Title: Foundations and future directions for causal inference in ecological research

Authors: Katherine Siegel[1,2], Laura E. Dee[3]

Affiliations:
[1] Cooperative Institute for Research in Environmental Sciences, University of Colorado-Boulder; katherine.j.siegel@colorado.edu
[2] Department of Geography, University of Colorado-Boulder
[3] Department of Ecology & Evolutionary Biology, University of Colorado-Boulder; Laura.Dee@colorado.edu

Corresponding author: Katherine Siegel, 4001 Discovery Drive, Boulder, Colorado, 80303; katherine.j.siegel@colorado.edu

Statement of authorship: KS and LD both conceived of the paper, wrote the paper, and created the figures and RMarkdown tutorials.

Data accessibility statement: Data used for the tutorials in the Supporting Information are available at 1) https://github.com/katherinesiegel/intro_causal_inf, along with accompanying code, and 2) the Open Science Framework at DOI: 10.17605/OSF.IO/3XVQG.

Abstract
Ecology often seeks to answer causal questions, and while ecologists have a rich history of experimental approaches, novel observational data streams and the need to apply insights across naturally occurring conditions pose opportunities and challenges. Other fields have developed causal inference approaches that can enhance and expand our ability to answer ecological causal questions using observational or experimental data. However, the lack of comprehensive resources applying causal inference to ecological settings and jargon from multiple disciplines create barriers. We introduce approaches for causal inference, discussing the main frameworks for counterfactual causal inference, how causal inference differs from other research aims, and key challenges; application of causal inference in experimental and quasi-experimental study designs; appropriate interpretation of the results of causal inference approaches given their assumptions and biases; foundational papers; and the data requirements and trade-offs between internal and external validity posed by different designs. We highlight that these designs generally prioritize internal validity over generalizability. Finally, we identify opportunities and considerations for ecologists to further integrate causal inference with synthesis science and meta-analysis and expand the spatiotemporal scales at which causal inference is possible. We advocate for ecology as a field to collectively define best practices for causal inference.

Introduction
Questions about causal relationships are common in ecology: we seek to understand the effect of biodiversity on ecosystem functioning (Tilman *et al.* 2001, 2014), the impacts of climate change and disturbance regimes on ecosystems (García Criado *et al.* 2020; Halofsky *et al.* 2020), the effects of anthropogenic activities on animal behavior (Gaynor *et al.* 2018), the effects of different abiotic variables on plant productivity across ecosystem types (Smith *et al.* 2024), and the effectiveness of restoration and conservation (Geldmann *et al.* 2019; Suding 2011). These are

fundamentally causal questions: they seek to isolate and estimate the effect of a causal variable on an outcome (**Box 1**) and rule out alternative explanations for the estimated effects (**Table 1**).

Table 1. Matching distinct research aims with methods. Ecological studies that use observational data can have different aims, which require different methodological techniques. In addition to estimating causal relationships, we are often interested in description and prediction, or the ability to estimate outcomes outside of the observed data. Description, causal inference, and prediction ask fundamentally different questions and require different methods (Hernán et al., 2019). For instance, some methods ecologists use to assess the performance of their models are appropriate for predictive aims but not causal analysis (Addicott et al., 2022; Arif & MacNeil, 2022*a*; Pichler & Hartig, 2023)**.** Here, we demonstrate the different data needs and methods required to answer descriptive, predictive, and causal questions in ecology (table adapted from Hernán et al. (2019) and Laubach et al. (2021)).

	Description	**Prediction**	**Causal analysis**
Urban ecology (Locke et al., 2021)			
Question	How does historical redlining relate to current patterns of tree canopy cover?	Can historical redlining predict current tree canopy cover?	What is the effect of historical redlining on current patterns of tree canopy cover?
Data	● Redlining polygons ● Current tree cover	● Redlining polygons ● Current tree cover	● Redlining polygons ● Current tree cover ● Current neighborhood-level socioeconomic characteristics and zoning ● Spatial data on tree-planting efforts
Methods	Summary statistics on tree cover in redlined vs. non-redlined neighborhoods	Regression analysis predicting tree cover as a function of presence of historical redlining	Regression discontinuity design comparing tree cover at boundaries of redlined vs. non-redlined neighborhoods
Invasion ecology (Knapp & Matthews, 2000)			
Question	What are the population trends of introduced fish species and endemic amphibians in alpine	Which lakes are likely to provide suitable habitat for both introduced fish and endemic	Does the increase in populations of introduced fish species cause a decline in endemic

	lakes?	amphibians?	amphibian populations in alpine lakes?
Data	● Population of introduced fish species over time ● Population of endemic amphibians over time	● Presence/absence or abundance data for introduced fish species ● Presence/absence or abundance data for endemic amphibians ● Lake-level data on nutrient levels, elevation, surface area, maximum depth, substrate composition, solar radiation input, and isolation from other lakes	● Population of introduced fish species over time ● Population of endemic amphibians over time ● Lake-level data on nutrient levels, elevation, surface area, maximum depth, substrate composition, solar radiation input, and isolation from other lakes
Methods	Summary trends over time for both taxa	Species distribution models	Difference-in-difference comparing population trends in lakes with and without introduced fish species, before and after their introduction
Protected areas (Xu et al., 2022)			
Question	Do protected forests have different land surface temperatures than unprotected forests?	How is climate change likely to change land surface temperature in protected and unprotected forests?	Do protected areas buffer against climate change impacts on land surface temperature?
Data	● Protected area polygons ● Land cover maps ● Land surface temperature data	● Protected area polygons ● Land cover maps ● Downscaled climate projections	● Protected area polygons ● Land cover maps ● Land surface temperature data ● Site-level data on elevation, topographic roughness,

			distance to roads, distance to cities, and forest type
Methods	Compare the average land surface temperatures of protected and unprotected forests	Use process-based models to project land surface temperature in forests under different climate change scenarios	Matching protected and unprotected forests, then regression analysis to estimate the effect of protection on land surface temperature

Identifying and quantifying causal relationships, however, pose challenges in complex ecological systems. Many factors impact an outcome of interest, and confounding variables – which affect both the causal variable and the outcome – can bias estimates of causal relationships. For example, precipitation is a confounding variable when estimating the effect of plant species richness on grassland productivity, by affecting both richness and productivity (Dee *et al.* 2023). Failure to account for precipitation in our model would lead to incorrect conclusions about the significance, magnitude, and/or direction of the effect of species richness on productivity. Confounding variables occur frequently in ecological systems: as researchers, we may be aware of and able to measure some but not all of them (e.g., we may lack data on some confounding variables, or our model may be misspecified, causing us to omit a confounder). This creates challenges for understanding causal relationships in ecology.

To answer causal questions, ecologists have traditionally used randomized experiments or pseudo-experiments (Christie *et al.* 2019). However, many ecological questions face logistical and ethical challenges to experimentation, such as inability to replicate natural disturbances or ethical issues regarding manipulation of endangered or non-native species). Furthermore, experiments can be imperfect and do not always meet the assumptions required for causal inference: unexpected, non-random processes may pose challenges for their causal interpretation (Arif & Massey 2023; Kimmel *et al.* 2021). Other fields facing similar barriers, including public health and economics, have extended the foundations underlying experimental design to develop frameworks for inferring causal relationships from observational data (Greenstone & Gayer 2009; Little & Rubin 2000). These frameworks include statistical approaches for overcoming the challenges posed by experimental and observational data, emphasizing clear articulation of the assumptions required for causal interpretations of estimated effects (Hernán & Robins 2016). While the conservation impact evaluation field has embraced these approaches, particularly to assess the effectiveness of protected areas (Ferraro & Pattanayak 2006; Jones & Shreedhar 2024), causal inference approaches are less widely adopted in ecology. Encouragingly, recent reviews have provided introductions to causal inference geared towards ecologists (Butsic *et al.* 2017; Larsen *et al.* 2019), and ecological studies have increasingly applied quasi-experimental approaches (Box 1) (Dee *et al.* 2023; Ramsey *et al.* 2019; Wu *et al.* 2023) and used causal graphs (Arif & MacNeil 2023; Grace *et al.* 2016; Shipley 1999) in empirical settings.

These approaches to causal inference can improve our ability to investigate causal relationships using both experimental and observational data. Stronger integration of causal inference into

ecology can enable new insights by 1) strengthening experimental design and clarifying the assumptions required for deriving causal inference from experiments (Kimmel *et al.* 2021) and 2) advancing rigorous assessment of causal relationships from observational data (Butsic *et al.* 2017; Larsen *et al.* 2019). These approaches can enable ecologists to leverage novel data streams from remote sensing, long-term monitoring, or citizen/community science to test ecological theory in natural, non-experimental ecosystems (Dee *et al.* 2016, 2023; Larsen & Noack 2020) and ask ecological questions at management-relevant spatial and temporal scales at which randomized controlled experiments are not possible (Ratcliffe *et al.* 2022, 2024; Siegel *et al.* 2022a, b; Simler-Williamson & Germino 2022). This integration has not yet reached its full potential, as applying these approaches appropriately requires an in-depth understanding of the assumptions, strengths, and limitations of causal inference.

As ecologists, we face significant jargon and disciplinary barriers to adopting causal inference approaches, despite the recent proliferation of applications to ecology and open-source software tools. Quasi-experimental approaches to causal inference are not part of most graduate curricula in ecology, and experimental design courses may not equip students with tools to interpret their results when the assumptions underlying randomized experiments are violated. Exploring causal inference using texts from multiple other disciplines (e.g., Angrist & Pischke (2008, 2015), Cunningham (2021)), ecologists may struggle to find intuitive, applicable examples. Different fields' jargon also creates obstacles (**Box 1** provides a glossary). For example, other fields use "panel data" to describe what an ecologist might call "longitudinal data," and "fixed effects" has a different – nearly opposite – meaning in ecology than in econometrics (Byrnes & Dee 2024). These barriers raise the risk of misusing methods and missed opportunities to advance basic and applied ecology. The growth of machine learning highlights the urgency of clarifying best practices in the field of causal inference, as these popular methods may not be the best approach to answering causal questions (Pichler & Hartig 2023).

To help ecologists overcome these barriers, we provide an accessible translation of causal inference study designs by building intuition around the assumptions, strengths, and limitations of different approaches. We present the underlying frameworks of causal inference; the assumptions upon which causal inference – both from experimental and observational approaches – rest and how our interpretation of "arguably causal" results should reflect the assumptions underlying the approaches we use; and applications to ecological research. We highlight that studies are not simply "causal" or "not causal": there is a spectrum based on the strength of assumptions given the study design, data context, and research question (Kimmel *et al.* 2021). Throughout, we introduce readers to foundational texts. For additional, self-guided study, we provide a curated reading list and reproducible demonstrations of individual causal inference approaches (**Supporting Information**), drawing on our experiences teaching a graduate-level causal inference course for ecologists (**Box 2**). Building from previous introductions (e.g., Arif & MacNeil (2022b), Butsic *et al.* (2017), Fick *et al.* (2021), Grace (2021), and Ramsey *et al.* (2019)), we discuss how strengths of causal inference approaches in terms of reducing bias – internal validity – can be weaknesses in terms of generalizability and emphasize that these approaches require substantial amounts of data to detect effects. To increase generalizability, we discuss potential integrations of causal inference with synthesis science and meta-analysis and highlight how the use of new data streams (e.g., from remote sensing) can increase both the scale of inference and sample sizes for causal inference. We end with a

forward-looking view for the field to collectively define best practices for causal inference in ecology.

Causal inference frameworks

Causal analysis – including experiments and quasi-experiments – must contend with the fundamental problem that we can only observe one state of the world (Hernan 2004). We cannot directly observe how a change (e.g., a treatment, exposure, or altered condition) affects the same individual unit (e.g., person, plant, place) under both treatment and control conditions simultaneously (Holland 1986). In other words, we cannot directly observe the counterfactual: if a given unit received the treatment, we cannot observe the alternative scenario in which that same unit did not (Box 1). To address this, two complementary frameworks for causal inference have emerged: the potential outcomes (PO) framework (Rubin 1972) and the structural causal model (SCM) (Pearl 2009). In both, and throughout this paper, we define a treatment as a potential manipulation or "intervention" by humans or nature. Treatments can be binary (e.g., species presence/absence), categorical (e.g., ecosystem type), or continuous (e.g., precipitation levels). Treatments may be the result of active manipulation by humans (e.g., species introductions) or nature (e.g., beavers' transformation of hydrology) or a characteristic of a system (e.g., edaphic gradients) (Holland 1986).

The PO framework defines a causal effect based on a set of potential outcomes that could be observed in alternative states of the world (Rubin 1972, 2005): the causal effect is the difference in potential outcomes across two states of nature (**Figure 1**). The unobserved potential outcomes are counterfactuals (Morgan & Winship 2014). Counterfactuals, or well-defined alternatives to the outcomes that we observe in the world, are central to causal inference (Ferraro 2009). Different approaches are used to construct a counterfactual, all of which – including experiments, where control groups are often the counterfactual – require assumptions (Kimmel *et al.* 2021).

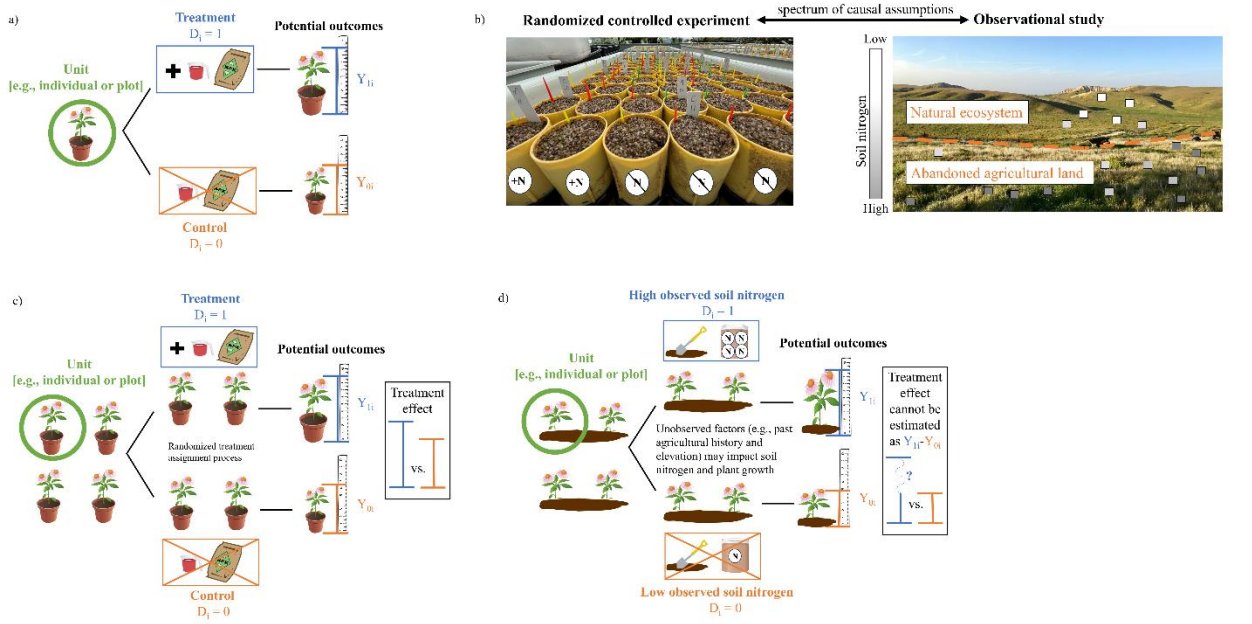

Figure 1. The fundamental problem of causal inference poses a challenge for experimental and observational studies. **a)** We cannot observe the outcomes of different treatment scenarios – receiving the treatment ($D_i = 1$) *and* not receiving the treatment ($D_i = 0$) – for a single unit (Splawa-Neyman 1923). In this example assessing nitrogen's effect on plant growth, because of the fundamental problem of causal inference, we can only observe the outcomes Y_{1i} when $D_i = 1$ and Y_{0i} when $D_i = 0$. The individual treatment effect is $Y_{1i} - Y_{0i}$, which is the causal effect of the treatment for unit i. **b)** Different approaches to causal inference range in the strength of the assumptions they make to estimate causal effects, from randomized controlled experiments (which make the weakest assumptions) to purely observational studies (which make stronger assumptions). **c)** Randomization of treatment assignment ensures there is no systematic relationship between treatment assignment and underlying characteristics of the unit that could otherwise affect the outcome, allowing for estimation of the treatment (or causal) effect as the average difference between the outcomes for the different treatments. **d)** Observational data, lacking randomization, poses challenges for causal inference. In this example, the sample plots vary in their background characteristics (e.g., past land use, elevation), which affect soil nitrogen and plant growth, complicating our ability to estimate the potential outcomes. Icons from Saxby *et al*. (2024). Photo credits: N. Emery and K. Siegel.

The other dominant causal inference framework is the SCM (Pearl 2009, 2010), which is related and complementary to the PO framework (Malinsky *et al.* 2019; Pearl 2009; Richardson & Robins 2013). The SCM framework combines counterfactual causality from PO with graphical model approaches (Spirtes *et al.* 2001), generalizing structural models more common in ecology, with roots in path analysis (Wright 1921). Recent reviews introduce the SCM to ecologists (Arif & MacNeil 2023; Laubach *et al.* 2021). Briefly, the SCM uses directed acyclic graphs (DAGs) to quantify the effects of interventions (Pearl 2009). Drawing on domain knowledge, previous research, and ecological theory, DAGs are causal diagrams that map causal relationships among variables as directional arrows or paths in a graph (**Figure 2**). DAGs make transparent our assumptions about the relationships in our study system (Pearl 2009). DAGs include all known potential confounding variables (**Box 1**) (Arif & MacNeil 2023) – whether or not they are observed in our data – and can clarify variables that fall on the causal path (mediators) or that create other sources of bias (e.g., colliders, **Box 1**) (**Figure 2**). DAGs thus provide a useful starting point for clarifying and articulating assumptions about causal relationships based on prior knowledge (**Figure 3a**) and for thinking through the spatial and temporal scales of the dynamics and variables of interest. We recommend drawing a DAG before performing an analysis, and ideally before data collection. Arif & MacNeil (2023) provide guidance for ecologists on developing a DAG and testing its consistency with the underlying data, including R code.

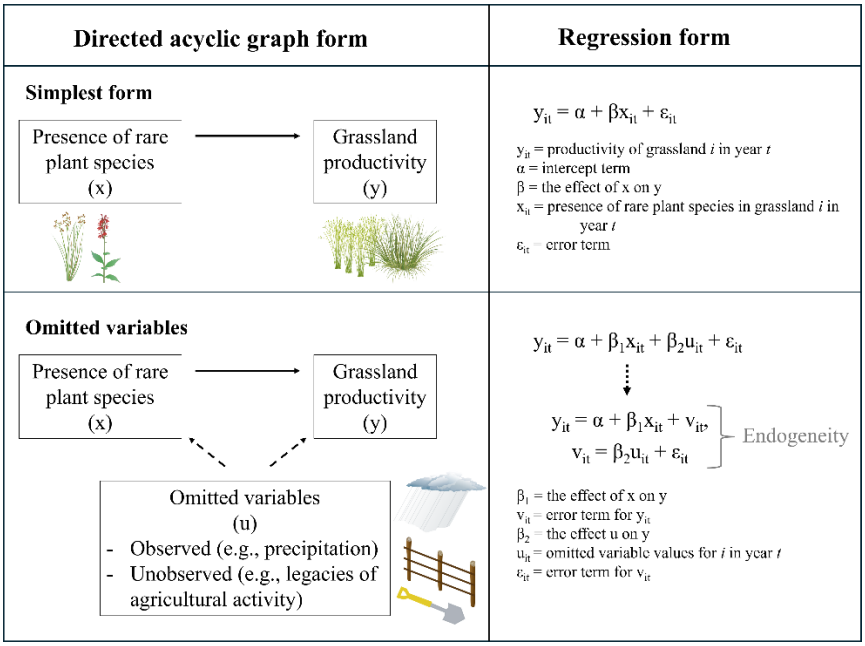

Figure 2. When confounding variables are not accounted for, endogeneity occurs: the treatment term is correlated with the error term, yielding biased estimates of the treatment effect. We demonstrate this issue using directed acyclic graphs (DAGs) to visualize a hypothesized causal effect of the presence of rare plant species on grassland productivity. We show the regression equations corresponding with each DAG to demonstrate how omission of observed or unobserved confounding variables (e.g., precipitation, historical land use) leads to biased estimates of the treatment effect. We overcome challenges to endogeneity by conditioning on all confounding variables: this is equivalent to applying the back-door criterion (i.e., blocking all back-door paths). This can be challenging: all paths must be specified and correct, and all confounding variables must be controlled for and measured without error (Huntington-Klein 2022). Icons from Saxby *et al.* (2024).

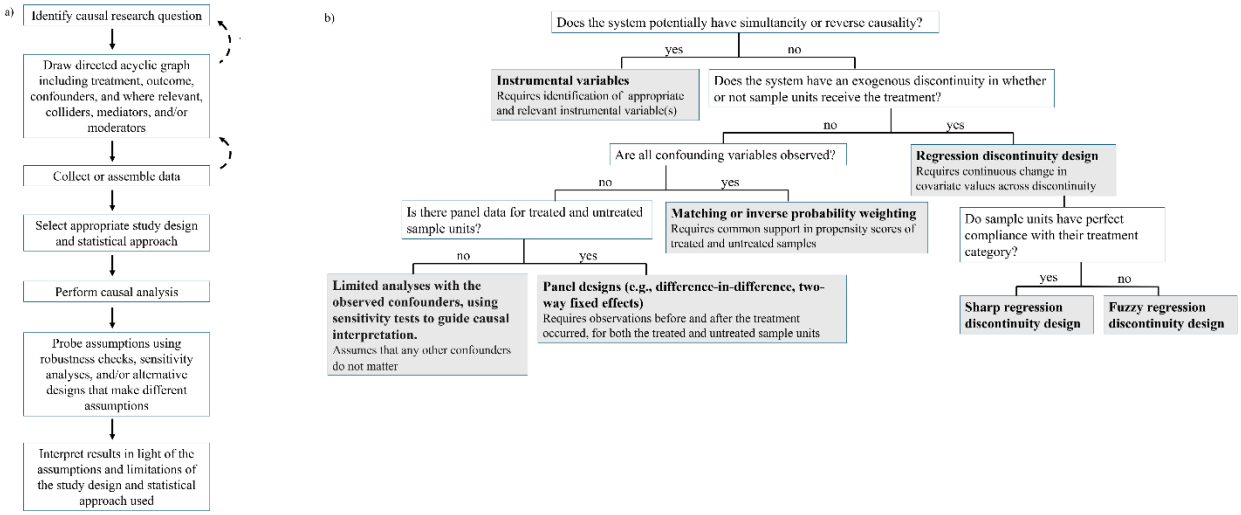

Figure 3. a) A workflow for causal inference in ecology. Dashed arrows indicate steps that may require iteration. For example, the process of drawing a DAG may lead us to modify our research

question by clarifying the outcome we believe, based on prior knowledge, would actually be impacted by the treatment we have identified. Similarly, the process of collecting or assembling data may change our DAG by forcing us to use proxies for important confounders. **b)** A decision tree for choosing a study design and statistical approach for causal inference. The research question, domain knowledge about the study system (e.g., understanding whether there are issues of simultaneity or compliance), and properties of the available data (e.g., presence of panel data) all shape the decision about which, if any, approaches will be appropriate and feasible. For each method in a gray box, we note additional key assumptions or requirements that must be met.

Challenges for causal inference

As previously introduced, the frequent occurrence of confounding variables makes causal analysis difficult in ecological systems. Confounders pose challenges for experimental and quasi-experimental approaches to causal inference: failure to account for confounding variables can bias estimates of the treatment effect (i.e., the estimated effect will differ from the true effect) because if confounders are omitted from the model, the model error will be correlated with the treatment (**Figure 2**). This phenomenon, where the treatment variable is correlated with the error term, is called endogeneity (conversely, if the treatment term is not correlated with the error term, then it is exogenous). Endogeneity can arise from other causes, like reverse or bidirectional causality or measurement error (**Box 1**) in the explanatory variable, but the challenge of confounding variables is especially pertinent in ecology. When confounding variables are not accounted for and thus cause bias in the estimator (**Box 1**), this is called omitted variable bias. Notably, omitted variable bias is an issue regardless of the sample size: increasing the sample size does not reduce the bias in the estimate. Thus, confounding variables threaten causal inference. Note that we discuss regression-based approaches to estimating causal effects, but other approaches exist (Pearl 2010).

DAGs can help identify whether we have measured or unmeasured confounding variables (so-called "back-door paths" that introduce endogeneity and lead to spurious correlations and bias) (Rohrer 2018). To satisfy Pearl's "back-door criterion," we can use DAGs to identify which confounding variables to control for so that the effect of our causal variable of interest is conditionally independent (or *d*-separated) given this control (Arif & MacNeil 2022*b*; Pearl 2009). The back-door criterion must be completed for each pathway of interest to interpret the results causally.

As nonparametric causal graphs, DAGs encode our assumptions about causal relationships in a system to help guide the choice of variables to include or not when estimating their effects (e.g., in regression analyses). On their own, however, they do not quantify or estimate the magnitude of causal effects. For causal estimation, we next describe statistical designs for causal inference that fall along a spectrum from those that require the weakest assumptions for causal interpretation, to approaches that require much stronger assumptions.

Experimental designs

The counterfactual model of causality described above was at the heart of Fisher's randomized controlled experiments (Fisher 1935). Randomized controlled experiments, or randomized control trials (RCTs), compare treated units to control units: control units serve as the

counterfactual. Randomization – or random treatment assignment – ensures that every unit has the same probability of receiving the treatment and therefore that there is no systematic relationship between the outcomes and observed or unobserved confounding variables (**Figure 1b**). Randomization makes the treatment independent of confounders (**Figure 1c**), and the expected potential outcome for the control units is the same as the expected potential outcome for the entire population. Thus, random assignment makes two or more comparable groups. With perfect randomization, groups should be identical on average prior to the treatment because every unit has an equal probability of being treated. In an ideal randomized controlled experiment, the effect of confounding variables is eliminated and the key assumptions of causal inference are met (Kimmel *et al.* 2021). Then, we can compare the differences-in-means of treatment groups to estimate an average treatment effect (**Box 1**) of the population in the experiment.

Randomized controlled experiments require the fewest and weakest assumptions for causal inference (Fernainy *et al.* 2024). However, even in experiments, several key assumptions must be met for potential outcomes – and thus counterfactuals – to be well-defined. First, experiments assume that the treatment *T* does not affect the outcome *Y* except through its effect on *X*, the cause being studied (the "excludability" assumption): the treatment is solely responsible for the different outcomes observed, and there are no confounding variables. In addition, experiments must satisfy the stable unit treatment value assumption (SUTVA), an assumption common to all causal inference approaches we discuss. SUTVA has two key components: no interference (a unit's outcome is only conditional on whether it received treatment) and no multiple versions of the treatment (there is only a single, well-defined version of each treatment level). Finally, experiments assume that there is no "non-compliance": units have or maintain the treatment they were assigned (e.g., in a seed addition experiment, the planted seeds emerge, no other species invade, and no species fail to emerge). However, these assumptions can be challenging to meet; thus, experiments can deviate from perfect randomization and compliance, highlighting the need to engage explicitly with causal thinking when interpreting the results of experiments (Kimmel *et al.* 2021). Furthermore, while randomized controlled experiments are viewed as the gold standard for causal inference in terms of internal validity (or the extent to which a study accurately estimates a causal relationship within a study population), generalizing from experiments and creating experiments that replicate the conditions and scales of processes found in nature pose challenges.

Quasi-experimental designs

Without randomized treatment assignment and experimental control, quasi-experimental designs can facilitate causal inference but require more assumptions – many of which are inherently untestable – to be met (Imbens 2024). These approaches require careful probing and justification of their assumptions based on system-specific knowledge to support interpretation of arguably causal relationships. Quasi-experiments can be used at any spatial and temporal scale, while randomized, controlled ecological experiments in the field and lab are mostly restricted to smaller scales. Quasi-experiments often use specific data structures, such as cross-sectional and panel data. Cross-sectional data are observations from multiple units at a single point in time, facilitating comparison of treatment effects across individuals. Panel data are observations of multiple units across multiple time points.

Quasi-experimental approaches to causal inference must often contend with selection bias resulting from non-random treatment assignment. For example, we may be interested in the effect of land-based nutrient pollution on kelp cover (Krumhansl *et al*. 2016) (**Figure 4a**). To answer this question, we might relate remotely-sensed water quality data to long-term kelp cover monitoring data. However, distance to human settlements and the coast is likely an important confounding factor: it affects the amount of land-based pollution to which a kelp forest is exposed (the treatment), and it also affects fishing pressure on predators of sea urchins, which in turn affect kelp cover (Ling *et al*. 2009). If we simply compared kelp forests with high vs. low levels of nutrient pollution, we might attribute observed differences in kelp cover to pollution without accounting for the confounding effect of remoteness on fishing pressure. This example demonstrates selection bias: kelp forests exposed to the treatment are systematically different from untreated kelp forests in ways that affect the outcome. Kelp forests with the highest nutrient pollution are likely closer to coastal areas with high human population densities and thus also subject to higher fishing pressure, while kelp forests with minimal pollution are far from the coast with less accessible fishing grounds (Witman & Lamb 2018). Selection bias stems from non-random treatment assignment: the units exposed to the treatment we wish to study are not randomly selected, which introduces confounding. When study designs fail to account for selection bias, the estimated difference in the mean outcomes for the treated and untreated groups actually represents the average causal effect *plus* the effect of selection bias.

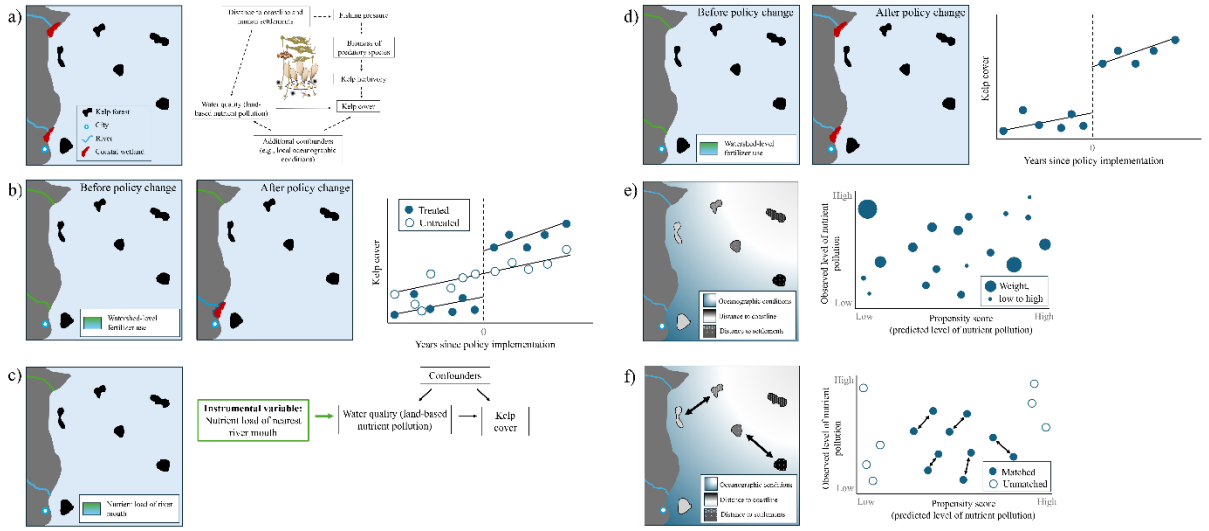

Figure 4: Illustrations of quasi-experimental methods. **a)** A DAG illustrates assumed causal relationships and confounders for a hypothetical study of land-based nutrient pollution's impact on kelp cover. **b)** Difference-in-differences compare treated and untreated units before and after treatment implementation (here, a policy improving water quality discharged by rivers by reducing fertilizer use and restoring wetlands). **c)** Instrumental variables isolate treatment effects through variables that impact the treatment but only influence the outcome through their relationship with the treatment. Here, green outlining indicates the instrumental variable (nearest river mouth's nutrient load). **d)** Regression discontinuity designs compare units on either side of interventions (here, implementation of the policy from 4b). **e)** Inverse probability of treatment weighting uses propensity scores to weight units based on the likelihood that their treatment status is the status predicted by their observable confounders. **f)** Matching uses propensity scores to identify treated and untreated units with comparable confounding variables.

Confounding variables may be observable (i.e., factors that the investigator has identified as potential confounders *and* measured) or unobservable (i.e., factors that are known and not measured, or unknown). It may not be possible to measure unobservable variables: the study site may lack historical records (Butsic *et al.* 2017), data may not be publicly available, or collecting these data may be prohibitively expensive. Different quasi-experimental methods take different approaches to dealing with, and make different assumptions about, the presence and importance of observable and unobservable confounders. And while quasi-experimental designs require more assumptions than randomized experiments to derive arguably causal findings, these approaches have the benefit of observing natural conditions (rather than conditions manipulated by the researcher) and enable analyses at broader scales.

We review the main approaches to quasi-experimental causal inference, categorizing them according to whether or not they condition on unobservable confounders in addition to observables. We discuss each approach's assumptions, strengths, limitations, and data requirements (**Figure 3b**). Butsic *et al.* (2017) and Larsen *et al.* (2019) provide further introductions to these approaches. Across all these approaches, we recommend, as a first step, drawing a DAG – based on knowledge of the study system and ecological theory – with the treatment, outcome, and all potential confounders, and mediators and moderators, when they are relevant to the research question (**Box 1**).

Conditioning on observable and unobservable confounders

Among quasi-experimental approaches to causal inference, approaches that condition on both observable and unobservable confounders require the weakest assumptions for causal interpretation, by relaxing the assumption that we have observed all confounders (**Figure 4b-d**). These approaches can yield arguably causal interpretations even if we cannot measure or do not know all confounding variables in our system, or if we have drawn an incorrect DAG and thus do not know the true data-generating process. We briefly review the core ideas and assumptions, applications to ecology, and recent trends for these approaches, focusing on difference-in-difference designs, panel regressions, instrumental variables, and regression discontinuity designs. In **Table 2**, we highlight more recent extensions for these designs.

We start with difference-in-difference (DiD) designs – similar conceptually to before-after control-impact (BACI) and thus familiar to ecologists (Green 1979; Stewart-Oaten & Bence 2001) – which compare the differences in control and treated groups before and after an intervention or exposure (reviewed, with extensions, in (Wauchope *et al.* 2021)). This approach compares the differences between the (*treated group after − treated group before*) - (*untreated group after − untreated group before*) to estimate how much more the treated group changed as compared to how much the untreated group changed (**Figure 3b**; **Figure 4b**). To create a counterfactual, difference-in-difference relies on the untestable assumption that the trends in time for these groups would be the same (or parallel) without the treatment. While most textbook examples consider binary treatments, difference-in-difference also applies to continuous treatments or treatments of different intensities (Callaway *et al.* 2024). This field is rapidly evolving, with emerging methodological extensions for cases where the parallel trends assumption is violated and effects are not homogenous (reviewed in Roth *et al.* (2023)).

Similarly, panel regressions or "within" estimators make comparisons within groups, such as sites, individuals, or time periods. Panel regression controls for fixed differences across units and time-specific effects, or variables that affect all sites in a unit of time) (Wooldridge 2010) (**Figure 3b**). Time-invariant characteristics of sites can be confounding (e.g., more remote kelp forests also tend to be less impacted by land-based pollution (**Figure 4**)); these across-site differences are "between variation." To control for these differences, panel approaches use "fixed effects" – dummy variables for each group to control for time-invariant, confounding differences across groups, whether or not the confounding variables are observed. Here, a fixed effect has a different meaning than its use in mixed effect and hierarchical modeling, which instead considers a fixed effect to be a parameter that does not vary by group (Bolker *et al.* 2009). With this approach, we can track how, within a location, kelp cover changes through time in response to other variables that change through time, like sea surface temperature. Thus, we can compare sites to themselves at different treatment levels (e.g., levels of nutrient pollution) observed at different points of time as the counterfactual (Dee *et al.* 2023).

These approaches differ from, and make weaker assumptions for causal identification than, mixed effects models using random effects (Byrnes & Dee 2024) or conditioning on observable confounding variables along (Dee *et al*. 2023). The downside is panel approaches "throw out" the between variation (both confounding and otherwise) and require large panel datasets because they estimate a coefficient for each group and time (Angrist & Pischke 2008; Wooldridge 2010). Nested sampling designs can exploit cross-sectional data with multiple plots sampled across multiple sites and retain between-group variation (reviewed in Byrnes & Dee (2024) and Wooldridge (2010)). These approaches are increasingly used in ecology (e.g., Dudney *et al.* (2021), Ratcliffe *et al.* (2022) Suskiewicz *et al.* (2024)) and are straightforward to implement in R (Bergé 2018): see Dee *et al.* (2023) and Byrnes & Dee (2024) for tutorials.

Instrumental variables (IV) regression can eliminate sources of bias from all forms of confounding variables (including the time-varying confounding variables missed in DiD), measurement error, reverse causality, and simultaneity (**Box 1**; **Figure 3b**). IV regression uses a third variable (an "instrument", *Z*) that is related to the treatment, *X*, but not to the outcome, *Y*, except through its effect on *X* (or at least, after controlling for other variables in the system) (Angrist & Krueger 2001; Imbens 2014) (**Figure 4c**). An IV in a regression mimics what an experiment's randomization process would do, where the randomly assigned treatment process is independent of *Y*. The IV must be strongly related to the treatment, but not with the outcome (after controlling for other covariates). When these two assumptions are met, IV regression yields a local average treatment effect (**Box 1**). The challenge is finding a valid and relevant IV.

For example, Macdonald & Mordecai (2019) used IV regression to isolate the effects of deforestation on malaria transmission and *vice versa* in the Amazon; because their effects are simultaneous, isolating one from another is challenging with standard methods such as mixed effect models. They used dry season aerosol pollution as an IV to isolate the effects of annual deforestation on malaria transmission from the reverse relationship. For causal interpretation, a key, untestable assumption is that dry season aerosol pollution and deforestation are strongly related (because most deforestation occurs and cleared forests are burned in the dry season) but that dry season aerosol pollution does not directly affect annual malaria transmission after

controlling for other factors.

The final quasi-experimental approach to controlling for observable and unobservable confounders we discuss is regression discontinuity design (RDD) (**Figure 4d**). RDD is an option when there is a spatial, temporal, or policy discontinuity that separates treated from untreated units (Hahn *et al.* 2001; Imbens & Lemieux 2008); **Figure 3b**). In RDDs, the treated and untreated units are sorted according to their position relative to a threshold in the "running" variable (which defines the location of the discontinuity): on one side of the threshold, all units receive the treatment, while all units on the other side are untreated. RDDs compare the outcomes of units located directly on either side of the threshold to estimate the treatment effect (Cattaneo *et al.* 2019).

RDDs assume that the location of the discontinuity is exogenous: all observable and unobservable confounding variables are constant or continuous on either side of the threshold, without jumps in their values. As a result, units located directly on either side of the threshold are very similar to one another (there are generally no units observed directly at the threshold). In ecological systems, it can be difficult to identify appropriate, exogenous discontinuities in the absence of policy changes and management interventions (Englander 2019), although temporal discontinuities (e.g., before and after a disturbance event) may meet the assumptions of RDDs (Grainger & Costello 2014). RDDs assume that in the absence of the treatment, the outcome would not change discontinuously at the threshold (Hahn *et al.* 2001). To assess the validity of this assumption, RDDs require sufficient data on both sides of the threshold (Wuepper & Finger 2023). We also assume that all unobserved confounders are either correlated with the running variable or not discontinuous across the threshold. Generally, RDDs estimate the treatment effect using a narrow bandwidth of units on either side of the threshold to avoid making assumptions about the shape of the underlying regression functions.

RDDs also assume that the probability of treatment changes discontinuously at the threshold (Cattaneo *et al.* 2019). In a sharp RDD, we assume perfect compliance: all units above the threshold receive the treatment, while none of the units below the cutoff receive it (**Figure 3b**). We can relax this assumption and use fuzzy RDD, which merely assumes that the probability of treatment changes discontinuously at the threshold (Wuepper & Finger 2023). Fuzzy RDDs allow for treatment noncompliance: the value of the running variable is a predictor of whether a unit received the treatment but does not completely determine its treatment status. The value of the running variable relative to the threshold thus functions as an IV that affects the outcome solely through its effect on the likelihood of treatment. RDDs also assume that there is no endogenous sorting of units: units do not seek to be on one side of the threshold (Lee 2008). In ecological applications, endogenous sorting may occur where animal behavior comes into play – e.g., the landscape of fear shapes animal movement (Gaynor *et al.* 2019) – or where treatments cause spatial spillovers – e.g., protected area establishment increases extractive activities directly outside reserve boundaries (Ewers & Rodrigues 2008).

A strength of RDD is that many of the underlying assumptions can be tested visually (Cattaneo & Titiunik 2022). We can test whether the discontinuity is exogenous by plotting the values of confounding variables across the threshold and checking for discontinuous change (Cattaneo *et al.* 2019). Plotting the data using placebo thresholds can reveal whether there are locations with

similar treatment effects in the absence of a treatment discontinuity (Noack *et al.* 2022; Wuepper & Finger 2023). Density tests that check for increased sample unit density on one side of the threshold can test for endogenous sorting (McCrary 2008).

Conditioning on observable confounders

Quasi-experimental designs that condition on observable confounders (**Figure 3b**) make the strong, untestable assumption that all important confounders are observable. Two such approaches are inverse probability of treatment weighting (hereafter, "weighting") and matching (**Figure 4e and 4f**). Both use observable confounders to calculate propensity scores, or the probability of a unit receiving a treatment based on that unit's covariate values (Rosenbaum & Rubin 1983; Stuart 2010). In matching, we develop a set of control and treated units by identifying the control units with propensity scores closest to those of the treated units (**Figure 4f**). We discard untreated units that do not have similar propensity scores to treated units and *vice versa*, maintaining only units with sufficient overlap in their covariate values (i.e., common support). In weighting, each unit is weighted based on its propensity score such that treated units with high propensity scores and untreated units with low propensity scores have lower weights than other units (**Figure 4e**). Weighting retains all units. We use the weights in the subsequent regression model to estimate the treatment effect. In both matching and weighting, including the covariates used to calculate the propensity score in the subsequent regression model increases the robustness of the treatment effect estimate (Jones & Lewis 2015). Matching is more commonly used with a binary treatment, although continuous treatments are sometimes stratified, and there is ongoing development of approaches for continuous treatments (Brown *et al.* 2021; Fong *et al.* 2018; Hirano & Imbens 2004).

Weighting and matching integrate well with regression methods and work with both panel and cross-sectional data. Estimated treatment effects are also less sensitive to mis-specified models (Butsic *et al.* 2017), and propensity scores reduce bias from measurement error in the covariates (Austin 2010). There are also simple diagnostics to assess the quality of matches, including comparison of standardized mean differences pre- and post-matching. Matching can also reveal where there is not sufficient common support to make plausible causal claims (Ho *et al.* 2011). Finally, if unobserved confounders are correlated with the observed confounders, these approaches can adjust for unobservables.

The essential assumption of weighting and matching is that selection bias is caused by observable confounders or unobserved confounders that are correlated with observed variables: they suffer from omitted variable bias when there are unobservable confounders (Simler-Williamson & Germino 2022). Compared to experimental designs and quasi-experimental approaches that condition on unobservables, weighting and matching require stronger assumptions for causal interpretation. Finally, matching has several distinct limitations: it relies on sufficient common support for treated and untreated units, and it reduces the variation in the dataset because units without quality matches are dropped. The results must be interpreted in the context of the reduced dataset: the estimated treatment effect is valid for the range of units included in the matched dataset, but it may not be appropriate to extrapolate the estimated effect to the full dataset.

In addition, and complementary to, these quasi-experimental designs rooted in the PO framework

are approaches from the SCM framework. In SCM, conditioning on all confounding variables is equivalent to applying Pearl's back-door criterion (blocking all back-door paths). This makes two assumptions: that all paths are specified correctly, and that confounding variables are observable and measured (Huntington-Klein 2022). When this cannot be achieved, an alternative is the front-door criterion, which adjusts for a mediator that is uncorrelated with the confounding variables of concern (Bellemare *et al.* 2024; Pearl 1995, 2009). Controlling for an exogenous mediator blocks the effect of omitted confounding variables and isolates the effect of the causal variable of interest (Pearl & Mackenzie 2018). To implement the front door criterion, one first estimates the effect of the treatment on the mediator without confounders, then estimates the effect of the mediator on the outcome. These two effects are multiplied to get the total effect of the treatment on the outcome (Arif & MacNeil 2023). However, identifying situations in which the front door-criterion works is challenging, so it is less frequently used (Huntington-Klein 2022)

Discussion

Causal questions are central to ecological understanding, and ecology has a rich tradition of experiments to address causal questions and estimate the magnitude of causal effects. In recent years, ecological literature reviewing or applying causal inference approaches that complement experimental approaches has exploded, highlighting a variety of approaches that can exploit new data streams to extend ecological understanding to broader spatial and temporal scales. However, making sense of how and when to apply these approaches, and navigating the wide-ranging, rapidly evolving, technical, and jargon-filled fields that causal inference spans still pose challenges. In response, we review key challenges for causal inference using experimental and observational data in ecology, quasi-experimental approaches to answering causal questions, and the key assumptions underlying these approaches. Building on previous reviews (e.g., Butsic *et al.* (2017) and Larsen *et al.* (2019)), we explicitly define quasi-experimental designs in terms of their treatment of unobservable confounding variables. We believe that this distinction is very important for ecologists, as approaches that select on both observable and unobservable variables require weaker assumptions for causal interpretation. We demonstrate how the use of PO and SCM frameworks can be complementary and provide a workflow for moving from a causal question, to a DAG, to the appropriate methodological approach, to interpretation of results (**Figure 3**). We also provide resources for self-guided study, including reproducible code with accompanying data and a curated reading list (**Supporting Information**).

Causal inference is not as straightforward as following a recipe or implementing a pre-existing software package. Robust causal inference requires careful combination of pre-existing knowledge (formalized in DAGs), appropriate data, study design, and interpretation of estimated effects in light of key assumptions. Adding nuance, approaches for causal inference pose trade-offs and require different assumptions, some of which may be more or less plausible in particular contexts (Grace 2024). In addition, the approaches reviewed here emphasize carefully estimating one causal effect at a time, rather than estimating all causal effects in a system at once, although causal inference can contribute to the goal of building system-level knowledge (**Box 4**). To navigate these important nuances, we synthesize some critical considerations: the spectrum of weak to strong assumptions required for causal interpretation of estimated effects, different designs' trade-offs between internal and external validity, and data requirements for causal inference. We offer recommendations for overcoming these limitations and outline future

research needs.

Tradeoffs between internal and external validity

Causal inference designs exist along a spectrum from true randomization to purely observational: this spectrum reflects both the strength of assumptions needed for causal interpretation and trade-offs in internal and external validity. While internal validity refers to accurate estimates of causal relationships within a study population, external validity is the extent to which a study's results can be applied beyond the study sample (Spake *et al.* 2022). Quasi-experimental approaches and randomized controlled experiments prioritize internal validity: researchers rigorously eliminate sources of bias in their estimates of the treatment effect (Desjardins *et al.* 2021). Much like experiments with different treatments, different quasi-experimental designs yield distinct estimands, with varying implications for external validity. On one end of the spectrum are ideal, randomized controlled experiments, which prioritize internal validity and require the weakest assumptions for causal interpretation of effect sizes. However, experiments may struggle with external validity, as the controlled conditions and specific populations involved can limit the generalizability of findings to broader, real-world contexts (Dee *et al.* 2023) or other forms of a treatment (Wolkovich *et al.* 2012). Moving further along the spectrum, quasi-experiments and imperfect experiments also have trade-offs between internal and external validity (Kowalski 2023). Both RDD and IV estimate local average treatment effects (LATE) rather than the average treatment effect (ATE) (**Box 1**). RDDs estimate the LATE using units located directly on either side of the discontinuity (Baker & Lindeman 2024): it may not be appropriate to extrapolate this LATE to units located far from the discontinuity, although emerging methods allow researchers to assess RDDs' external validity (Wing & Bello-Gomez 2018; Wuepper & Finger 2023). Similarly, the estimated causal effects of IV designs only apply to compliers (the units that vary in response to the IV, **Box 1**) (Imbens, 2010).

However, we often seek generalizability, or external validity, to extend our findings beyond the units and spatiotemporal scale that we studied (Spake *et al.* 2022). Moving further along the spectrum to observational studies that condition only on observables, approaches like matching estimate average treatment effects or average treatment effects on the treated (**Box 1**) but make stronger assumptions about our ability to identify and include all confounders. Still, in matching, because we exclude unmatched units, we cannot assume that the estimated treatment effect would apply to units whose covariate values fall outside the area of common support (Crump *et al.* 2009; Stuart 2010). For example, Siegel *et al.* (2022*a*) use matching to estimate the effect of federal vs. private land ownership on wildfire probability in western US forests. However, because federal wilderness areas tend to be at higher elevations than private forests, the matched dataset includes relatively few wilderness units (<7% of federal units in the matched dataset). It would thus be inappropriate to naively extrapolate their findings to high-elevation wilderness forests.

Data considerations

Available data also determine generalizability, the choice of causal inference design (and therefore internal validity), and the statistical power to detect an effect. Quasi-experimental designs have specific data requirements: their appropriateness will depend on both the research question and the data context. As noted previously, cross-sectional and panel data are common dataset structures in quasi-experimental approaches. Difference-in-difference and panel designs

require panel data, while IV and RDD require some plausibly exogenous sources of variation. With only cross-sectional data, design options are more limited (e.g., matching or weighting), and it is harder to flexibly control for confounding variables, particularly unobserved variables.

Cross-sectional versus panel data also may reflect different effects and degrees of generalizability. Cross-sectional data captures a snapshot in time and can be used for space-for-time comparisons, which are critiqued in applications such as climate change ecology for issues with generalizability (Lovell *et al.* 2023). While cross-sectional data allow us to examine how a particular treatment (e.g., exposure to reduced precipitation) affects multiple units (e.g., grassland plots in different locations), panel data allow us to examine trends over time across the treated and untreated units and generalize to multiple time points. This facilitates the study of ecologically interesting questions such as time lags in treatment effects, the effects of varying levels of treatment exposure over time, and interactions between the treatment and covariates over time. However, there may also be trade-offs in existing datasets in terms of spatial extent versus resampling through time. Furthermore, a reliance on panel data that includes the pre-treatment period is ecologically limiting, as we are less likely to have these data for processes such as climate change impacts, species introductions, and unexpected disturbances. The realities of funding and data collection logistics may also restrict availability of panel data.

Sample size is a related consideration; many quasi-experimental approaches require relatively large datasets for sufficient statistical power to detect effects. Thus, ecologists working with limited datasets from field-based observations may not have sufficient data to leverage causal inference methods or enough power to detect a treatment effect (Kimmel *et al.* 2023; Lemoine *et al.* 2016). When there are interactions between the treatment and other covariates, the required sample size increases. New data streams can not only scale up ecological understanding and inferences when coupled with quasi-experimental approaches, but also increase statistical power.

Synthesis and meta-analyses can help expand external validity by combining multiple internally valid studies covering a range of naturally occurring conditions (Spake *et al.* 2022). Meta-analysis is a common approach to quantitative synthesis in ecology, especially of experiments. However, when meta-analyses include original studies with biased estimands, they can yield biased estimates and inaccurate results. This limitation is true for observational designs and imperfect experiments (Kimmel *et al.* 2021). Further, the estimands may not be the same across studies, muddying quantitative comparisons. Similarly, if the original studies feeding into a meta-analysis focus on different subpopulations with heterogeneous treatment effects, it becomes difficult to combine and generalize the estimated effects (Spake *et al.* 2022). Study eligibility criteria can reduce the probability of including original studies with bias: we recommend that the field develop eligibility criteria based on the treatment of confounding variables and other sources of bias. We may need to develop other approaches to account for remaining endogeneity in the original studies (Mathur & VanderWeele 2022) and for comparisons when different estimands and subpopulations are involved.

More generally, synthesis science combines datasets from disparate sources and often seeks to disentangle causal relationships (Carpenter *et al.* 2009; Halpern *et al.* 2020). Quasi-experimental approaches can expand and accelerate synthesis science's contributions to ecological knowledge, but measurement error (the difference between the true vs. recorded value of a variable) presents

a challenge. Synthesis approaches combine multiple data sources, each with their own sources of measurement error. When the extent and types of measurement error differ across studies, error and uncertainty can propagate through models using synthesized data.

Another opportunity is using large-scale datasets, such as time series derived from satellite imagery, combined with causal inference approaches. The volume of data from Earth observations, community science programs, and other distributed surveys and monitoring networks is rapidly increasing, expanding observations of ecological systems at larger scales and the sample sizes available for data-hungry approaches. More observations of ecosystems under a wider variety of time points, conditions, and scales will also increase generalizability of inferences and enable us to test new theories that span different spatial and temporal scales of causal relationships.

A challenge posed by these expanding data streams, however, is mismatches in spatial and temporal resolution between the treatment, confounders, and outcome. For example, remote sensing data can facilitate analysis at broader spatial scales but often are available at coarser resolutions, which can obscure understanding of highly localized processes (Alix-García & Millimet 2023; Jain 2020). If we were interested in the effect of artificial nighttime lights on large mammal behavior, for example, we might have outcome data at fine spatio-temporal scales (e.g. multiple data points per hour, accurate within several meters) from radio-collars, treatment data at 30x30 meter resolutions in the form of daily nighttime light data (Román *et al.* 2018), and data on environmental and socioeconomic confounders at various resolutions. However, newer remote sensing techniques and products (such as LiDAR) can generate data at finer spatial and temporal resolutions over larger spatial extents, helping to overcome issues with scale mismatches. There is also a growing literature examining the unique challenges when using remote sensing data for causal inference, as these data are often derived using machine learning. This magnifies challenges for controlling for confounders and of measurement error. For instance, if confounders are included in the machine learning model that predicts the data, that can introduce bias. Analogously, errors from machine learning model predictions are a form of measurement error in subsequent causal models that can introduce bias (Alix-García & Millimet 2023; Gordon *et al.* 2023; Jain 2020; Proctor *et al.* 2023). New methodological and conceptual advances are needed to reconcile these challenges and facilitate larger scale ecological causal understanding (Van Cleemput et al. 2024).

Establishing shared best practices for causal inference in ecology

Further integration of these approaches into the research design and statistics curriculum for graduate students in ecology can help us harness the power of causal inference (**Box 2**). Many ecologists report a desire for more statistical training and a mismatch between their formal training and current best-practices (Barraquand *et al.* 2014; Touchon & McCoy 2016). In our experience, graduate students are eager to learn new approaches. Through an emphasis on building students' intuitions regarding the strengths, limitations, and underlying assumptions of causal inference designs, focused coursework can strengthen students' research design and statistical skills. As the use of causal inference in ecology becomes more popular, we also need careful, critical reviewers to evaluate and provide input into these studies. Fortunately, a growing body of ecological studies, applications, and general resources can contribute to self-guided and course-based learning (Heiss 2022; Huntington-Klein 2022) (**Supporting Information**). As

more ecologists gain a working understanding of how causal inference can integrate with ecological research, we can develop a collective and evolving set of best-practices as a field (**Box 3**).

Finally, this synthesis is not exhaustive. **Table 2** summarizes additional topics and references. While we focus on counterfactual-based causal inference approaches, to build on ecology's rich history of experimentation, there are alternative notions of causality, such as causal detection (Munch *et al.* 2023; Runge *et al.* 2023; Sugihara *et al.* 2012) or causal discovery (Spirtes *et al.* 2001). Lastly, the designs we present can also be estimated using structural equation modeling (Shipley 1999, 2009) and Bayesian approaches (Li *et al.* 2023; Oganisian & Roy 2021).

Table 2: A summary of some additional topics in causal inference that could inform ecological research, including recent methodological advances. For each topic, we provide a brief description and some key readings for further self-guided study. Key readings include texts discussing the fundamentals of a method (denoted with a *) and texts that demonstrate an application of the particular method (denoted with a ǂ).

Topic	Description	Key readings
Experimental design and techniques to deal with imperfect experiments		
Challenges for experimental design and interpretation: non-compliance and attrition	Issues that arise when units do not receive the treatment they were assigned to or when units initially included in the sample are lost or otherwise not included in the analysis	Gerber & Green, 2012*
Challenges for experimental and observational design and interpretation: interference and spillovers between units	SUTVA assumes no interference between units, but in ecological settings, there may be interactions and spillovers between units	Ogburn & VanderWeele, 2014; Tchetgen & VanderWeele, 2012* Ferraro et al., 2018; Reich et al., 2021ǂ
Understanding mechanisms		
Mediation analysis and experimental design for mediation	Methods for assessing the direct effect of a treatment on a response and the indirect effect of the treatment, which is due to a mediator on the causal pathway from treatment to outcome	VanderWeele, 2015; Pirlott & MacKinnon, 2016* Huberman et al., 2020ǂ
Moderators, heterogeneous treatment effects, and conditional average treatment effects	Methods for assessing when and how different units may respond differently to the treatment (e.g., when the effect of one variable on the response differs depending on	Athey & Imbens, 2015*; Wager & Athey, 2018 Ferraro & Hanauer, 2014; Miller, 2020; VanderWeele, 2015*

	the level of another variable, or moderator)	
Sensitivity analyses		
Partial identification	An approach to causal inference that uses weaker (and more plausible) assumptions to estimate the upper and lower bounds of a causal effect	Arriagada et al., 2012; Hazzah et al., 2014; McConnachie et al., 2016‡
Sensitivity analyses and placebo designs	Methods for testing the robustness of estimated causal effects to violations of underlying assumptions	Eggers et al., 2023; VanderWeele & Ding, 2017; Cattaneo & Titiunik, 2022; Liu et al., 2013*
Generalizability, reproducibility, and transportability of effects		
Meta-analysis and generalizability	Approaches to generalizing results from studies with varying levels of external validity	Spake et al., 2022; Nakagawa et al., 2023; Spake et al., 2023*
Replication and pre-registration	Improving reproducibility (e.g., through defining research questions and approaches a priori)	Nosek et al., 2018; Strømland, 2019; Filazzola & Cahill, 2021; Kimmel et al., 2023*
Extensions: emerging tools for quasi-experimental approaches		
Staggered treatments, heterogeneity, and robust difference in difference	Extensions to difference-in-difference methods that can accommodate units that enter treatment at different times and relax assumptions above parallel trends	Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021*
Synthetic control methods	Methods of developing control units that use weighted averages of all potential control units to develop counterfactuals that are as comparable to the treated units as possible	Abadie et al., 2011* Abadie & Gardeazabal, 2003; Sills et al., 2015; West et al., 2020‡
Causal inference with measurement error	Challenges posed by measurement error; methods for accounting for measurement error	Alix-García & Millimet, 2023‡
Time series and dynamic panel models	Methods that allow for time lags, feedbacks, and changing	Arellano & Bond, 1991*

	relationships between variables over time	
Causal discovery		
Causal discovery using graphical models	Data-driven approaches to learning causal relationships from large datasets	Glymour et al., 2019; Runge et al., 2019, 2023; Spirtes et al., 2001*
Machine learning		
Machine learning fusion with causal inference	Recent advances blend machine learning approaches with causal inference (e.g. causal forests for heterogeneous treatment effects)	Athey, 2015; Athey & Imbens, 2015; Athey, 2017; Athey & Imbens, 2019; Pichler & Hartig, 2023*

Conclusion

With growing interest in and use of causal inference techniques, best practices – that are decided on and adopted by the field of ecology – are needed. This will allow us to effectively and constructively evaluate each other's work and build on it. Transparency is key, as causal analyses rely on assumptions at multiple stages, from study design to estimation of treatment effects and causal interpretation. By clearly stating and justifying our assumptions, we can create more credible estimates of causal effects and more reproducible results, enabling others to build on existing studies through improvements in data and methods. Ongoing and future improvements in estimation and identification (reviewed in Athey & Imbens (2019) and Roth *et al.* (2023)) – which are rapidly evolving in diverse fields, including ecology – can potentially weaken the underlying assumptions required for causal interpretations (Roth *et al.* 2023). Transparency about underlying assumptions can also help readers interpret the estimand, determine whether they believe a causal interpretation is appropriate, and understand the limits of a result's generalizability (Spake *et al.* 2022). Finally, transparency ensures that the approaches used are appropriate for the question at hand. Credible causal estimates will enable us to advance basic and applied ecology, informing ecological theory and ecosystem management at broad scales.

Acknowledgements

L.D. acknowledges support from the US National Science Foundation NSF CAREER #2340606 and NASA BioSCape #80NSSC 22K0796. K.S. acknowledges support from a NOAA Climate and Global Change Postdoctoral Fellowship. We thank Jon Chase, A. Simler-Williamson, and two anonymous reviewers for their feedback. We thank the students in our course, Causal Inference in Ecological Data, in Spring 2023 at the University of Colorado-Boulder. We thank Andrew Heiss, Paul Ferraro, and Van Butsic for inspiring teaching materials.

Box 1: Key terms in causal inference

As different disciplines have contributed to the development of causal inference, the field has accumulated a dizzying array of jargon. These specialized terms pose barriers to ecologists seeking to engage with the literature. We provide definitions for some key terms, with an extended glossary in Appendix 1.

- **Average treatment effect (ATE):** the average difference in the outcome variables between the treated and control populations (**Figure 1**).
- **Bias:** the difference between the estimated effect and the true value of the effect.
- **Collider:** a variable that is affected by both the treatment and the outcome. Conditioning on a collider can lead to incorrect estimates of the direction of the effect.
- **Complier:** a sample unit that received the treatment to which it was assigned: a unit that was assigned to the treated group and received the treatment, or a unit that was assigned to the untreated group and did not receive the treatment.
- **Conditioning:** an approach to isolating the effect of the treatment on the outcome of interest by considering the values of all other variables in a model given a certain value of the variable on which the model is conditioned. Also referred to as "adjusting."
- **Confounder:** a variable that affects both the treatment and the outcome. Failing to account for confounding variables biases estimates of the treatment effect.
- **Control:** the untreated units in an experiment or quasi-experiment.
- **Counterfactual:** well-defined alternative(s) to what we observe in the world.
- **Endogeneity:** correlation between the treatment variable and the error term, arising due to omitted confounding variables, reverse causality, simultaneity, or measurement error in the explanatory variable.
- **Estimand:** the effect of the treatment compared to the control for a specific population (e.g., average treatment effect, average treatment effect of the treated, local average treatment effect, and conditional average treatment effect (**Supporting Information**)).
- **Estimator:** a statistical approach to estimating the value of a model parameter.
- **Exogeneity:** the condition in which the treatment variable is not correlated with or causally influenced by other model parameters.
- **Local average treatment effect (LATE):** the treatment effect for units that were assigned to the treated group and did in fact receive the treatment, ignoring the effect of non-compliance.
- **Measurement error:** the difference between the true and recorded/observed value of a variable. Measurement error in the treatment variable biases the estimates, while measurement error in the outcome variable adds noise to the model without biasing the estimates.
- **Mediator:** a variable that lies on the causal pathway between the treatment and the outcome.
- **Moderator:** a variable that affects the magnitude of the causal effect, often implemented in statistical analyses and regression as an interaction term.
- **Omitted variable bias:** bias in estimates of the treatment effect that occurs when study designs do not account for confounding variables.
- **Panel data:** data collected for the same sample units over multiple time periods (i.e., longitudinal data).
- **Outcome:** the value of the response variable.
- **Quasi-experiments:** study designs that assess causal relationships in the absence of randomization, using variation in units' exposure to treatment(s).
- **Random assignment:** an approach to treatment assignment in which all units have an equal probability of receiving the treatment, regardless of underlying characteristics.

Randomization ensures that there are no systematic differences between the treated and control units, allowing for unbiased estimation of the treatment effect.

- **Reverse causality:** the outcome variable affects the treatment, rather than the treatment affecting the outcome.
- **Selection bias:** when the units that are exposed to the treatment are not randomly selected, there may be systematic differences between the treated and control samples, biasing the estimate of the treatment effect.
- **Simultaneity:** the treatment affects the outcome and the outcome affects the treatment.
- **Stable unit treatment value assumption (SUTVA):** the assumption that there is no interference in the system (the treatment status of one unit cannot influence the outcome of another unit) and that for each unit, there are not different versions of each treatment level or hidden variation in the treatment.
- **Treatment:** a potential manipulation by humans or nature. Causal inference focuses on treatments/causes where we could hypothetically imagine an ideal controlled experiment with randomized treatment assignment.

Box 2: Teaching causal inference

Formal coursework can increase ecologists' understanding of causal inference. To contribute to the development of causal inference curricula for ecologists, we developed and taught a graduate-level course on causal inference for ecology in the spring of 2023 in the Department of Ecology & Evolutionary Biology at University of Colorado-Boulder, USA. The course attracted participants from diverse fields, including PhD students and postdoctoral scholars in ecology, evolutionary biology, microbiology, geography, and environmental studies, as well as project scientists from academic research groups and a government agency. More than 90% of the course participants are now integrating causal inference methods into their dissertations, as side projects, or in their work in government agencies.

Course participants had different levels of statistical training, ranging from undergraduate-level statistics to extensive previous coursework in graduate-level biostatistics and econometrics. There was a similar diversity in experience and comfort with programming in R, the software language the course used. To meet the needs of this student body, we emphasized developing an intuitive understanding of the methods we taught, rather than stressing the underlying mathematics. For those with more technical training in statistics, we also provided key references for deeper dives into the math underlying these methods.

Our overall objectives were for students to gain an understanding of the main frameworks for counterfactual causal inference and how causal inference differs from other empirical research aims; familiarity with how causal inference is applied in experimental and quasi-experimental study designs; and experience reading the published literature with a critical eye towards appropriate use of methods for identifying causal relationships. Specifically, students learned to a) summarize key threats to causal inference and identify these threats when evaluating study designs; b) apply causal inference methods to real world research questions and datasets; c) identify the most appropriate study design(s) and methodology in non-experimental settings based on the available data and research question; d) implement these designs and methods using R and e) appropriately interpret the results and their potential biases; and f) communicate clearly about these methods, their results, and their assumptions.

The course consisted of lectures introducing key topics and methods, demonstrations of how to implement quasi-experimental methods in R using simulated and real datasets, student-led discussions of publications that used different approaches, and semester-long individual projects. Students demonstrated their understanding of the applications of different causal inference approaches, the underlying assumptions, and strengths and limitations of different methods through their projects: students identified a causal question, developed a DAG and revised it based on feedback, compiled the necessary data , conducted a preliminary analysis using a quasi-experimental method introduced in class, and interpreted the results in the context of the method's underlying assumptions (**Figure 3**). The projects gave the students an opportunity to apply causal inference to their own research areas, with a focus on understanding the underlying intuition and learning the mechanics of applying causal inference to real-world problems. They also gained experience in providing feedback on each other's analyses, practicing skills required for peer review.

Students readily adopted DAGs, but many struggled to align their datasets with quasi-experimental designs. They often found matching and weighting to be the most intuitive approaches, even though these methods make the strongest assumptions. They also gravitated towards these methods due to data constraints (e.g., a single time period of data with no clear discontinuity or instrumental variable). Students were familiar with randomized experiments but not the approaches available when an experiment does not go to plan. Students' uncertainty in determining which quasi-experimental method best fit their research question and data motivated us to create **Figure 3b**.

We provide a curated reading list from our course (Appendix 2) as a resource for those interested in developing similar courses or using the reading list to structure their own, self-guided learning. In our experience, a course on causal inference in ecology is useful for students familiar with ecological statistics and experimental design, but fundamental concepts of causal inference – such as underlying assumptions and issues with confounding – could be incorporated throughout research methods and study design curricula for ecologists.

In our experience, existing textbooks may not be well-suited as stand-alone texts for causal inference. Many textbook examples focus on binary treatments, while ecologists often encounter continuous or categorical treatments. This can create misconceptions about the applicability of quasi-experimental methods to ecological contexts (**Box 4**). Some textbooks pose issues for educators seeking to foster an inclusive and just classroom, as they may simplify complex social issues (e.g., positioning gender as a binary treatment).

Box 3: Best practices for causal inference in ecology

To take advantage of causal inference approaches responsibly and effectively, some understanding of their underlying assumptions and the contexts in which one study design is more robust than another is needed. To interpret an estimated effect or correlation as causal, assumptions are always required, even in randomized experiments (Kimmel *et al.* 2021). Indeed, just using a randomized experiment or quasi-experimental approach does not guarantee that the interpretation holds causal meaning. While we believe best practices should emerge collectively,

we suggest some (non-exhaustive) guidelines to move this process forward, based on our experiences teaching, applying, and reviewing papers on causal inference using ecological data.

Data collection and pre-analysis planning. Best practices start before data collection and study design enable more robust causal inference. An emerging focus is on best practices for reproducibility in ecology prior to data analysis (Kimmel *et al.* 2023; Parker *et al.* 2016).

- Create a DAG based on domain knowledge and use it to guide data collection or assembly of existing data.
- Where possible, collect or assemble pre-treatment data to expand the study designs available for your analysis (e.g., difference-in-difference, temporal regression discontinuity designs, or other panel design), in both experimental and quasi-experimental settings.
- Use pre-registration or pre-analysis plans to define your study design in advance. This enhances reproducibility, clarifies assumptions, and reduces the likelihood of p-hacking.
- Perform power analyses, particularly in field studies with low replication, datasets with small sample sizes, or when aiming to estimate interactions (moderator) effects, or site-specific effects.

Study and statistical design. When possible, use designs that make weaker assumptions, and triangulate results using complementary methods and sensitivity tests.

- Use DAGs to guide your analysis choices and include them in a pre-analysis plan.
- Where possible (based on the research question and data), use methods that make fewer and weaker assumptions about confounding variables (e.g., methods that condition on observable and unobservable confounders: IV, RDD, difference-in-difference, within-estimator, and other panel designs), because even with extensive domain knowledge, ecological surprises are still possible.
- Use multiple, complementary designs to assess the robustness of estimates to different assumptions about observed and unobserved confounding (e.g., Dee *et al.* (2023)).
- Use sensitivity tests to assess the robustness of estimates to the presence of confounding variables that have not been controlled for (Andraczek *et al.* 2024; Liu *et al.* 2013).

Communicating assumptions. Clarity around assumptions allows others to understand the study's strengths and limitations and can help identify further avenues for research.

- Include your DAG(s) in published analyses and interpret your model results in relation to your DAG. This will allow others to understand and critique your causal model, and we expect published DAGs to also help the field identify commonly unobserved confounders that may require new data streams to address.
- Clearly discuss the assumptions of your study and statistical design, including in randomized experiments, and how they are met as well as caveats.
- Explain how your design accounts for observable and unobservable sources of confounding.

Interpreting results. Careful interpretation of results given your study design's limitations and assumptions increases transparency and credibility.

- Interpret your findings in the context of your study's assumptions and all forms of bias that may remain in your estimates. If your model does not consider unobserved

confounders, note that your estimates of causal effects may include the effect of both the causal variable and unobserved confounders.
- Explicitly discuss the limitations of your analysis in terms of potential reasons the assumptions required for causal interpretation could be violated.
- Interpret analyses based on their external validity and avoid over-generalizing your inferences.

Changing incentives and norms. As a community, a cultural shift that prioritizes transparency and robustness in publishing would improve the credibility of causal inference. As reviewers, we recommend viewing explicit acknowledgement of a study's assumptions and limitations (e.g., transparency around issues with internal validity given potential violations of assumptions and limits to the generalizability of findings) as a strength, rather than a weakness or a justification for rejecting a study's findings. Current publishing incentives, which may dissuade transparency around assumptions and limitations, put the robustness and credibility of causal inference at risk. More transparency around these could help the science to build on itself (e.g., by collecting new data to overcome assumptions, or developing new methods to relax them).

Box 4: Common misconceptions about causal inference
We clarify some misconceptions about causal inference approaches and their applicability to ecology.

1. Causal inference is purely statistical and does not rely on domain knowledge. Researchers using statistical designs for causal inference do not apply these methods in a vacuum, but rather draw on ecological knowledge to shape their research questions, hypotheses, study design, and importantly, the interpretation and caveats surrounding the estimated effects and reported causal relationships. Drawing a DAG is a useful first step for formalizing prior knowledge (Figure 3).

2. Quasi-experiments can only handle binary treatments. Quasi-experiments can accommodate continuous and multivalue treatments (e.g., the effect of fire severity categories on forest biomass). They can also estimate heterogeneous effects (Table 2).

3. Causal inference does not provide information on mechanisms. The causal inference approaches we review can examine causal mechanisms, either by including moderators (e.g., interaction terms in a regression) or mediators on the causal path, using mediation analyses (VanderWeele 2015; Huberman *et al.* 2020) (Table 2, Figure 4).

4. The structural causal model (SCM) and the potential outcomes (PO) framework are competing and non-overlapping frameworks. SCM and PO are complementary frameworks and have been unified and translated from one to another (Malinsky *et al.* 2019; Richardson & Robins 2013). Both seek to achieve unbiased estimates and make assumptions for causal interpretation transparent.

5. Prior ecological knowledge can tell us the size and direction of a causal relationship and the bias associated with its estimate. Ecological systems are complex, and while prior knowledge allows us to form hypotheses about the direction and magnitude of causal relationships, our knowledge is limited. We may be incorrect in our assumptions about the size and direction of the

bias in our estimates. Bias could mean a true effect is masked (i.e., appears to be zero in an analysis) or the estimated effect is a mirage (i.e., a spurious effect, when there is no true effect). Bias can also lead to the incorrect sign of an effect or assumed relationship. Approaches such as the "useful approximation standard" (Grace 2024), which suggest that an estimated causal effect must simply be "predominantly causal" (i.e., the causal component of the estimate is greater than the bias component), thus make very strong assumptions, because the size of the bias versus that of the true causal effect is unknowable. Under these strong assumptions, researchers may run the risk of allowing confirmation bias to guide the interpretation of their results.

6. Studies or estimated effects are either causal or not causal. Causal inference designs fall along a spectrum based on the strength of the assumptions they make. Applying a quasi-experimental, experimental, or structural modeling approach does not guarantee that the estimated effect reflects a true causal effect. Causal interpretation of empirical estimates relies on assumptions and domain knowledge about whether those assumptions are met. While causal inference methods attempt to reach unbiased estimates, complete lack of bias is almost always unachievable. However, we believe that unbiased estimates should still be the goal, as there is no rigorous and reproducible definition of a "good enough" estimate (*sensu* Grace 2024). For instance, if a researcher could fully randomize their experimental treatment, they would not opt to only partially randomize it. Interpretation of effect sizes and relationships from causal analyses in ecology requires transparency about their assumptions and limitations. Use of robustness tests can help assess the strength of findings (**Box 3**).

7. Causal inference methods do not allow for generalizability. While causal inference approaches prioritize internal validity, statistical designs such as matching provide more general estimands, and causal inference approaches can be integrated with meta-analysis to generalize their findings.

8. Quasi-experiments do not seek to understand how a system works. The approaches we discuss seek to build up ecological understanding through estimations of each individual element and process in the system. Often, multiple analyses and DAGs are needed to advance understanding of multiple relationships within that system, as an individual DAG may have many assumed relationships but not identify all causal pathways (i.e., an individual DAG may not satisfy the backdoor criterion for all relationships in the system).

9. There is a silver bullet for causal understanding. There is no one-size-fits-all approach or formula to follow for causal inference. Instead, causal inference is a process that iteratively integrates prior knowledge, data, and causal assumptions. The choice of approach is based on best available knowledge, methods, and data – which are all evolving as science progresses.

References

Abadie, A., Diamond, A. & Hainmueller, J. (2011). Synth: An R Package for synthetic control methods in comparative case studies. *J Stat Softw*, 42.

Abadie, A. & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 93, 113–132.

Addicott, E.T., Fenichel, E.P., Bradford, M.A., Pinsky, M.L. & Wood, S.A. (2022). Toward an improved understanding of causation in the ecological sciences. *Front Ecol Environ*, 20, 474–480.

Alix-García, J. & Millimet, D.L. (2023). Remotely incorrect? Accounting for nonclassical measurement error in satellite data on deforestation. *J Assoc Environ Resour Econ*, 10, 1335–1367.

Andraczek, K., Dee, L.E., Weigelt, A., Hinderling, J., Prati, D., Le Provost, G., *et al.* (2024). Weak reciprocal relationships between productivity and plant biodiversity in managed grasslands. *Journal of Ecology*, 112, 2359–2373.

Angrist, J.D. & Krueger, A.B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15, 69–85.

Angrist, J.D. & Pischke, J.-S. (2008). *Mostly Harmless Econometrics*. Princeton University Press, Princeton, NJ.

Angrist, J.D. & Pischke, J.-S. (2015). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, Princeton, NJ.

Arellano, M. & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev Econ Stud*, 58, 277.

Arif, S. & MacNeil, M.A. (2022*a*). Predictive models aren't for causal inference. *Ecol Lett*, 25, 1741–1745.

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

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

Arif, S. & Massey, M.D.B. (2023). Reducing bias in experimental ecology through directed acyclic graphs. *Ecol Evol*.

Arriagada, R.A., Ferraro, P.J., Sills, E.O., Pattanayak, S.K. & Cordero-Sancho, S. (2012). Do payments for environmental services affect forest cover? A farm-level evaluation from Costa Rica. *Land Econ*, 88, 382–399.

Athey, S. (2015). Machine learning and causal inference for policy evaluation. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, pp. 5–6.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science (1979)*, 355, 483–485.

Athey, S. & Imbens, G.W. (2015). *Machine Learning for Estimating Heterogeneous Casual Effects* (No. 3350).

Athey, S. & Imbens, G.W. (2019). Machine Learning Methods That Economists Should Know About. *Annu Rev Econom*, 11, 685–725.

Austin, P.C. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*, 29, 2137–2148.

Baker, S.G. & Lindeman, K.S. (2024). Multiple Discoveries in Causal Inference: LATE for the Party. *CHANCE*, 37, 21–25.

Barraquand, F., Ezard, T.H.G., Jørgensen, P.S., Zimmerman, N., Chamberlain, S., Salguero-Gómez, R., *et al.* (2014). Lack of quantitative training among early-career ecologists: a survey of the problem and potential solutions. *PeerJ*, 2, e285.

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*.
Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*.
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.
Brown, D.W., Greene, T.J., Swartz, M.D., Wilkinson, A. V. & DeSantis, S.M. (2021). Propensity score stratification methods for continuous treatments. *Statistical Medicine*, 40, 1189–1203.
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.
Byrnes, J.E.K. & Dee, L.E. (2024). Causal inference with observational data and unobserved confounding variables. *bioRxiv*.
Callaway, B., Goodman-Bacon, A. & Sant'Anna, P.H. (2024). *Difference-in-differences with a continuous treatment*. Cambridge, MA.
Callaway, B. & Sant'Anna, P.H.C. (2021). Difference-in-Differences with multiple time periods. *J Econom*, 225, 200–230.
Carpenter, S.R., Armbrust, E.V., Arzberger, P.W., Chapin, F.S., Elser, J.J., Hackett, E.J., *et al.* (2009). Accelerate synthesis in ecology and environmental sciences. *Bioscience*, 59, 699–701.
Cattaneo, M.D., Idrobo, N. & Titiunik, R. (2019). *A Practical Introduction to Regression Discontinuity Designs*. Cambridge University Press.
Cattaneo, M.D. & Titiunik, R. (2022). Regression discontinuity designs. *Annu Rev Econom*, 14, 821–851.
Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I. & Sutherland, W.J. (2019). Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *Journal of Applied Ecology*, 56, 2742–2754.
Crump, R.K., Hotz, V.J., Imbens, G.W. & Mitnik, O.A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96, 187–199.
Cunningham, S. (2021). *Causal Inference: The Mixtape*. 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. *Proceedings of the Royal Society B: Biological Sciences*, 283, 20161435.
Desjardins, E., Kurtz, J., Kranke, N., Lindeza, A. & Richter, S.H. (2021). Beyond standardization: Improving external validity and reproducibility in experimental evolution. *Bioscience*, 71, 543–552.
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.
Eggers, A.C., Tuñón, G. & Dafoe, A. (2023). Placebo tests for causal inference. *Am J Pol Sci*.
Englander, G. (2019). Property rights and the protection of global marine resources. *Nat Sustain*, 2.

Ewers, R.M. & Rodrigues, A.S.L. (2008). Estimates of reserve effectiveness are confounded by leakage. *Trends Ecol Evol*, 23, 113–116.
Fernainy, P., Cohen, A.A., Murray, E., Losina, E., Lamontagne, F. & Sourial, N. (2024). Rethinking the pros and cons of randomized controlled trials and observational studies in the era of big data and advanced methods: a panel discussion. *BMC Proc*, 18, 1.
Ferraro, P.J. (2009). Counterfactual thinking and impact evaluation in environmental policy. In: *Environmental Program and Policy Evaluation: Addressing Methodological Challenges* (eds. Birnbaum, M. & Mickwitz, P.). Jossey-Bass, San Francisco, CA, pp. 75–84.
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. & Pattanayak, S.K. (2006). Money for nothing? A call for empirical evaluation of biodiversity conservation investments. *PLoS Biol*, 4, 482–488.
Ferraro, P.J., Sanchirico, J.N. & Smith, M.D. (2018). Causal inference in coupled human and natural systems. *Proceedings of the National Academy of Sciences*, 116, 5311–5318.
Fick, S.E., Nauman, T.W., Brungard, C.C. & Duniway, M.C. (2021). Evaluating natural experiments in ecology: using synthetic controls in assessments of remotely sensed land treatments. *Ecological Applications*, 31.
Filazzola, A. & Cahill, J.F. (2021). Replication in field ecology: Identifying challenges and proposing solutions. *Methods Ecol Evol*, 12, 1780–1792.
Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd.
Fong, C., Hazlett, C. & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Ann Appl Stat*, 12.
García Criado, M., Myers-Smith, I.H., Bjorkman, A.D., Lehmann, C.E.R. & Stevens, N. (2020). Woody plant encroachment intensifies under climate change across tundra and savanna biomes. *Global Ecology and Biogeography*, 29, 925–943.
Gaynor, K.M., Brown, J.S., Middleton, A.D., Power, M.E. & Brashares, J.S. (2019). Landscapes of Fear: Spatial Patterns of Risk Perception and Response. *Trends Ecol Evol*, 34, 355–368.
Gaynor, K.M., Hojnowski, C.E., Carter, N.H. & Brashares, J.S. (2018). The influence of human disturbance on wildlife nocturnality. *Science (1979)*, 360, 1232–1235.
Geldmann, J., Manica, A., Burgess, N.D., Coad, L. & Balmford, A. (2019). A global-level assessment of the effectiveness of protected areas at resisting anthropogenic pressures. *Proceedings of the National Academy of Science*, 1–7.
Gerber, A.S. & Green, D.P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton & Company, New York, NY.
Glymour, C., Zhang, K. & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Front Genet*, 10.
Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *J Econom*, 225, 254–277.
Gordon, M., Ayers, M., Stone, E. & Sanford, L.C. (2023). Remote control: Debiasing remote sensing predictions for causal inference. In: *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*.
Grace, J.B. (2021). Instrumental variable methods in structural equation models. *Methods Ecol Evol*, 12, 1148–1157.
Grace, J.B. (2024). An integrative paradigm for building causal knowledge. *Ecol Monogr*, 94.

Grace, J.B., Anderson, T.M., Seabloom, E.W., Borer, E.T., Adler, P.B., Harpole, W.S., *et al.* (2016). Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature*, 529, 390–393.
Grainger, C.A. & Costello, C.J. (2014). Capitalizing property rights insecurity in natural resource assets. *J Environ Econ Manage*, 67, 224–240.
Green, R.H. (1979). *Sampling Design and Statistical Methods for Environmental Biologists*. Wiley, New York.
Greenstone, M. & Gayer, T. (2009). Quasi-experimental and experimental approaches to environmental economics. *J Environ Econ Manage*, 57, 21–44.
Hahn, J., Todd, P. & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69, 201–209.
Halofsky, J.E., Peterson, D.L. & Harvey, B.J. (2020). Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA. *Fire Ecology*, 16.
Halpern, B.S., Berlow, E., Williams, R., Borer, E.T., Davis, F.W., Dobson, A., *et al.* (2020). Ecological synthesis and its role in advancing knowledge. *Bioscience*.
Hazzah, L., Dolrenry, S., Naughton, L., Edwards, C.T.T., Mwebi, O., Kearney, F., *et al.* (2014). Efficacy of two lion conservation programs in Maasailand, Kenya. *Conservation Biology*, 28, 851–860.
Heiss, A. (2022). *Program Evaluation for Public Service*. Available at: https://evalf22.classes.andrewheiss.com/. Last accessed 12 June 2024.
Hernan, M.A. (2004). A definition of causal effect for epidemiological research. *J Epidemiol Community Health (1978)*, 58, 265–271.
Hernán, M.A., Hsu, J. & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, 32, 42–49.
Hernán, M.A. & Robins, J.M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*, 183, 758–764.
Hirano, K. & Imbens, G.W. (2004). The Propensity Score with Continuous Treatments. In: *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (eds. Meng, X.-L. & Gelman, A.). Wiley Series in Probability and Statistics, New York, NY, pp. 73–84.
Ho, D.E., Imai, K., King, G. & Stuart, E.A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *J Stat Softw*, 42, 1–28.
Holland, P.W. (1986). Statistics and causal inference. *J Am Stat Assoc*, 81, 945–960.
Huberman, D.B., Reich, B.J., Pacifici, K. & Collazo, J.A. (2020). Estimating the drivers of species distributions with opportunistic data using mediation analysis. *Ecosphere*, 11.
Huntington-Klein, N. (2022). *The Effect: An Introduction to Research Design and Causality*. Chapman & Hall.
Imbens, G.W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J Econ Lit*, 48, 399–423.
Imbens, G.W. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science*, 29, 323–358.
Imbens, G.W. (2024). Causal inference in the social sciences. *Annu Rev Stat Appl*, 11, 123–152.
Imbens, G.W. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *J Econom*, 142, 615–635.

Jain, M. (2020). The benefits and pitfalls of using satellite data for causal inference. *Rev Environ Econ Policy*, 14, 157–169.
Jones, J.P.G. & Shreedhar, G. (2024). The causal revolution in biodiversity conservation. *Nat Hum Behav*.
Jones, K.W. & Lewis, D.J. (2015). Estimating the counterfactual impact of conservation programs on land cover outcomes: The role of matching and panel regression techniques. *PLoS One*, 10, 1–22.
Kimmel, K., Avolio, M.L. & Ferraro, P.J. (2023). Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nat Ecol Evol*, 7, 1525–1536.
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.
Knapp, R.A. & Matthews, K.R. (2000). Non-native fish introductions and the decline of the mountain yellow-legged frog from within protected areas. *Conservation Biology*, 14, 428–438.
Kowalski, A.E. (2023). How to examine external validity within an experiment. *J Econ Manag Strategy*, 32, 491–509.
Krumhansl, K.A., Okamoto, D.K., Rassweiler, A., Novak, M., Bolton, J.J., Cavanaugh, K.C., *et al.* (2016). Global patterns of kelp forest change over the past half-century. *Proceedings of the National Academy of Sciences*, 113, 13785–13790.
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.
Larsen, A.E. & Noack, F. (2020). Impact of local and landscape complexity on the stability of field-level pest control. *Nat Sustain*, 4, 120–128.
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. *Proceedings of the Royal Society B: Biological Sciences*.
Lee, D.S. (2008). Randomized experiments from non-random selection in U.S. House elections. *J Econom*, 142, 675–697.
Lemoine, N.P., Hoffman, A., Felton, A.J., Baur, L., Chaves, F., Gray, J., *et al.* (2016). Underappreciated problems of low replication in ecological field studies. *Ecology*, 97, 2554–2561.
Li, F., Ding, P. & Mealli, F. (2023). Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381.
Ling, S.D., Johnson, C.R., Frusher, S.D. & Ridgway, K.R. (2009). Overfishing reduces resilience of kelp beds to climate-driven catastrophic phase shift. *Proceedings of the National Academy of Sciences*, 106, 22341–22345.
Little, R.J. & Rubin, D.B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annu Rev Public Health*, 21, 121–145.
Liu, W., Kuramoto, S.J. & Stuart, E.A. (2013). An Introduction to Sensitivity Analysis for Unobserved Confounding in Nonexperimental Prevention Research. *Prevention Science*, 14, 570–580.
Locke, D.H., Hall, B., Grove, J.M., Pickett, S.T.A., Ogden, L.A., Aoki, C., *et al.* (2021). Residential housing segregation and urban tree canopy in 37 US Cities. *npj Urban Sustainability*, 1, 15.

Lovell, R.S.L., Collins, S., Martin, S.H., Pigot, A.L. & Phillimore, A.B. (2023). Space-for-time substitutions in climate change ecology and evolution. *Biological Reviews*, 98, 2243–2270.
MacDonald, A.J. & Mordecai, E.A. (2019). Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing. *Proc Natl Acad Sci U S A*, 116, 22212–22218.
Malinsky, D., Shpitser, I. & Richardson, T. (2019). A potential outcomes calculus for identifying conditional path-specific effects. In: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* . PMLR, Naha, Japan.
Mathur, M.B. & VanderWeele, T.J. (2022). Methods to address confounding and other biases in meta-analyses: Review and recommendations. *Annu Rev Public Health*, 43, 19–35.
McConnachie, M.M., Romero, C., Ferraro, P.J. & van Wilgen, B.W. (2016). Improving credibility and transparency of conservation impact evaluations through the partial identification approach. *Conservation Biology*, 30, 371–381.
McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *J Econom*, 142, 698–714.
Morgan, S.L. & Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press.
Munch, S.B., Rogers, T.L. & Sugihara, G. (2023). Recent developments in empirical dynamic modelling. *Methods Ecol Evol*, 14, 732–745.
Nakagawa, S., Yang, Y., Macartney, E.L., Spake, R. & Lagisz, M. (2023). Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environ Evid*, 12, 8.
Noack, F., Larsen, A., Kamp, J. & Levers, C. (2022). A bird's eye view of farm size and biodiversity: The ecological legacy of the iron curtain. *Am J Agric Econ*, 104.
Nosek, B.A., Ebersole, C.R., DeHaven, A.C. & Mellor, D.T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115, 2600–2606.
Oganisian, A. & Roy, J.A. (2021). A practical introduction to Bayesian estimation of causal effects: Parametric and nonparametric approaches. *Stat Med*, 40, 518–551.
Ogburn, E.L. & VanderWeele, T.J. (2014). Causal Diagrams for Interference. *Statistical Science*, 29.
Parker, T.H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J.D., Chee, Y.E., *et al.* (2016). Transparency in ecology and evolution: Real problems, real solutions. *Trends Ecol Evol*, 31, 711–719.
Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82, 669–688.
Pearl, J. (2009). *Causality*. Cambridge University Press.
Pearl, J. (2010). An introduction to causal inference. *Int J Biostat*, 6.
Pearl, J. & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA.
Pichler, M. & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods Ecol Evol*, 14, 994–1016.
Pirlott, A.G. & MacKinnon, D.P. (2016). Design approaches to experimental mediation. *J Exp Soc Psychol*, 66, 29–38.
Proctor, J., Carleton, T. & Sum, S. (2023). *Parameter recovery using remotely sensed variables*. Cambridge, MA.
Ramsey, D.S.L., Forsyth, D.M., Wright, E., McKay, M. & Westbrooke, I. (2019). Using propensity scores for causal inference in ecology: Options, considerations, and a case study. *Methods Ecol Evol*, 10, 320–331.

Ratcliffe, H., Ahlering, M., Carlson, D., Vacek, S., Allstadt, A. & Dee, L.E. (2022). Invasive species do not exploit early growing seasons in burned tallgrass prairies. *Ecological Applications*, 32.
Ratcliffe, H., Kendig, A., Vacek, S., Carlson, D., Ahlering, M. & Dee, L.E. (2024). Extreme precipitation promotes invasion in managed grasslands. *Ecology*, 105.
Reich, B.J., Yang, S., Guan, Y., Giffin, A.B., Miller, M.J. & Rappold, A. (2021). A Review of Spatial Causal Inference Methods for Environmental and Epidemiological Applications. *International Statistical Review*, 89, 605–634.
Richardson, T.S. & Robins, J.M. (2013). *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality* (No. 128). Seattle, WA.
Rohrer, J.M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Adv Methods Pract Psychol Sci*, 1, 27–42.
Román, M.O., Wang, Z., Sun, Q., Kalb, V., Miller, S.D., Molthan, A., *et al.* (2018). NASA's Black Marble nighttime lights product suite. *Remote Sens Environ*, 210, 113–143.
Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
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. (1972). Estimating causal effects of treatments in experimental and observational studies. *ETS Research Bulletin Series*, 1972.
Rubin, D.B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J Am Stat Assoc*, 100, 322–331.
Runge, J., Gerhardus, A., Varando, G., Eyring, V. & Camps-Valls, G. (2023). Causal inference for time series. *Nat Rev Earth Environ*.
Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci Adv*, 5.
Saxby, T., Hawkey, J. & Anderson, S. (2024). Integration and Application Network Media Library.
Shipley, B. (1999). Testing causal explanations in organismal biology: Causation, correlation and structural equation modelling. *Oikos*, 86, 374–382.
Shipley, B. (2009). Confirmatory path analysis in a generalized multilevel context. *Ecology*, 90, 363–368.
Siegel, K.J., Larsen, L., Stephens, C., Stewart, W. & Butsic, V. (2022a). Quantifying drivers of change in social ecological systems: land management impacts wildfire probability in forests of the western US. *Reg Environ Change*, 22, 98.
Siegel, K.J., Macaulay, L., Shapero, M., Becchetti, T., Larson, S., Mashiri, F.E., *et al.* (2022b). Impacts of livestock grazing on the probability of burning in wildfires vary by region and vegetation type in California. *J Environ Manage*, 322.
Sills, E.O., Herrera, D., Kirkpatrick, A.J., Brandão, A., Dickson, R., Hall, S., *et al.* (2015). Estimating the impacts of local policy innovation: The synthetic control method applied to tropical deforestation. *PLoS One*, 10, e0132590.
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.

Smith, M.D., Wilkins, K.D., Holdrege, M.C., Wilfahrt, P., Collins, S.L., Knapp, A.K., *et al.* (2024). Extreme drought impacts have been underestimated in grasslands and shrublands globally. *Proceedings of the National Academy of Sciences*, 121.

Spake, R., Bowler, D.E., Callaghan, C.T., Blowes, S.A., Doncaster, C.P., Antão, L.H., *et al.* (2023). Understanding 'it depends' in ecology: a guide to hypothesising, visualising and interpreting statistical interactions. *Biological Reviews*, 98, 983–1002.

Spake, R., O'Dea, R.E., Nakagawa, S., Doncaster, C.P., Ryo, M., Callaghan, C.T., *et al.* (2022). Improving quantitative synthesis to achieve generality in ecology. *Nat Ecol Evol*, 6, 1818–1828.

Spirtes, P., Glymour, C. & Scheines, R. (2001). *Causation, Prediction, and Search*. 2nd edn. MIT Press, Cambridge, MA.

Splawa-Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles. Section 9. *Annals of Agricultural Sciences*, 1–51.

Stewart-Oaten, A. & Bence, J.R. (2001). Temporal and spatial variation in environmental impact assessment. *Ecol Monogr*, 71, 305–339.

Strømland, E. (2019). Preregistration and reproducibility. *J Econ Psychol*, 75, 102143.

Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1–21.

Suding, K.N. (2011). Toward an era of restoration in ecology: Successes, failures, and opportunities ahead. *Annu Rev Ecol Evol Syst*, 42, 465–487.

Sugihara, G., May, R., Ye, H., Hsieh, C.-H., Deyle, E., Fogarty, M., *et al.* (2012). Detecting causality in complex ecosystems. *Science (1979)*, 338, 496–500.

Suskiewicz, T.S., Byrnes, J.E.K., Steneck, R.S., Russell, R., Wilson, C.J. & Rasher, D.B. (2024). Ocean warming undermines the recovery resilience of New England kelp forests following a fishery-induced trophic cascade. *Ecology*, 105.

Tchetgen, E.J.T. & VanderWeele, T.J. (2012). On causal inference in the presence of interference. *Stat Methods Med Res*, 21, 55–75.

Tilman, D., Isbell, F. & Cowles, J.M. (2014). Biodiversity and ecosystem functioning. *Annu Rev Ecol Evol Syst*, 45, 471–493.

Tilman, D., Reich, P.B., Knops, J., Wedin, D., Mielke, T. & Lehman, C. (2001). Diversity and productivity in a long-term grassland experiment. *Science (1979)*, 294, 843–845.

Touchon, J.C. & McCoy, M.W. (2016). The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere*, 7.

Van Cleemput, E., Adler, P.B., Suding, K.N., Rebelo, A.J., Poulter, B., & Dee, L.E. (2024). Scaling-up ecological understanding with remote sensing and causal inference. *Trends Ecol Evol*.

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, Oxford, UK.

VanderWeele, T.J. & Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med*, 167, 268.

Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*, 113, 1228–1242.

Wauchope, Hannah.S., Amano, T., Geldmann, J., Johnston, A., Simmons, B.I., Sutherland, W.J., *et al.* (2021). Evaluating impact using time-series data. *Trends Ecol Evol*, 36, 196–205.

West, T.A.P., Börner, J., Sills, E.O. & Kontoleon, A. (2020). Overstated carbon emission reductions from voluntary REDD+ projects in the Brazilian Amazon. *Proc Natl Acad Sci U S A*, 117, 24188–24194.
Wing, C. & Bello-Gomez, R.A. (2018). Regression discontinuity and beyond. *American Journal of Evaluation*, 39, 91–108.
Witman, J.D. & Lamb, R.W. (2018). Persistent differences between coastal and offshore kelp forest communities in a warming Gulf of Maine. *PLoS One*, 13, e0189388.
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*, 485, 494–497.
Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2nd edn. MIT Press, Cambridge, MA.
Wright, S. (1921). Correlation and causation. *J Agric Res*, 20.
Wu, X., Sverdrup, E., Mastrandrea, M.D., Wara, M.W. & Wager, S. (2023). Low-intensity fires mitigate the risk of high-intensity wildfires in California's forests. *Sci Adv*, 9.
Wuepper, D. & Finger, R. (2023). Regression discontinuity designs in agricultural and environmental economics. *European Review of Agricultural Economics*.
Xu, X., Huang, A., Belle, E., De Frenne, P. & Jia, G. (2022). Protected areas provide thermal buffer against climate change. *Sci Adv*, 8, eabo0119.

Supporting Information for "Foundations and future directions for causal inference in ecological research," by Siegel and Dee.

This file contains:

Appendix 1: Extended glossary
Appendix 2: Curated reading list for causal inference
Appendix 3: RMarkdown tutorials

Appendix 1: Extended glossary

- **Average treatment effect of the treated (ATT):** the average treatment effect for the units that actually received the treatment.
- **Common support:** sufficient overlap in covariate values of treated and untreated units (and their propensity scores).
- **Complier average causal effect (CACE):** equivalent to the local average treatment effect (LATE). This is the treatment effect for units that were assigned to the treated group and did in fact receive the treatment, ignoring the effect of non-compliance.
- **Conditional average treatment effect (CATE):** the average treatment effect for a defined subgroup of the population.
- **Directed acyclic graph (DAG):** a diagram that maps the causal relationships between variables in a system as directional arrows or paths.
- **External validity:** the extent to which a study's results can be applied beyond the study sample.
- **Internal validity:** the extent to which a study accurately estimates a causal relationship within a study population.
- **Natural experiments:** events (i.e., not randomized experimental interventions) that divide a population into a treated and untreated group. These events are not guaranteed to have exogenous variation.
- **Non-complier:** While compilers are sample units that actually received the treatment to which they were assigned, non-compliers can take several forms: **never takers** (units that do not receive the treatment, regardless of whether they are assigned to the treated or control group), **always takers** (units that receive the treatment, regardless of whether they are assigned to the treated or control group), and **defiers** (units that receive the opposite of their treatment status).
- **Precision:** a metric of the similarity of multiple estimates to each other. Precision is distinct from bias: a biased estimate can be very precise.
- **Treatment effect (or causal effect):** the average causal effect of a variable on an outcome variable of scientific or policy interest. In the potential outcomes framework, we define a treatment effect as the difference in potential outcomes across two alternative states of the world. The average treatment effect is often reported as the treatment effect.
- **Two-way fixed effects:** models that incorporate both unit-specific (e.g., using an intercept per site, with no assumption that they are drawn from a common distribution) and time-specific dummy variables (e.g., a categorical variable for each year).
- **Unit:** the individual (e.g., physical site or organism) whose outcome is measured at a specific time following either exposure to treatment or non-exposure to treatment.

Appendix 2: Curated reading list for causal inference

Recommended textbooks

Core textbooks

Note: Some of these texts use examples to illustrate their points that are problematic (e.g., treating gender as a binary or studying post-colonial economic development without considering the violence of colonialism). We do not agree with the assumptions underlying these examples and we acknowledge the problems with them. The descriptions of methods are still some of the easiest to read out there.

- Angrist, JD & JS Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. (Princeton, NJ: Princeton University Press).
- Angrist, JD & JS Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.
- Cunningham, S. 2021. *Causal Inference: The Mixtape*. New Haven, CT: Yale University Press. https://mixtape.scunning.com/.
- Gerber, AS & DP Green. 2012. *Field Experiments: Design, Analysis and Interpretation*. New York, NY: W.W. Norton & Company.
- Huntington-Klein, N. 2022. The Effect: An Introduction to Research Design and Causality. Chapman & Hall. https://theeffectbook.net/.
- McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN, 2nd edition*. Boca Raton, FL: CRC Press.
- Morgan, SL & C Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.
- Rosenbaum, P. 2010. *Observational Studies*. New York, NY: Springer.

Texts with more technical discussions of causality:

- Holland, PW. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.2307/2289064.
- Heckman, JJ. 2000. Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, *115*(1), 45–97.
- Pearl, J. 2009. *Causality* (2nd ed.). Cambridge: Cambridge University Press.

A book geared towards a popular science audience:

- Pearl, J, & D Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA.

Readings for each key topic covered in the main text

Introduction to causal inference and counterfactual causality

- Hernán, MA, J Hsu, & B Healy. 2019. A second chance to get causal inference right: a classification of data science tasks. *Chance*, *32*(1). https://doi.org/10.1080/09332480.2019.1579578

- Angrist, JD and JS Pischke. 2008. Chapter 1: Questions about questions. In: *Mostly Harmless Econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.
- Gerber, AS & DP Green. 2012. Chapter 1: Introduction. *Field Experiments: Design, Analysis and Interpretation*. New York, NY: W.W. Norton & Company.

Introduction to the main frameworks for counterfactual and design-based causal inference

- Angrist, JD and JS Pischke. 2008. Chapter 2: The experimental ideal. In: *Mostly Harmless Econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.
- Hernán, MA. 2016. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, *26*(10), 674–680. https://doi.org/10.1016/j.annepidem.2016.08.016

Potential outcomes

Overview/methods

- Little, RJ & DB Rubin. 2000. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health*, *21*(1), 121–145. https://doi.org/10.1146/annurev.publhealth.21.1.121
- Rubin DB. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Applications and reviews for ecologists

- Larsen AE, K Meng, & BE Kendall. 2019. Causal analysis in control-impact ecological studies with observational data. *Methods in Ecology & Evolution*, *10*(7), 924–934. https://doi.org/10.1111/2041-210X.13190

DAGs and the structural causal model framework

Overview/methods

- Morgan, SL and C Winship. 2007. Chapter 3: Causal graphs. In: *Counterfactuals and Causal Inference: methods and principles for social research*. Cambridge, UK: Cambridge University Press.
- Cunningham, S. 2021. Chapter 3: Directed acyclic graphs. In: *Causal Inference: The Mixtape* (New Haven, CT: Yale University Press).
- Rohrer, JM. 2018. Thinking clearly about correlations and causation: graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Applications and reviews for ecologists

- Arif, S & MA MacNeil. 2023. Applying the structural causal model framework for observation causal inference in ecology. *Ecological Monographs*, *93*(1), e1554. https://doi.org/10.1002/ecm.1554
- Laubach, ZM, EJ Murray, KL Hoke, RJ Safran, & W Perng. 2021. A biologist's guide to model selection and causal inference. *Proceedings of the Royal Society B*, *288*, 20202815. https://doi.org/10.1098/rspb.2020.2815

Software and R package(s) for drawing and analyzing DAGs

- ggdag: https://cran.r-project.org/web/packages/ggdag/vignettes/intro-to-ggdag.html
- Shinydag: https://www.gerkelab.com/project/shinydag/
- TETRAD: https://sites.google.com/view/tetradcausal
- DAG program: https://hsz.dife.de/dag/
- dagR: Breitling, LP. 2010. dagR: A Suite of R Functions for Directed Acyclic Graphs. *Epidemiology*, *21*(4), 586-587. DOI: 10.1097/EDE.0b013e3181e09112

Randomized controlled experiments (RCTs) and experimental design (not comprehensive)

- Gerber, AS & DP Green. 2012. Chapter 2: Causal inference and experimentation. In: *Field Experiments: Design, Analysis and Interpretation*. New York, NY: W.W. Norton & Company.
- Kimmel K, LE Dee, ML Avolio, & PJ Ferraro. 2021. Causal assumptions and causal inference in ecological experiments. *Trends in Ecology & Evolution*, *36*(12), 1141–1152. https://doi.org/10.1016/j.tree.2021.08.008

Observational data and counterfactuals

- Imbens, GW. 2021. Causality in econometrics: methods in conversation with practice. Nobel Prize Lecture. https://www.nobelprize.org/prizes/economic-sciences/2021/imbens/lecture/
- Ferraro PJ. 2009. Counterfactual thinking and impact evaluation in environmental policy. *New Directions for Evaluation*, *2009*(122), 75–84. https://doi.org/10.1002/ev.297

Introduction to quasi-experimental methods

- Butsic V, DJ Lewis, VC Radeloff, M Baumann, & T Kuemmerle. 2017. Quasi-experimental methods enable stronger inferences from observational data in ecology. *Basic & Applied Ecology*, *19*, 1–10. https://doi.org/10.1016/j.baae.2017.01.005
    - The supplementary materials of this paper include R scripts demonstrating how the methods described in the paper work.
- Larsen AE, K Meng, & BE Kendall. 2019. Causal analysis in control-impact ecological studies with observational data. *Methods in Ecology & Evolution*, *10*(7), 924–934. https://doi.org/10.1111/2041-210X.13190
- Arif S & MA MacNeil. 2022. Utilizing causal diagrams across quasi-experimental approaches. *Ecosphere*, *13*(4), e4009. https://doi.org/10.1002/ecs2.4009

Difference-in-difference (DiD)

Overview/methods

- Angrist, JD and JS Pischke. 2015. Chapter 5: Differences-in-Differences. In *Mastering 'Metrics: The Path from Cause to Effect* (Princeton, NJ: Princeton University Press).
- Wauchope, HS, T Amano, J Geldmann, A Johnston, BI Simmons, WJ Sutherland, JPG Jones. 2021. Evaluating impact using time-series data. *Trends in Ecology & Evolution*, *36*(3), 196–205. https://doi.org/10.1016/j.tree.2020.11.001

Applications

- Simler-Williamson, AB & MJ Germino. 2022. Statistical considerations of nonrandom treatment applications reveal region-wide benefits of widespread post-fire restoration action. *Nature Communications*, *13*, 3472. https://doi.org/10.1038/s41467-022-31102-z
- McDermott, GR, KC Meng, GG McDonald, CJ Costello. 2018. The blue paradox: preemptive overfishing in marine reserves. *Proceedings of the National Academy of Sciences*, *116*(12), 5319-5325. https://doi.org/10.1073/pnas.1802862115

Extensions

- Callaway, B & PHC Sant'Anna. 2021. Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*, 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001
- D'Haultfœuille, X, S Hoderlein, Y Sasaki. 2023. Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. *Journal of Econometrics*, *234*, 664–690. https://doi.org/10.1016/j.jeconom.2022.07.003
- Goodman-Bacon, A. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*, 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

R package(s)

- did: Callaway, B & P Sant'Anna. 2021. did: Difference in differences. R package version 2.1.2, https://bcallaway11.github.io/did/.

Panel methods

Overview/methods

- Angrist, JD and JS Pischke. 2008. Chapter 5: Parallel worlds: Fixed effects, differences-in-differences, and panel data. In: *Mostly Harmless Econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press.
- Wooldridge, JM. (2010). Econometric analysis of cross section and panel data. MIT Press.
- Wooldridge, JM. (2015). Introductory econometrics: A modern approach. Cengage learning.
- Wooldridge, JM. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. https://dx.doi.org/10.2139/ssrn.3906345

Applications

- Dee, LE, PJ Ferraro, CN Severen, *et al.* 2023. Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. *Nature Communications*, *14*, 2607. https://doi.org/10.1038/s41467-023-37194-5
- Byrnes, JEK & LE Dee. 2024. Causal inference with observational data and unobserved confounding variables. *bioRxiv*. https://doi.org/10.1101/2024.02.26.582072
- Larsen, AE. 2013. Agricultural landscape simplification does not consistently drive insecticide use. *Proceedings of the National Academy of Sciences*, *110*(38), 15330–15335. https://doi.org/10.1073/pnas.1301900110
- Dudney, J, CE Willing, AJ Das, *et al.* 2021. Nonlinear shifts in infectious rust disease due to climate change. *Nature Communications*, *12*, 5102. https://doi.org/10.1038/s41467-021-25182-6

- Ratcliffe, H, A Kendig, S Vacek, D Carlson, M Ahlering, LE Dee. 2023. Extreme precipitation promotes invasion in managed grasslands. *Ecology*, *105*(1), e4190. https://doi.org/10.1002/ecy.4190.

R package(s)

- Bergé L. 2018. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*.

Instrumental variables

Overview/methods

- Angrist, JD and JS Pischke. 2015. Chapter 3: Instrumental Variables. In: *Mastering 'Metrics: The Path from Cause to Effect.* Princeton, NJ: Princeton University Press.
- Kendall, BE. 2015. A statistical symphony: instrumental variables reveal causality and control measurement error. In Fox, GA, S Negrete-Yankelevich, & VJ Sosa (eds.) *Ecological Statistics: Contemporary Theory and Application*. Oxford, UK: Oxford Academic.
- Imbens, GW. 2014. Instrumental variables: An econometrician's perspective. *Statistical Science*, *29*(3), 323–358. https://doi.org/10.1214/14-STS480

Applications

- Sims, ERE. 2010. Conservation and development: Evidence from Thai protected areas. Journal of Environmental Economics & Management, *60*(2), 94–114. https://doi.org/10.1016/j.jeem.2010.05.003
- MacDonald, AJ & EA Mordecai. 2020. Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing. *Proceedings of the National Academy of Sciences*, *116*(44), 22212–22218. https://doi.org/10.1073/pnas.1905315116
- Dee, LE, PJ Ferraro, CN Severen, *et al.* 2023. Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. *Nature Communications*, *14*, 2607. https://doi.org/10.1038/s41467-023-37194-5

R package(s)

- AER: Kleiber, C & Zeileis A. 2008. *Applied Econometrics with R*. Springer-Verlag, New York. https://CRAN.R-project.org/package=AER.
- ivreg: Fox, J, C Kleiber, & A Zeileis. 2023. ivreg: Instrumental-variables regression by '2SLS', '2SM', or '2SMM', with diagnostics. R package version 0.6-2, https://CRAN.R-project.org/package=ivreg.

Regression discontinuity designs (RDD)

Overview/methods

- Angrist, JD and JS Pischke. 2015. Chapter 4: Regression Discontinuity Designs. In *Mastering 'Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.
- Keele, L.J. & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, *23*, 127–155.
- Lee, D.S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *20*, 281–355.

Applications

- Englander G. 2019. Property rights and the protection of global marine resources. *Nature Sustainability*, *2*, 981–987. https://doi.org/10.1038/s41893-019-0389-9
- Noack, F, A Larsen, J Kamp, & C Levers. 2021. A bird's eye view of farm size and biodiversity: The ecological legacy of the iron curtain. *American Journal of Agricultural Economics*, *104*(4), 1460–1484. https://doi.org/10.1111/ajae.12274
- Burgess, R, FJM Costa, & BA Olken. 2019. National borders and the conservation of nature. *SocArXiv*. doi:10.31235/osf.io/67xg5.

R package(s)

- rdrobust: Calonico, S, MD Cattaneo, MH Farrell, & R Titiunik. 2022. rdrobust: Robust data-driven statistical inference in regression-discontinuity designs. R package version 2.1.1, https://CRAN.R-project.org/package=rdrobust.
- rddensity: Cattaneo, MD, M Jansson, & X Ma. 2023. rddensity: Manipulation testing based on density discontinuity. R package version 2.4, https://CRAN.R-project.org/package=rddensity.

Pre-regression matching and weighting

Overview/methods

- Ramsey, DSL, DM Forsyth, E Wright, M McKay, & I Westbrooke. 2019. Using propensity scores for causal inference in ecology: Options, considerations, and a case study. *Methods in Ecology & Evolution*, *10*(3), 320–331. https://doi.org/10.1111/2041-210X.13111
- Stuart, EA. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. doi: 10.1214/09-STS313.
- Schleicher, J, J Eklund, MD Barnes, J Geldmann, JA Oldekop, & JPG Jones. 2020. Statistical matching for conservation science. *Conservation Biology*, *34*(3), 538-549. https://doi.org/10.1111/cobi.13448

Applications

- Andam, KS, PJ Ferraro, A Pfaff, GA Sanchez-Azofeifa, & JA Robalino. 2008. Measuring the effectiveness of protected area networks in reducing deforestation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 16089–16094. https://doi.org/10.1073/pnas.0800437105
- Siegel, KJ, L Larsen, C Stephens, W Stewart, & V Butsic. 2022. Quantifying drivers of change in social-ecological systems: land management impacts wildfire probability in forests of the western US. *Regional Environmental Change*, *22*, 98. https://doi.org/10.1007/s10113-022-01950-y
- Siegel, KJ, L Macaulay, M Shapero, T Becchetti, S Larson, FE Mashiri, L Waks, L Larsen, V Butsic. 2022b. Impacts of livestock grazing on the probability of burning in wildfires vary by region and vegetation type in California. *Journal of Environmental Management*, *322*, 116092. https://doi.org/10.1016/j.jenvman.2022.116092
- Geldmann, J, A Manica, ND Burgess, L Coad, & A Balmford. 2019. A global-level assessment of the effectiveness of protected areas at resisting anthropogenic pressures. *PNAS*, *116*(46), 23209-23215. https://doi.org/10.1073/pnas.1908221116

R package(s)

- MatchIt: Ho DE, K Imai, G King, & EA Stuart. 2011. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1-28. https://doi.org/10.18637/jss.v042.i08
- WeightIt: Greifer N. 2022. WeightIt: Weighting for covariate balance in observational studies. R package version 0.13.1, https://CRAN.R-project.org/package=WeightIt.
- ipw: van der Wal, WM & RB Geskus. 2011. ipw: An R package for inverse probability weighting. *Journal of Statistical Software*, *43*(13), 1-23. DOI: 10.18637/jss.v043.i13

Generalizability

- Spake, R, RE O'Dea, S Nakagawa, *et al.* 2022. Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution*, *6*, 1818–1828. https://doi.org/10.1038/s41559-022-01891-z
- Korell, L, H Auge, JM Chase, C Harpole, & TM Knight. 2020. We need more realistic climate change experiments for understanding ecosystems of the future. *Global Change Biology*, *26*(2), 325–327. https://doi.org/10.1111/gcb.14797
- Spake R, AS Mori, M Beckmann, PA Martin, AP Christie, MC Duguid, & CP Doncaster. 2021. Implications of scale dependence for cross-study syntheses of biodiversity differences. *Ecology Letters*, *24*, 374–390. https://doi.org/10.1111/ele.13641

Appendix 3: RMarkdown tutorials

We provide RMarkdown tutorials demonstrating selected quasi-experimental methods for causal inference in a public GitHub repository: https://github.com/katherinesiegel/intro_causal_inf. The data used in the tutorials is stored in the same repository. A README file describes the contents of the repository.

The file "Matching_Weighting.Rmd" demonstrates the use of matching and weighting methods, using data compiled by Siegel et al. (2022) on the relationship between land ownership and burn probability in forests of the western US. Please refer to the original paper for details on the data sources (Siegel et al., 2022).

The file "Fixed_Effects" demonstrates the use of panel data and fixed effects models, using data compiled by Dee et al. (2023) on the relationship between grassland species richness and productivity. Please refer to the original paper for details on the data sources (Dee et al., 2023).

References

Dee, L. E., Ferraro, P. J., Severen, C. N., Kimmel, K. A., Borer, E. T., Byrnes, J. E. K., Clark, A. T., Hautier, Y., Hector, A., Raynaud, X., Reich, P. B., Wright, A. J., Arnillas, C. A., Davies, K. F., MacDougall, A., Mori, A. S., Smith, M. D., Adler, P. B., Bakker, J. D., … Loreau, M. (2023). Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. *Nature Communications*, *14*(1), 2607. https://doi.org/10.1038/s41467-023-37194-5

Siegel, K. J., Larsen, L., Stephens, C., Stewart, W., & Butsic, V. (2022). Quantifying drivers of change in social ecological systems: land management impacts wildfire probability in forests of the western US. *Regional Environmental Change*, *22*, 98. https://doi.org/https://doi.org/10.1007/s10113-022-01950-y