature.

The site mean temperature coefficient, called a **contextual effect** (Antonakis *et al.* 2021) in the Group Mean Covariate design, shows how changing the mean temperature of a site – and all properties that correlate with site mean temperature – would affect snail abundance were the temperature within a plot to stay the same. For example, *if our plot was 10 degrees C, what would snail abundance be if said plot was in a site with an average temperature of 5 degrees C versus 20 degrees C*? If the contextual effect is 0, then we can conclude that a simple mixed model would have sufficed and that omitted variable bias was not a problem in this particular analysis (Antonakis *et al.* 2021).

Finally, we must account for correlation in the error term due to between-site variability that does not come from confounders when estimating standard errors used for statistical tests. As our group mean contains the variation due to cluster-level confounded variables, we can now use a random effect to account for cluster-level (i.e., site-level) variability. This random effect now encompasses variation due to cluster without variation due to confounders. By using a random effect for cluster while spending one degree of freedom to estimate a coefficient for our group mean predictor, we gain significantly in efficiency over the fixed effects model.

The Group Mean Covariate model design will run into problems, however, if the correlation between our causal variable of interest and its cluster-level mean is too high. To solve this, we can use a design that transforms our causal variable to remove this correlation. We accomplish this with a **Group Mean Centering** design, which subtracts the cluster-level mean

from the causal variable of interest. In our example, this would mean that for each year at each site, we subtract the observed temperature from that site's average temperature across the whole time series. Figure 6d shows the DAG for this design and the similarities and key differences with the previous designs. After this transformation, the variation we use in cluster-level anomalies (i.e., within cluster variability) as our predictor variable alongside a cluster level mean as follows:

$$y_{ij} = \beta_0 + \beta_1\left(x_{ij} - \overline{x_i}\right) + \beta_2\overline{x_i} + \delta_i + \epsilon_{ij} \quad (8)$$

The coefficient of the site mean of temperature, β_2, is the between-site effect of a driver of interest and confounders, and the anomaly from the site mean coefficient, β_1, is the within-site temperature effect. Thus, equation 8 decomposes our causal variable of interest into between- and within-cluster terms, an approach already in use in ecology (van de Pol & Wright 2009). Here, the interpretation of β_2 is different than in the Group Mean Covariate design. β_2 for our snail example is now a **between estimator** of the combined effect of moving across gradients in temperature and correlated drivers between the sites. If $\beta_2 = \beta_1$, omitted variables are not meaningfully influencing snail abundances; both the between and within site differences are due solely to temperature or multiple confounders have perfectly cancelled one another out.

The Group Mean Covariate, Group Mean Centered, and Fixed Effects designs all differ in structure, but they will yield the same point estimates of β_1 under most conditions and balanced data, as they all rely on within-site variation in temperature (see simulations below and Wooldridge 2010). Thus, one might ask: *which statistical design should I use*? This decision depends on the structure and size of one's data (e.g., how many coefficients do you have the power to estimate given your sample size) and the question of interest (e.g., are you interested in between-site differences?). For example, do you have many sites and are only interested in the causal effect of temperature? Fixed effects design. Do you want to know how plot-level snail abundance would change if the average site temperature changes, but plot temperature stays the same? Group Mean Covariate design. Do you want to understand the effects of temperature while examining the net effect of many variables shaping between-site gradients? Group Mean Centered design. Each design can further be extended to cases where the magnitude of the causal

669 variable of interest's effect is moderated by the level of confounding variables (i.e., an
670 interaction or "heterogeneous" causal effect, see Box 2).

671 *What a Difference Differencing Makes*

672 Our examples thus far have focused on confounding variables that are unobserved and
673 vary across space (i.e., between sites). We have not yet discussed omitted confounding variables
674 that differ across time. In the case of omitted confounders varying solely across time and not
675 space (e.g., sites vary randomly in recruitment across space, but year-to-year regional variation
676 in recruitment is correlated with year-to-year regional variation in temperature), we can extend
677 the frameworks presented above, using years rather than sites as clusters as in our example. If
678 time-varying confounders are uniform across sites (i.e., are additive with spatial confounders),
679 then we can use an econometric fixed effect of time and an econometric fixed effect of space (a
680 two-way fixed effect or TWFE model design, Wooldridge 2021) and extensions (Roth *et al.*
681 2023) or a site-average of predictors and a time-average of predictors (a Two-Way Mundlak
682 model design; Wooldridge 2021).

683 If, however, temporal confounders differ by site, we need a more general solution. If
684 omitted confounders vary spatiotemporally, we can extend our previous framework further using
685 the same principles (see Box 3 and Dee *et al.* 2023). If, however, temporal confounders merely
686 vary in strength from one site to the next, the **first and second difference** statistical model
687 designs provide easy solutions. These statistical model designs deal with both spatial
688 confounders and site-varying temporal confounders. To illustrate, consider extending our
689 example so that, in addition to site-level oceanographic recruitment effects, the abundance of
690 snails is influenced by coastal development over time at each site (Fig. 7A). However, rates of
691 development are not the same across all sites. As such, separating the effect of local coastal
692 development from the effect of local temperature variability on snail abundance is difficult. We
693 can see this in a small modification to the dynamics of our system from eq. 1:

694 $$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma w_i + \lambda_i j + e_{ij} \quad (9).$$

695 In this scenario, there are both site-specific confounders, represented by γu_i and temporal
696 confounders. All terms are the same as in eqn. 1, except we now have a temporal confounder that
697 varies by site (λ_i) that drives change in snails over time, (j). If there is also a trend in temperature

over time (e.g., climate change), our estimation of β_1 in any model that did not include our temporal-confounder would suffer from Omitted Variable Bias. One solution to this scenario is to fit a model with an econometric fixed effect of site to account for spatial confounders and a site by time effect - a separate temporal trend for each site - to account for the confounding variables that are site specific and vary through time at the site level (e.g., Dee *et al.* 2016). This would lead to twice the number of parameters as the number of sites, however, assumes linear trends in each site, and might not be a feasible given data and power constraints. Or, with more data, and replicated samples within years, one could have a categorical site by year effect. See Box 3 for more on these more difficult problems.

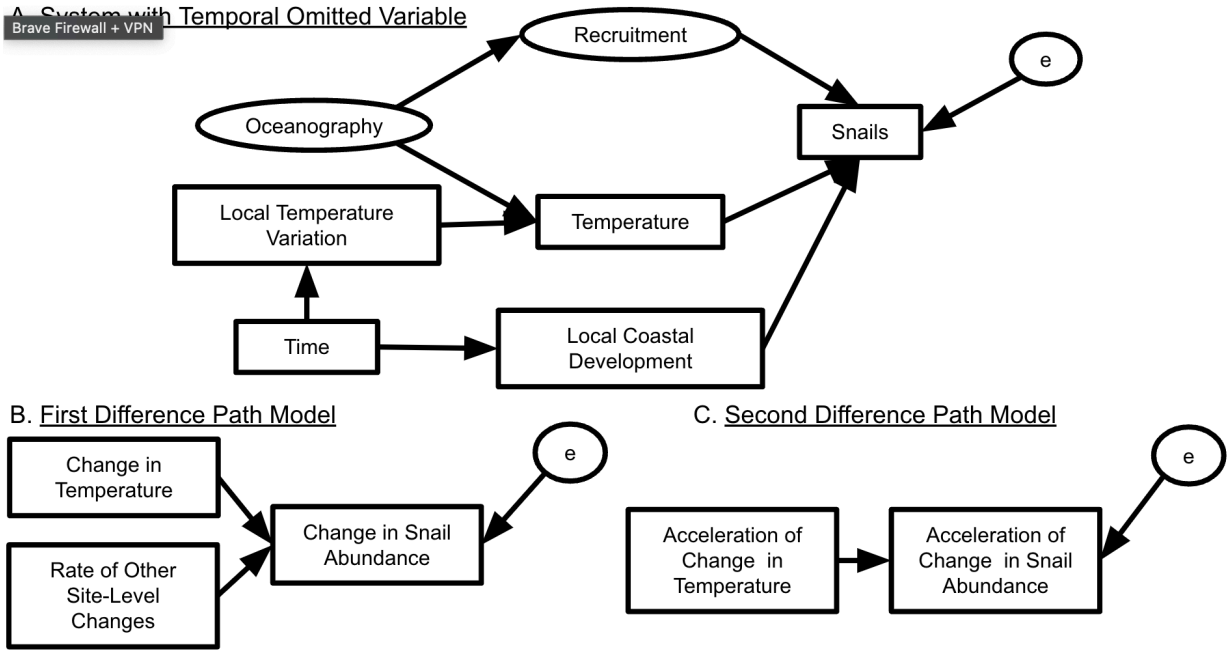

Figure 7. Directed Acyclic Graphs representing the way that different designs handle spatial and temporal omitted variables. (A) shows the true data generating process of this hypothetical system, with not only recruitment confounded with temperature, but also with local coastal development increasing over time alongside local temperature also increasing over time. Thus, we have both temporal and spatial confounding variables. In **(B)** we see the causal diagram represented by the first differences model, where due to differencing we separate temporal confounders from the signal of change in temperature. In **(C)** the second difference model, temporal confounders are removed, and now we are estimating the effects of acceleration in change in temperature on acceleration in change in snail abundance. Also note that the fixed effects design in equations [5] and [6] can be extended to a "two-way fixed effects" design to control for both site and year.

Fortunately, there is a simple solution similar to the fixed effects transformation: **temporal differencing**. For each time point in our data, if we subtract the previous time point, we produce a model evaluating the relationship between change in our response variable versus change in our causal variable of interest. Like the fixed effects transformation, the confounding

effects of site-level omitted variables that do not have a temporal trend (e.g. are time invariant over the study period) are eliminated. After the transformation, λ_i remains as a term to be estimated with an econometric fixed effect, and it will recover an estimate of the trend for each site. This approach has the added benefit of sweeping up other unknown site-level trends into our estimate of λ_i. Our first difference model design, represented as a path diagram in Figure 7B, translates to the following with means model notation:

$$\Delta y_{ij} = \beta_1 \Delta x_{1ij} + \lambda_i + \Delta \epsilon_{ij} \quad (10)$$

Here β_1 estimates the effect of temperature as before with λ_i estimating the site-level trend of other drivers. Note that confounders at the site level, γw_i above, are removed algebraically in this design. If there is no temporal trend in temperature, and as such there is no correlation with other site-level trends, we *could* use random effects for the site term. We caution, however, that this adds back the random effects assumption which is unlikely to be met. Note that if the time between sampling events is unequal across sites, we can divide change by time between samples to model change per unit time. Finally, if we are uninterested in site-specific trends, we can calculate the second difference $\Delta^2 y_{ij} = \Delta y_{ij} - \Delta y_{i,j-1}$ which eliminates the need to estimate λ_i. Note that β_1 in this second differences statistical model design model is estimating the relationship between acceleration in change in temperature and acceleration in change in snails. This formulation eliminates the need to make assumptions about multiple forms of confounding at the cost of two time-points worth of data.

Using either temporal differencing design has several advantages. We again remove the effect of omitted confounders at the site level. We also control for or remove the effects of temporal confounders at the cluster level that have similar trends to our causal variable of interest. Thus, our estimate of a temperature effect is again causally identified. Our analysis is robust to omitted variable bias from two sources of unknown confounders. In contrast, it requires variation within sites. Like the econometric fixed effect design, differencing relies on *within site* variation as it removes the confounded *between site* variation. Further, it cannot be estimated if the causal variable of interest is time invariant; it yields imprecise estimates if the variable changes little over time. The final main drawback of the differencing approach is the reduced sample sizes; we lose observations from one or two time periods which can become problematic

for small numbers of sites relative and only a few years. This reduction in sample size reduces power and can lead to less precise standard errors, especially in the case of the second difference design. However, this reduction in sample size could be outweighed by more robust and flexible control over confounders, both measured and unmeasured. There are other cases where how temporal and spatial confounders affect a system can create a need for more careful designs – interactions between the two, non-linearities, spatiotemporal confounders, and more. We outline how other designs can also be extended these cases in Box 3.

Box 2: A Difficult Slope: Omitted Variables that Cause Variation in the Magnitude of the Causal Effect

An omitted confounder might not merely contaminate our estimate of a causal effect but can also lead to model misspecification in the form of missed heterogeneity in the causal effect. This occurs when the causal effect of our variable of interest depends on the level of the confounder itself (i.e., it modifies the causal effect – an interaction effect). In our example, thermal effects on snail abundance could depend on levels of recruitment because dense aggregations of intertidal organisms are often better at retaining water and thus resisting desiccation or other forms of thermal stress (e.g., Silliman *et al.* 2011). This dependence is problematic if we have not measured recruitment. In a naive mixed model, we might incorporate this heterogeneity as a random slope. As before, however, the random effects assumption is violated, so a random effects estimator will be biased. To deal with the problem of omitted variable bias in this case, we present two solutions. First, we can use a fixed effects design and include an interaction term between the site dummy variable and our causal variable of interest, allowing us to estimate site-specific temperature effects. Given that we now have site-level slopes, the number of parameters can blow up, leading to this approach being highly inefficient and not advisable for small sample sizes. Instead, we could use correlated random effects approaches with an interaction between the group mean and our causal variable of interest. For example, for a group mean covariate design (i.e. Mundlak device), we would use the following equation:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 \bar{x_i} + \beta_3 x_{ij} \bar{x_i} + \delta_i + \epsilon_{ij}$$

This design allows us to examine how site-level confounders – known and unknown – can lead to variation in the effect of our causal variable of interest. It can also show that they have no

effect if the estimand for β_3 is not different from 0. We could use a similar model for the group mean centered design if deemed appropriate. If we suspect that the magnitude of the temperature effect varied with other non-confounded covariates, we could instead use a random slope. In general, models with interactions representing moderators can provide powerful insights into both the effect of the causal driver of interest as well as how those effects vary.

Comparison of Approaches

To demonstrate the utility the preceding solutions, and the consequences of not using them, we fit a variety of models to simulated data based on a longitudinal study of snail populations at multiple sites based on Figure 3. For a single simulation run, we created a system as follows:

- Oceanography is a variable with a mean of 0 and standard deviation of 1.
- Site mean recruitment is -2 multiplied by the oceanography variable and then rescaled to have a mean of 10 individuals per plot (e.g., so it does not go negative). It is the same in a site across all years.
- Site mean temperature is calculated as twice the oceanography variable and then rescaled to have a mean of 15C.
- Site temperature in year t is determined by site mean temperature and additional variation, with a mean of 0 and standard deviation of 1.

Snail abundance at site *i* in year *t* is then determined in a given year as in Fig. 3, where snails are a function of recruitment, temperature and other drivers. We simulated data where the effect of temperature on snails is 1 as is the effect of recruitment on snails; the effect of other drivers varying with a mean of 0 and standard deviation of 1. We then simulate sampling 10 sites over 10 years. We provide the code to generate and results from 100 simulated data sets in Appendix A and at https://github.com/jebyrnes/ovb_yeah_you_know_me. We analyzed each simulation run using all the statistical model designs described above, compared to naive models with no site effect. We also included group mean covariate and group mean centered models with and without a random effect to demonstrate the role of a random effect in these models with respect to influencing parameter standard errors and handling unbalanced data. Appendix B and Supplementary Data 1 walk through the analysis of a single data set. For a more interactive

exploration of this and the full suite of simulated data and parameters, see the web applications written using R Shiny provided as Appendix C (for one simulated run alone) and Appendix D (with 100 or more replicate simulations exploring the distributions of parameters).

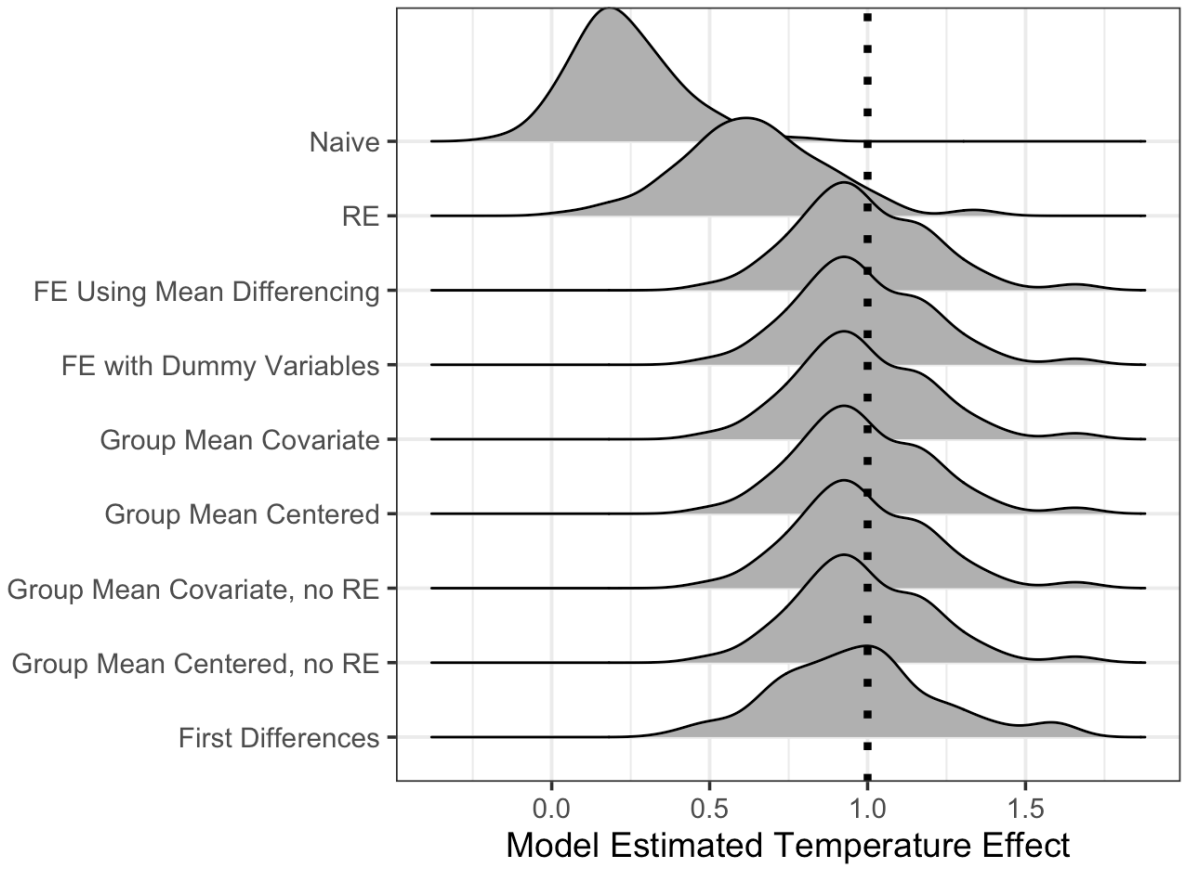

Figure 8. Distribution of point estimates of temperature effects from different models across all 100 simulations. The true effect size (= 1) is highlighted with a dotted line. The y-axis labels correspond to the Naïve model in equation 1, Random Effects model in equation 3, the Fixed Effects models in equations 5 and 6, the Group Mean Covariate models in equation 7and the Group Mean Centered models to equation 8, and the First Differences model to equation 10. The Naïve and random effects models produce biased coefficient estimates on average, in contrast to all other methods.

Broadly, our simulations show that the random effects (RE) model – what ecologists typically use – is consistently biased in these simulations. The point estimates from RE model are well-below the estimates from both the other designs and the true effect size (Fig. 8, Table 1). Further, not only is the estimated coefficient of the RE model always biased compared to other estimators in our simulations, it is more often within 2SE of 0 (i.e., would fail to reject a null hypothesis) in comparison to all other model designs. More worrying, in the majority (54%) of simulations, the 95% confidence intervals of the RE model do not contain the true slope,1, of the

temperature effect (Table 1). Other than the naïve and random effects model, the other designs show similar estimates with balanced data (in Table 1). The first differences model underperforms with respect to its CI not containing the true parameter value relative to other designs, but it is not even close to the random effects or naïve model design. However, the relative performance of first difference versus econometric fixed effect depends on the structure of data, whether it has more time periods, or more units (Wooldridge 2010), which likely explains this discrepancy.

Table 1. Summary simulation results. Mean and SD of point estimates of temperature effects from different models in the first two columns. Fraction of simulated runs where the mean +/- 2 SE of the temperature effect either overlapped 0 (i.e., high likelihood of committing a type II error) or did not contain the true effect of temperature in the final columns. Models are as in Fig. 8.

Model Type	Mean Estimate	SD Estimate	Fraction Sims where 95% CI Contains 0	Fraction Sims where 95% CI does Not Contain 1
Naïve	0.231	0.165	0.56	0.99
Random Effects (RE)	0.640	0.232	0.08	0.54
FE† Using Mean Differencing	0.985	0.215	0.00	0.05
FE† with Dummy Variables	0.985	0.215	0.00	0.05
Group Mean Covariate	0.985	0.215	0.00	0.05
Group Mean Centered	0.985	0.215	0.00	0.05
Group Mean Covariate, no RE	0.985	0.215	0.01	0.04
Group Mean Centered, no RE	0.985	0.215	0.01	0.04
First Differences	0.971	0.259	0.01	0.12

†FE = econometric fixed effects

Additional explorations show that, in line with the benefits of random effects in mixed models, a site-level random effect is crucial for Group Mean Centered or Group Mean Covariate models when either the study design is unbalanced or there is site-level variation that is uncorrelated with temperature (for more details, see Appendix A). If our simulation has no site-level variation other than temperature and our confounder, a random effect does not improve either models' ability to estimate the effect of our causal variable of interest with respect to bias or precision. This assumption is unrealistic for most real data sets, however. As such, we highlight the need for a site level random effect with either of these two designs or a clustered

standard error. For estimating standard errors, in general, we urge researchers to incorporate random effects or clustered robust standard errors as needed to accommodate clustering in the error, per the study design, recognizing the tradeoffs of using both and appropriate context (*reviewed in* Oshchepkov & Shirokanova 2022).

Box 3: Reality Bites: Coping with spatiotemporal omitted confounders

Spatiotemporal confounding variables – those that are site (or plot) specific and vary through time – pose challenges, and the solutions require more thoughtful study and statistical model design. To illustrate, we consider a scenario where recruitment, a confounding variable related to both snail abundance and temperature, is not static through time but instead varies by site and year (as in a realistic case). For example, sites that experience strong cold-water pulses in a year also experience unusually snail high recruitment in those same years due to oceanographic drivers. The sampling designs for coping with spatiotemporal omitted variables are based on the same principles as cross-sectional and longitudinal sampling, only now we combine the two to include plots within sites that are sampled through time.

With longitudinal data with multiple plots sampled within a site through time, we can flexibly control for spatiotemporal confounding at the site level by extending the two-way fixed effect designs discussed above. We can add a site-by-time fixed effect, η_{ij} , to our model, in addition to a fixed effect of plot, λ_k, where *k* is a fixed plot within site resampled over time (see below for a discussion of fixed versus re-randomized plots). This produces the following means model:

$$y_{ijk} = \beta_1 x_{1ijk} + \lambda_k + \eta_{ij} + \epsilon_{ijk}$$

From this equation, we can see that λ_k captures time invariant plot-level confounding effects while η_{ij} captures the effects of spatiotemporal omitted variables at the site by time level. Note, there could be additional spatial or temporal only confounders. This design sweeps their effects onto the spatiotemporal term.

In small datasets, the above model design can consume degrees of freedom rapidly. In datasets with insufficient power, we can instead use the correlated random effects (e.g., a variation on the Two-way Mundlak model design *sensu* Wooldridge 2021) which are more

874 efficient. Correlated random effect use site-year means ($\overline{x_{ij}}$) and plot means ($\overline{x_k}$) for the entire
875 survey to control for spatiotemporal and plot confounding respectively:

876 $$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + \beta_2 \overline{x_k} + \beta_3 \overline{x_{ij}} + \delta_k + \delta_{ij} + \epsilon_{ijk}$$

877 Here the δ_k and δ_{ij} terms are random effects for plot and unique site-time combinations
878 respectively.

879 When sampling to handle spatiotemporal confounders, should plots within sites over time
880 be permanent or randomly placed each year? The above models assume permanent plots, so we
881 can eliminate confounding variables at the plot-level that is time invariant over the study period.
882 For this reason, permanent plots help us cope with within-site OVB issues and have higher
883 power to detect change over time (Urquhart & Kincaid 1999). Logistically, however, permanent
884 plots within sites might not be possible. As such, the above models can be modified to drop plot
885 effects; however, they would then assume that there are no confounding differences across plots
886 and could have lower power to detect effects of drivers. We emphasize that the choice of fixed or
887 random plot placement with these designs is a balancing act, however, as fixed plots can lead to
888 a lower sample size due to logistical considerations in many environments, and direct readers to
889 other explorations of this topic (e.g. Gomes 2022).

890 Finally, without variation within sites as well as through time – e.g., plots within sites
891 resampled over years – we cannot include a site by year effect as in the above models. We can
892 attempt to use site-level time trends (e.g., as linear or polynomial trends) or trends generated
893 from Generalized Additive Models (Wood 2017) to approximate site-by-time effects see Dee *et*
894 *al.* 2016a), but, this requires knowledge of how the confounder varies at sites over time and
895 extensive testing for robustness to these assumptions. In the many cases this is not possible or
896 inadvisable given the likelihood of creating incorrect causal inference. In those cases, without
897 multiple plots per site over time, "nothing to be done" (Beckett 1954).

898 In general, we urge caution when dealing with spatiotemporal omitted variables, and
899 careful use of causal diagrams to ensure that we are controlling for a confounder at the relevant
900 spatiotemporal scale. This topic is one that that deserves far more exploration in Ecology. More
901 from other disciplines on this tricky class of problem and approaches can be found from

literature outside of the scope of this paper (Athey & Imbens 2017; e.g., Ferraro & Hanauer 2014; Oster 2019).

Discussion

We hope that our introduction to statistical and sampling designs that address the problem of omitted variable bias and causal inference with observational data has opened up new possibilities for our readers. At the core of these and other causal inference techniques is building an *a priori* causal model of how a system works and clarity in assumptions required for a causal interpretation of an estimated effect. *A priori* models can also guide your choice of sampling designs and statistical methods for causal questions. Further, the techniques presented here for addressing omitted variable bias are well within the standard statistical abilities of most modern ecologists (see Appendix B for implementation). The inferences made from designs that can better control for unobserved confounding variables can be profound for our ability to understand biological systems, as seen in our toy example.

We hope that Ecologists can see the concepts presented here as part of a generalizable approach to handling confounding variables using clustered or hierarchical data. While we use sites and years, the same concepts apply to studies with cohort effects, individual effects, or other lower levels of clustering as well as to larger-scale studies with not just sites and years but regions and decades. The general suite of approaches remains the same, and potential confounding variables at these different scales can be identified in causal diagrams. Cross-sectional and longitudinal sampling designs are also generalizable beyond the simple case presented in our simulation example. For instance, one could adapt the above designs if temperature and recruitment varied through time at a regional rather than site scale (e.g., sampling plots within a single or many sites over many years to leverage spatial and temporal variation in temperature) or for spatiotemporal designs (*see* Box 3). Combining these sampling designs with others, such as a stratified random sampling design (Foster *et al.* 2018; Grafström & Lundström 2013; Kermorvant *et al.* 2019; Robertson *et al.* 2013; Stevens & Olsen 2004), will allow for the analyses that can both improve causal identification and also provide more precision in estimation over multiple environmental gradients. How to design a study to fully

account for confounders, however, will hinge on a causal structure of the system and a researcher's ability to be humble in the face of what they might not know.

Box 4: Clustered Robust Standard Errors: An Underutilized Tool in Ecology

While the focus of this paper is on bias not precision, many of the issues discussed overlap with issues of non-independence that could generate incorrect standard errors and statistical tests for inference. In light of that, we recommend the use of clustered robust standard errors. Clustered robust standard errors offer a flexibly way to accommodate clustered data, heteroskedasticity, serial correlation between time points, and other arbitrary correlation structures within the data (Abadie *et al.* 2017; Cameron & Miller 2015) yet are not commonly used in Ecology (but see examples in Dee *et al.* 2016; Dudney *et al.* 2021; and code in the appendices of Dee *et al.* 2023). While random effects, autocorrelation structures, and more, can address some of the same issues, clustered robust standard errors make weaker assumptions about the exact form of the correlation structures in the error, thus providing a simpler solution when the structure of the data are not of interest to the research question but important to account for in inferences. However, there are tradeoffs – weaker assumptions also mean less efficiency. Thus, we recommend looking at comparisons of approaches (e.g., Oshchepkov & Shirokanova 2022) or conducting sensitivity tests to check the robustness of inferences to choices of approach for modeling standard errors. A full discussion of clustered and robust standard errors is beyond the scope of this paper, and we refer applied researchers to the documentation for the '*sandwich'* package in R and other comprehensive reviews (e.g., Cameron & Miller 2015).

The important thing is to be transparent in how we deal – or do not deal – with the confounding variables. What are the assumptions you are making to interpret an effect as causal? Why did you control for some covariates and not others? Do you have a DAG or even a conceptual model of your system that might help a reader understand your thought process? If you are using mixed models, do you meet the random effects assumption, and why or why not? Have you evaluated your residuals to determine if you need to implement robust standard errors? Clarifying these types of decisions, even in a brief sentence if not a figure or full breakdown in a manuscript supplement (e.g., see Dee *et al.* 2023), will go far in terms of making your analyses

more transparent. This transparency will make the work easier to be built upon to advance science. On top of transparency, we also must be humble. We must accept that our models and knowledge are imperfect. Someday, someone will come along with a different approach that will produce different conclusions and yield new insights. This progression is part of the scientific process.

The approaches presented herein are not a panacea. They require assumption for causal inferences, as does any approach, including experiments (Kimmel *et al.* 2021). Some assumptions are shared with experiments: i.e., SUTVA – or the stable unit treatment value assumption which has two parts: 1) no interference or "spillovers" across units and 2) no multiple versions of or "hidden variations" in the causal variable of interest (reviewed in Kimmel *et al.* 2021). In addition, most of the statistical model designs presented here also include assumptions that expected effects are linear and additive (Imai & Kim 2021) and homogeneous across units and time periods. We have included some discussion of relaxing these assumptions via interactions (i.e., Box 2); however, there is a growing literature on estimating causal effects in these designs under more varied forms of heterogeneity and non-linearity (Callaway & Sant'Anna 2021; de Chaisemartin & D'Haultfœuille 2020; Goodman-Bacon 2021; Sun & Abraham 2021). Relaxing this assumption takes more thought and consideration of one's question of interest and the system dynamics from DAGs.

Further, all the approaches presented here make the parallel trends assumption. The assumption of parallel trends is most easily understood considering a binary causal driver of interest (i.e., if the driver is present or absent). It implies that, without driver being present, the *difference* in outcomes between different clusters (e.g., sites) after conditioning on covariates would be constant through time. This assumption is more likely met with fewer time periods (e.g., two time periods spanning before and after an impact, as in a before-after-control impact, or BACI design). The assumption can be tested in the pre-treatment period but is untestable after the treatment (for details, see Roth 2022). This assumption extends to continuous causal variables. There, we assume the response of interest across clusters would have followed parallel trajectories in the absence of a change in the causal variable and after adjusting for other observed covariates. The parallel trends assumption has come under a great deal of scrutiny recently (reviewed in Roth *et al.* 2023), particularly when changes in the causal variable of interest happen at different points in time across units (called "staggered treatments," see Baker

991 *et al.* 2022; Marcus & Sant'Anna 2021) and in the face of heterogeneous effects of causal
992 variables (for details, see Borusyak *et al.* 2023; de Chaisemartin & D'Haultfœuille 2020;
993 Goodman-Bacon 2021; Sun & Abraham 2021). This is a rapidly evolving literature, with many
994 proposed solutions (*reviewed in* Roth et al. 2023), including for heterogeneous effects (*see* Roth
995 *et al.* 2023), non-linear cases (Imai & Kim 2021), and continuous causal variables (Callaway *et*
996 *al.* 2021). Many of these solutions are already being implemented in standard software (*see* Roth
997 *et al.* 2023). Further, we suggest using the approaches reviewed here in concert with sensitivity
998 tests (Altonji *et al.* 2005; Cinelli & Hazlett 2020; Oster 2019; Rosenbaum 2002) and by
999 implementing multiple designs that make different assumptions to probe robustness of results
1000 (*see* Dee *et al.* 2023 for an ecological example).
1001 Finally, we emphasize that this paper provides an entry point into a broader,
1002 interdisciplinary literature on causal inference in observational data, longitudinal data analysis,
1003 and panel regression methods. Indeed, other methods, including quasi-experimental designs such
1004 as instrumental variables and regression discontinuity, can be used to eliminate omitted variable
1005 bias (see reviews and examples in Angrist *et al.* 1996; Arif & MacNeil 2022; Butsic *et al.* 2017;
1006 Dee *et al.* 2023; Grace 2021; Kendall 2015; Larsen *et al.* 2019; MacDonald & Mordecai 2019).
1007 Thoughtful uses of the front-door criterion – the use of mediators between a cause and effect that
1008 are unaffected by confounders – could also prove useful for ecology for identifying a causal
1009 relationship (Bellemare *et al.* 2024; Pearl *et al.* 2016); although, there are none to our knowledge
1010 in the Ecological literature yet. We urge ecologists, long grounded in experiments, to open
1011 themselves to writings in Econometrics, Epidemiology, Computer Science, Public Health, and
1012 other disciplines with rigorous approaches to causal inference in observational data. Embracing
1013 this transdisciplinary approach will enable us to enhance our knowledge of the tremendous
1014 advances in causal inference and explore questions currently beyond our reach. As an incomplete
1015 set of starting points for further reading, we recommend The Effect: An Introduction to Research
1016 Design and Causality (2021), Cunningham's Causal Inference: The Mixtape (2021), McElreath's
1017 chapters on causal diagrams in Statistical Rethinking (2020), Angrist and Pishke's Mostly
1018 Harmless Econometrics (2008), Morgan and Winship's Counterfactuals and Causal Inference
1019 (2015), Sloman's Causal Models (2005), and Pearl's Causal Inference in Statistics: A Primer
1020 (2016). We also suggest Ecologists interrogate the assumptions and interpretations of their
1021 experiments (Kimmel *et al.* 2021). Given how an experiment was designed and run, are its

results causally valid with respect to the purported mechanism? It is high time to critically interrogate how to get the robust causal inferences needed to grapple with our rapidly changing world.

Conclusion

"Correlation does not equal causation" rings in many of our heads from our Biostatistics 101 courses. One main reason behind this message is the specter of Omitted Variable Bias from unmeasured confounding variables. This fear has impeded the use of observational data for causal inference in Ecology for much of its recent history. We hope this review can lift some of that fear and, armed with the tools introduced here and knowledge of a literature beyond this piece, we can move forward as a discipline. With a massively growing volume of observational data, problems at continental to global scales demanding rapid answers, and now, new arrows in our Ecological data analysis quiver, we look forward to seeing the studies and insights from the next generation of Ecologists.

Acknowledgements

We thank the NCEAS LTER working group, "Scaling-up productivity responses to changes in biodiversity," for initiating the conversations and feedback that led to this paper, supported by the NSF Long-Term Ecological Research (LTER) Network Communications Office and DEB-1545288. This work was partially supported by the National Science Foundation as part of the PIE-LTER Program (award #1637630), Woods Hole Sea Grant, and the Stone Living Lab to J.B.; and by NSF OCE #2049360, NSF CAREER #2340606, and NASA BioSCape award #80NSSC22K0796 to L.E.D. We thank S. Elmendorf, B. Hobart, I. Rosenthal, R. Stevenson, A. Carter, and the UMB Stats Snack for helpful conversation and comments on early drafts of the manuscript.

References

Abadie, A., Athey, S., Imbens, G.W. & Wooldridge, J. (2017). *When Should You Adjust Standard Errors for Clustering?* (Working Paper No. 24003). Working Paper Series. National Bureau of Economic Research.

Abdallah, W., Goergen, M. & O'Sullivan, N. (2015). Endogeneity: How Failure to Correct for it can Cause Wrong Inferences and Some Remedies. *Br. J. Manag.*, 26, 791–804.

Altonji, J.G., Elder, T.E. & Taber, C.R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *J. Polit. Econ.*, 113, 151–184.

Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996). Identification of Causal Effects Using Instrumental Variables. *J. Am. Stat. Assoc.*, 29.

Angrist, J.D. & Pischke, J.-S. (2008). Mostly harmless econometrics. In: *Mostly Harmless Econometrics*. Princeton university press.

Antonakis, J., Bastardoz, N. & Rönkkö, M. (2021). On Ignoring the Random Effects Assumption in Multilevel Models: Review, Critique, and Recommendations. *Organ. Res. Methods*, 24, 443–483.

Antonakis, J., Bendahan, S., Jacquart, P. & Lalive, R. (2010). On making causal claims: A review and recommendations. *Leadersh. Q.*, Leadership Quarterly Yearly Review, 21, 1086–1120.

Arif, S. & MacNeil, M.A. (2022). Utilizing causal diagrams across quasi-experimental approaches. *Ecosphere*, 13, e4009.

Arif, S. & MacNeil, M.A. (2023). Applying the structural causal model framework for observational causal inference in ecology. *Ecol. Monogr.*, 93, e1554.

Athey, S. & Imbens, G.W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *J. Econ. Perspect.*, 31, 3–32.

Baker, A., Larcker, D.F. & Wang, C.C.Y. (2022). How Much Should We Trust Staggered Difference-In-Differences Estimates?

Beckett, S. (1954). *Waiting for Godot: tragicomedy in 2 acts*. Evergreen book. Grove Press, New York.

Bell, A., Fairbrother, M. & Jones, K. (2018). Fixed and random effects models: making an informed choice. *Qual. Quant.*, 55, 117.

Bellemare, M.F., Bloem, J.R. & Wexler, N. (2024). The Paper of How: Estimating Treatment Effects Using the Front-Door Criterion*. *Oxf. Bull. Econ. Stat.*

Benedetti-Cecchi, L. & Cinelli, F. (1997). Confounding in field experiments: direct and indirect effects of artifacts due to the manipulation of limpets and macroalgae. *J. Exp. Mar. Biol. Ecol.*, 209, 171–184.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., *et al.* (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.*, 24, 127–135.

Borusyak, K., Jaravel, X. & Spiess, J. (2023). Revisiting Event Study Designs: Robust and Efficient Estimation.

Broitman, B.R., Blanchette, C.A. & Gaines, S.D. (2005). Recruitment of intertidal invertebrates and oceanographic variability at Santa Cruz Island, California. *Limnol. Oceanogr.*, 50, 1473–1479.

Butsic, V., Lewis, D.J., Radeloff, V.C., Baumann, M. & Kuemmerle, T. (2017). Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic Appl. Ecol.*, 19, 1–10.

Callaway, B., Goodman-Bacon, A. & Sant'Anna, P.H.C. (2021). Difference-in-Differences with a Continuous Treatment.

Callaway, B. & Sant′Anna, P.H.C. (2021). Difference-in-Differences with multiple time periods. *J. Econom.*, Themed Issue: Treatment Effect 1, 225, 200–230.
Cameron, A.C. & Miller, D.L. (2015). A Practitioner′s Guide to Cluster-Robust Inference. *J. Hum. Resour.*, 50, 317–372.
Carpenter, S.R., Kitchell, J.F. & Hodgson, J.R. (1985). Cascading Trophic Interactions and Lake Productivity. *BioScience*, 35, 634–639.
de Chaisemartin, C. & D′Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *Am. Econ. Rev.*, 110, 2964–2996.
Cinelli, C. & Hazlett, C. (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 82, 39–67.
Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Suppl. J. R. Stat. Soc.*, 4, 102–118.
Cunningham, S. (2021). Causal inference. In: *Causal Inference*. Yale University Press.
Dee, L.E., Ferraro, P.J., Severen, C.N., Kimmel, K.A., Borer, E.T., Byrnes, J.E.K., *et al.* (2023). Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. *Nat. Commun.*, 14, 2607.
Dee, L.E., Miller, S.J., Peavey, L.E., Bradley, D., Gentry, R.R., Startz, R., *et al.* (2016). Functional diversity of catch mitigates negative effects of temperature variability on fisheries yields. *Proc. R. Soc. B Biol. Sci.*, 283, 20161435.
Dudney, J., Willing, C.E., Das, A.J., Latimer, A.M., Nesmith, J.C.B. & Battles, J.J. (2021). Nonlinear shifts in infectious rust disease due to climate change. *Nat. Commun.*, 12, 5102.
Efron, B. & Morris, C. (1975). Data Analysis Using Stein′s Estimator and its Generalizations. *J. Am. Stat. Assoc.*, 70, 311–319.
Eisenhart, C. (1947). The Assumptions Underlying the Analysis of Variance. *Biometrics*, 3, 1–21.
Ferraro, P.J. & Hanauer, M.M. (2014). Advances in Measuring the Environmental and Social Impacts of Environmental Programs. *Annu. Rev. Environ. Resour.*, 39, 495–517.
Ferraro, P.J. & Miranda, J.J. (2017). Panel Data Designs and Estimators as Substitutes for Randomized Controlled Trials in the Evaluation of Public Programs. *J. Assoc. Environ. Resour. Econ.*, 4, 281–317.
Fisher, R.A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.*, 52, 399–433.
Foster, S., Monk, J., Lawrence, E., Hayes, K., Hosack, G. & Przeslawski, R. (2018). Statistical considerations for monitoring and sampling.
Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
Gomes, D.G.E. (2022). Should I use fixed effects or random effects when I have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ*, 10, e12794.
Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *J. Econom.*, Themed Issue: Treatment Effect 1, 225, 254–277.
Gotelli, N.J. & Ellison, A.M. (2012). *A Primer of Ecological Statistics*. Second Edition. Oxford University Press, Oxford, New York.
Grace, J.B. (2021). Instrumental variable methods in structural equation models. *Methods Ecol. Evol.*, 12, 1148–1157.

Grace, J.B. & Irvine, K.M. (2020). Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology*, 101.
Grafström, A. & Lundström, N. (2013). Why Well Spread Probability Samples Are Balanced. *Open J. Stat.*, 3, 36–41.
Griffith, G.J., Morris, T.T., Tudball, M.J., Herbert, A., Mancano, G., Pike, L., *et al.* (2020). Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.*, 11, 5749.
Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E.D., *et al.* (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794.
Heckman, J.J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective*. *Q. J. Econ.*, 115, 45–97.
Henderson, B. (2006). Open Letter To Kansas School Board.
Hernan, M.A. & Robins, J.M. (2023). *Causal Inference: What If*. CRC Press, Boca Raton.
Holland, P.W. (1986). Statistics and Causal Inference. *J. Am. Stat. Assoc.*, 81, 945–960.
Huntington-Klein, N. (2021). *The effect: An introduction to research design and causality*. CRC Press.
Imai, K. & Kim, I.S. (2021). On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Polit. Anal.*, 29, 405–415.
Imbens, G.W. (2014). Instrumental Variables: An Econometrician's Perspective. *Stat. Sci.*, 29, 323–358.
Imbens, G.W. & Rubin, D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge.
Kendall, B.E. (2015). *A statistical symphony: instrumental variables reveal causality and control measurement error*.
Kermorvant, C., D'Amico, F., Bru, N., Caill-Milly, N. & Robertson, B. (2019). Spatially balanced sampling designs for environmental surveys. *Environ. Monit. Assess.*, 191, 524.
Kimmel, K., Dee, L.E., Avolio, M.L. & Ferraro, P.J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends Ecol. Evol.*, 36, 1141–1152.
Larsen, A.E. (2013). Agricultural landscape simplification does not consistently drive insecticide use. *Proc. Natl. Acad. Sci.*, 110, 15330–15335.
Larsen, A.E., Meng, K. & Kendall, B.E. (2019). Causal analysis in control–impact ecological studies with observational data. *Methods Ecol. Evol.*, 10, 924–934.
Larson, D., Grace, J. & Larson, J. (2008). Long-term dynamics of leafy spurge (*Euphorbia esula*) and its biocontrol agent, flea beetles in the genus Aphthona. *Biol. Control*, 47, 250–256.
Laubach, Z.M., Murray, E.J., Hoke, K.L., Safran, R.J. & Perng, W. (2021). A biologist's guide to model selection and causal inference. *Proc. R. Soc. B Biol. Sci.*, 288, 20202815.
Lubchenco, J. (1980). Algal Zonation in the New England Rocky Intertidal Community: An Experimental Analysis. *Ecology*, 61, 333–344.
MacDonald, A.J. & Mordecai, E.A. (2019). Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing. *Proc. Natl. Acad. Sci.*, 116, 22212–22218.
Marcus, M. & Sant'Anna, P.H.C. (2021). The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics. *J. Assoc. Environ. Resour. Econ.*, 8, 235–275.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
Morgan, S.L. & Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press.
Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, 46, 69–85.
Oshchepkov, A. & Shirokanova, A. (2022). Bridging the gap between multilevel modeling and economic methods. *Soc. Sci. Res.*, 104, 102689.
Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *J. Bus. Econ. Stat.*, 37, 187–204.
Paine, R.T. (1966). Food web compexity and species diversity. *Am. Nat.*, 100, 65–75.
Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82, 669–688.
Pearl, J. (2009). *Causality*. Cambridge university press.
Pearl, J., Glymour, M. & Jewell, N.P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
van de Pol, M. & Wright, J. (2009). A simple method for distinguishing within- versus between-subject effects using mixed models. *Anim. Behav.*, 77, 753–758.
Power, M.E. (1990). Effects of Fish in River Food Webs. *Science*, 250, 811–814.
Ratcliffe, H., Kendig, A., Vacek, S., Carlson, D., Ahlering, M. & Dee, L.E. (2023). Extreme precipitation promotes invasion in managed grasslands. *Ecology*, e4190.
Reichman, O.J. (1979). Desert Granivore Foraging and Its Impact on Seed Densities and Distributions. *Ecology*, 60, 1086–1092.
Rinella, M.J., Strong, D.J. & Vermeire, L.T. (2020). Omitted variable bias in studies of plant interactions. *Ecology*, 101, e03020.
Robertson, B.L., Brown, J.A., McDonald, T. & Jaksons, P. (2013). BAS: Balanced Acceptance Sampling of Natural Resources. *Biometrics*, 69, 776–784.
Robins, J. (1989). The control of confounding by intermediate variables. *Stat. Med.*, 8, 679–701.
Robinson, G.K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Stat. Sci.*, 6, 15–32.
Rosenbaum, P.R. (2002). *Observational Studies*. Springer Series in Statistics. Springer, New York, NY.
Roth, J. (2022). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *Am. Econ. Rev. Insights*, 4, 305–322.
Roth, J., Sant′Anna, P.H.C., Bilinski, A. & Poe, J. (2023). What′s trending in difference-in-differences? A synthesis of the recent econometrics literature. *J. Econom.*, 235, 2218–2244.
Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 688–701.
Rubin, D.B. (2005). Causal Inference Using Potential Outcomes. *J. Am. Stat. Assoc.*, 100, 322–331.
Ruesink, J. (2000). Intertidal mesograzers in field microcosms: linking laboratory feeding rates to community dynamics. *J. Exp. Mar. Biol. Ecol.*, 248, 163–176.
Schielzeth, H. & Nakagawa, S. (2012). Nested by design: model fitting and interpretation in a mixed model era. *Methods Ecol. Evol.*, 4, 14–24.

Shipley, B. (2016). *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference with R*. 2nd edn. Cambridge University Press, Cambridge.
Silliman, B.R., Bertness, M.D., Altieri, A.H., Griffin, J.N., Bazterrica, M.C., Hidalgo, F.J., *et al.* (2011). Whole-community facilitation regulates biodiversity on Patagonian rocky shores. *PloS One*, 6, e24502.
Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M. & Biss, P.M. (2006). The Park Grass Experiment 1856–2006: its contribution to ecology. *J. Ecol.*, 94, 801–814.
Simler-Williamson, A.B. & Germino, M.J. (2022). Statistical considerations of nonrandom treatment applications reveal region-wide benefits of widespread post-fire restoration action. *Nat. Commun.*, 13, 3472.
Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.
Stachowicz, J.J., Best, R.J., Bracken, M.E.S. & Graham, M.H. (2008). Complementarity in marine biodiversity manipulations: reconciling divergent evidence from field and mesocosm experiments. *Proc. Natl. Acad. Sci.*, 105, 18842–18847.
Stevens, D.L. & Olsen, A.R. (2004). Spatially Balanced Sampling of Natural Resources. *J. Am. Stat. Assoc.*, 99, 262–278.
Sun, L. & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econom.*, Themed Issue: Treatment Effect 1, 225, 175–199.
Textor, J., van der Zander, B., Gilthorpe, M.S., Liśkiewicz, M. & Ellison, G.T. (2016). Robust causal inference using directed acyclic graphs: the R package 'dagitty.' *Int. J. Epidemiol.*, 45, 1887–1894.
Underwood, A.J., Underwood, A.L., Underwood, A.J. & Wnderwood, A.J. (1997). *Experiments in ecology: their logical design and interpretation using analysis of variance*. Cambridge university press.
Urquhart, N.S. & Kincaid, T.M. (1999). Designs for Detecting Trend from Repeated Surveys of Ecological Resources. *J. Agric. Biol. Environ. Stat.*, 4, 404–414.
Wolkovich, E.M., Cook, B.I., Allen, J.M., Crimmins, T.M., Betancourt, J.L., Travers, S.E., *et al.* (2012). Warming experiments underpredict plant phenological responses to climate change. *Nature*.
Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. 2nd edn. Chapman and Hall/CRC, New York.
Wooldridge, J.M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
Wooldridge, J.M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.
Wooldridge, J.M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators.
Yund, P.O., Tilburg, C.E. & McCartney, M.A. (2015). Across-shelf distribution of blue mussel larvae in the northern Gulf of Maine: consequences for population connectivity and a species range boundary. *R. Soc. Open Sci.*, 2, 150513.