Quantifying impacts of policy and practice interventions on biodiversity and climate

Mark A. Bradford[1,2,3,*], Eli P. Fenichel[2], H. Dean Hosgood[4], Emily E. Oldfield[2,5], Alexander Polussa[1], Eric Potash[6], Luke Sanford[2], Oswald J. Schmitz[2], Sara E. Kuebbing[1,3]

[1]*The Forest School, Yale School of the Environment, Yale University, New Haven, CT, USA*

[2]*Yale School of the Environment, Yale University, New Haven, CT, USA*

[3]*Yale Center for Natural Carbon Capture, Yale University, New Haven, CT, USA*

[4]*Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA*

[5]*Environmental Defense Fund, 555 12th St., Suite 400, Washington D.C., USA*

[6]*Agroecosystem Sustainability Center, Institute for Sustainability, Energy, and Environment, University of Illinois Urbana-Champaign, Urbana, IL, USA*

*Author for correspondence: Mark A. Bradford, The Forest School, Yale School of the Environment, 360 Prospect St., New Haven, CT 06511, USA. Email: mark.bradford@yale.edu. Phone: +1 203 285 4921.

Submission type: **Perspective**

SUMMARY

There is urgent demand for ecosystem management interventions – targeted actions through policies and practices – that meaningfully address climate change and biodiversity loss while sustaining ecosystem delivery of water, food, fibre and fuel. Rigorous quantification of intervention outcomes is required for decision makers to identify, promote and scale effective interventions. Yet quantification of intervention effectiveness – i.e. their real-world impact – is hampered by limited use in ecology of causal approaches that generate counterfactual, empirical evidence at the scales of policy and practice actions. Here, we review the historical development of causal approaches and ecological experimentation, and emerging efforts to reunite the two. Reunification requires ecology to broaden its philosophical consideration of the validity and generalisability of evidence and to expand its experimental framework. Such an 'applied causal ecology' promises evidence that builds confidence that policy and practice interventions will sustain ecosystem services and achieve biodiversity and climate goals.

Keywords

applied ecology, counterfactual reasoning, coupled human and natural systems, ecosystem services, external validity, natural climate solutions, pragmatic studies, research design, translational research, science for policy.

INTRODUCTION

Humans rely on ecosystems for many services, including clean water, food, fibre and fuel. The continued supply of these services is dependent on sustaining and restoring the health of ecosystems[1]. Policy and practice interventions intended to achieve these goals, such as preventing forest conversion to other land uses and practicing regenerative agriculture, are viewed as essential for sustainable production and as solutions to help mitigate climate change and biodiversity losses[2]. As the human population grows, and quality of life increases demand more resources for each person, effective ecosystem management – through targeted policy and practice actions – is increasingly central to sustaining the health of people and the planet.

Ecosystems are complex entities, comprising a multitude of species and individuals interacting with one another and the physical environment. Ecologists study this complexity to understand how causes – such as precipitation amounts, forest fragmentation or species interactions – shape spatial and temporal changes in ecological populations, communities and elemental fluxes and stocks. Here, we use cause in a more specific manner, focusing on 'target causes' which we define as those that can be broadly manipulated through policy or practice interventions. For example, beyond some instances of cloud seeding, precipitation amount is not manipulable at scale, whereas broad scale policy and practice actions to prevent forest conversion or promote regenerative agriculture are common. Precipitation amount, however, may strongly modify or confound the impacts of these target causal interventions on ecological outcomes, such as soil carbon stocks[3,4]. 'Non-target causes' must then be accounted for in study design or analysis so that intervention effects are accurately estimated.

Collectively, quantifying impacts of ecosystem management interventions on ecological outcomes then demands (1) a focus on causes that are manipulable through policy and practice interventions, and (2) study designs that account for (a) influential non-target causes and (b) measurable counterfactual scenarios (e.g. unrealised cases where the intervention is not enacted). These outcomes include biodiversity, as well as ecosystem services such as crop yields and water quality.

Ecological outcomes of interest have expanded recently – beyond variables such as biodiversity, crop yields and water quality[1] – to include ecosystem-mediated removals and avoided emissions of greenhouse gases. This interest has resulted in a push for policy and practice interventions that manage ecosystems as natural climate solutions[5]. These interventions are needed, in addition to aggressive reductions in fossil fuel emissions, to limit warming[2]. These climate mitigation goals are increasingly being coupled with nature-restoration goals of ecosystem management across intergovernmental to sub-national levels[6]. The coupling recognizes that ecosystem services such as climate mitigation depend upon biodiversity and the resilience of ecosystems to stressors, which in turn depend on sustainable management practices[1]. Ecology is then faced with addressing questions about which policy and practice interventions are most effective, at real-world scales of ecosystem management, for sustaining healthy ecosystems and providing climate and other services.

The need to quantify the effectiveness of ecosystem management interventions to achieve biodiversity and climate goals has resulted in calls to conduct causal impact studies at the often massive, real-world scales at which policies and practices are enacted[7]. Studies of causal impact at such scales are routine in so-called 'causal fields' such as epidemiology, where

there is a recognized need for quantifying intervention outcomes under real-world conditions[8–10]. Researchers apply the PICO framework to quantify the effectiveness of interventions at such large scales. The framework involves defining the population (P) receiving the intervention, the intervention (I) as a cause that can be managed, control (C; i.e. counterfactual) scenarios which happen absent the intervention, and the outcome (O) to be measured[11,12].

The foundational sciences for mitigating biodiversity loss and climate change – ecology and biogeochemistry – rarely apply the PICO approach at real-world scales. These sciences instead favour approaches expressly designed to identify causes and their underlying mechanisms of action (i.e. causal pathways)[13]. Whereas such knowledge can inform ecosystem management, it does not provide direct evidence of the effectiveness of an intervention[14]. For example, using small-scale, controlled experiments to test whether genetic engineering results in more resilient, high-yielding crops uncovers mechanisms – such as how altering root systems affects plants' abilities to access soil water under drought conditions[15]. Yet, by itself, the approach falls short because it fails to provide understanding of impacts outside of controlled experimental conditions. Making the research fully policy and management relevant requires field trials and yield monitoring under commercial agricultural conditions to quantify the real-world effectiveness of population-level interventions[14–16]. For example, management effects on yields can differ by as much as 25-80% between controlled small-plot experiments and real-world scales because, in the latter, reduced management intensities and greater environmental variability may combine to reduce the effectiveness of interventions[17].

The most rigorous evidence for achieving scientific consensus on which interventions are likely most effective then requires mechanistic understanding and direct quantification[18–22].

That is, it requires knowledge both of important causal factors, including the mechanisms through which a target cause acts on the outcome, *and* the extent to which intervening to change that cause achieves the intended outcome at real-world scales. Ecology's primary approach of identifying causes and underlying mechanisms has left it incompletely equipped to inform policy and management interventions with the high level of confidence needed to affirm that real-world management interventions will be effective (Fig. 1).

To help redress this shortfall in knowledge generation for real-world effectiveness, we (1) review key differences in how ecology and causal fields treat core concepts about generalisability and validity of evidence; (2) present a brief history of how ecology departed, toward the end of the 1930s, from its roots in experimental causal impact studies; and (3) consider concepts that ecology must wrestle with to successfully translate the PICO framework. Such translation will expand the body of ecological knowledge – through development of an applied causal ecology – to generate the most comprehensive scientific evidence for selecting policies and practices with the greatest benefits for nature and people.

VALIDITY AND CAUSAL EVIDENCE

In applied causal inference, research questions address causal interventions that can be compared to a measurable counterfactual scenario, to quantify if the intervention has the intended outcome[23]. Such interventions can take the form of policy or practice changes. For instance, a causal policy question could ask whether direct payments are effective in encouraging farmers to adopt regenerative management practices. A causal practice question could ask how one of the management practices, such as cover cropping, might be best

126 adapted (e.g. through species choice) to reduce nitrous oxide – a potent greenhouse gas –
127 emissions from agricultural soils in different regions.

128 Different causal disciplines differ in their underlying terminologies, theory, statistical
129 concerns and objectives[11,20,24]. Such differences have made it challenging to develop a coherent
130 causal framework for ecology. For example, economics tends to emphasize the internal validity
131 of inference, whereas health sciences emphasize external validity[20,24]. Recent efforts to
132 translate causal approaches from these disciplines into ecology reflect these different
133 emphases[25,26], with consequent differences for the interpretation of evidence.

134 Specifically, internal validity is high when the investigator has strong evidence that the
135 intervention effect has been accurately quantified under the study conditions[20,27]. The
136 emphasis here is on correctly resolving the causal pathway – i.e. that the estimated effect on
137 the treated individuals is solely due to the intervention and not some alternate explanation.
138 The emphasis on internal validity reflects the challenge in quantitative social sciences on being
139 able to definitively associate outcomes with specific causal mechanisms in the face of the
140 complexity of socioeconomic systems. In contrast, external validity is high when there is strong
141 evidence that the measured effect is generalisable to the broader population, and in some
142 cases to other populations[27]. A focus on external validity helps decision makers understand
143 whether intervention effects are likely substantively meaningful[28]. Such knowledge is
144 imperative for public health interventions, such as vaccination campaigns. In health sciences,
145 controlled work to establish mechanism, followed by real-world trials to show intervention
146 effectiveness at population scales, are an integral part of regulatory impact assessments[28,29].

147 The different emphases on validity among causal disciplines are brought into sharp relief
148 through contrasting definitions of common causal terms, such as 'selection bias'[11]. In
149 economics, with a stronger focus on internal validity, the term relates to selection of individuals
150 from the study sample into treatment and control groups. Bias, introduced through systematic
151 differences between treatment and control groups that modify the intervention effect, reduces
152 the accuracy of the causal estimate for the study sample and hence internal validity that the
153 effect is caused by the intervention[12]. In contrast, selection bias in health disciplines such as
154 epidemiology concerns selection of the study sample before treatment assignment[11]. The focus
155 is instead on ensuring the sample is representative of – and hence the causal estimate has high
156 enough external validity to be generalisable to – the population that receives the intervention.
157 Here, we minimize use of causal terminology because, as the selection bias example shows, the
158 same term used differently among causal disciplines can have a decided effect on how policy
159 makers and practitioners should consider evidence.
160 We instead focus on how approaches to experimentation and generalisability affect
161 how evidence should be considered for decision-making. Ecology's treatment of internal and
162 external validity, for example, often departs from that of casual fields[27,30,31]. In economics, for
163 example, high internal validity is sought to correctly specify the causal pathway and hence
164 estimate the causal effect under specific conditions in a real-world context. By contrast, in
165 ecology high internal validity is sought by conducting experiments that heavily control for or
166 eliminate non-causal variation[30,31]. For example, high internal validity would be ascribed to
167 treatment effects in a greenhouse experiment in which potted wheat plants are grown under
168 regular versus drought conditions using natural (control) and compost-amended soils, where

169 compost is a climate-smart practice intended to build yield resilience to drought. Yet ecological
170 inference that the compost effect has high internal validity is incompatible with the definition in
171 a causal field. For the latter, the effect would need to be estimated for a sample of control and
172 compost-treated commercial farm fields occurring within a nexus of background variation of
173 non-target causes.

174 In ecology, external validity tends to be sought through what is often termed
175 observational studies[30,31]. For example, one might sample soils across the European Union from
176 different land uses such as arable fields and forests to identify causes that might explain
177 observed differences in soil carbon[32,33]. However, such a study design would not meet the
178 criteria of a causal study (Fig. 2). Specifically, it would merely explain variation in the outcome
179 and thus not provide sufficient evidence that a cause – such as land use change from cropland
180 to forest – would be effective at changing the target outcome[20,34–37]. In a causal discipline such
181 as epidemiology, this goal would be addressed by applying the PICO framework to specify the
182 intervention and population of interest. Applied to the ecological case study, a causal analysis
183 would need to consider an intervention such as reforestation and the population of arable
184 fields to which such a treatment will and will not be applied (Box 1). The study would need to
185 sample a representative subset of these fields, to establish control and intervention fields that
186 encompass population-level variation in non-target causes. Failure to account for such variation
187 in the sample could lead to biased quantification of the effectiveness of the reforestation
188 intervention if the non-target causes overwhelm the outcome disproportionally between the
189 sample and population, or the control and treated fields[38].

Representativeness ensures that the estimated effect of the intervention has high external validity. Such validity enables inferences to be generalised to the population of interest[27]. But here, again, a key concept – i.e. generalisability – has a different meaning in a causal discipline versus how it is commonly used in ecology[26]. Generalisability in ecology typically refers to whether fundamental mechanistic (i.e. process) understanding – such as evolution through natural selection – can become a principle that applies across multiple contexts[13]. In causal disciplines 'generalisability' is rooted in statistical inference—whether the estimated average treatment effect of the intervention can apply to the broader population (i.e. generalisability) and/or populations under other contexts (i.e. transferability)[26].

The key characteristics of evidence – captured by concepts of internal and external validity, and generalisability – clearly differ markedly in their general usage between ecology and causal disciplines. For the latter, concepts of validity and generalisability target the value of evidence for informing policy and practice actions. In ecology, usage of the same concepts creates barriers to the translation of approaches from causal disciplines. These barriers may help to explain why ecology – despite calls and demand for actionable ecological evidence[7,19,25,26,35] – remains limited in its ability to provide robust quantitative estimates that build confidence that interventions applied at real-world scales will be effective.

A BRIEF HISTORY OF ECOLOGY AND CAUSAL PHILOSOPHY

The causal philosophy underpinning accurate estimation of the effects of real-world interventions is absent from influential texts on the philosophical approaches used for knowledge generation in modern ecology[13]. Its absence even from recent ecology papers,

summarizing inferential approaches used in the discipline[39], emphasizes how disconnected ecology is from approaches to knowledge generation used in those causal disciplines that are effective at influencing policy and practice. Nascent efforts to connect ecology with these causal inference approaches reflect demand for rigorous applied evidence of the effectiveness of ecosystem interventions, at real-world scales, to achieve production, biodiversity and climate goals[40–45]. Such efforts have been centred on conservation science, but awareness of the need to expand causal inference to other ecological fields is growing[46–50]. For example, ecosystem ecology, the basis of efforts such as natural climate solutions, would seem an ideal candidate for approaches that quantify intervention effectiveness. Yet carbon accounting for agricultural and forest lands is dominated by 'indirect' quantification approaches, such as process-based models that 'scale' mechanistic knowledge to predict intervention effectiveness[14]. We suggest that the disconnect of ecology from causal inference reflects something much deeper than a lack of familiarity with the analytical approaches and concepts used in causal disciplines. Instead, we believe the disconnect is due to a fundamental departure of ecology and causal disciplines in what are valid ways to generate scientific evidence.

We trace this departure to the 1930s, when a long-running debate about the validity of experimental evidence between William Sealy Gosset and Ronald A. Fisher was cut-short by Gosset's sudden death in 1937. Gosset, the creator of Student's *t*-test and small-sample, causal analysis, was instrumental in advancing modern causal inference approaches[21,22]. Fisher, frequently lauded as the single most important figure in 20th century statistics, guided ecology – and many other scientific disciplines – away from experimentation at real-world scales to highly-controlled, small-plot studies[53]. Under these contrived conditions, Fisher had high

234 confidence that the observed effect of an experimental intervention could, from a single
235 experiment, be differentiated from zero[53]. This thinking, widely used by science and
236 amalgamated into the basis for null-hypothesis significance testing[54–56], identifies potential
237 causes but does not quantify the effect of the cause under real-world conditions.
238 The debate between the two scientists revolved around agricultural experimentation[57],
239 itself the predecessor of many contemporary experimental approaches used in ecology. In the
240 first decades of the 20th century, agricultural experimentation and applied statistics were
241 developing together to address pressing questions about food and beverage production. The
242 research was reported and discussed in leading statistical and general science journals,
243 including the *Journal of the Royal Statistical Society Series B*, *Biometrika,* and *Nature*[57–60]. Fisher
244 advocated for high levels of experimental control[53], likely in response to the lack of control and
245 replication in the decades of prior research made available to him when he began his tenure in
246 1919 at the Rothamsted Experimental Station, one of the oldest and renowned agricultural
247 institutions. Whereas the company Gosset worked for, the then *Arthur Guinness, Son & Co., Ltd*
248 – which likely funded at that time much of what was to become modern applied statistics –
249 made practice decisions only after data were collected under conditions representative of
250 agricultural operations[51,52].
251 Representative conditions, to Gosset and his employers, meant interventions applied at
252 the field level and replicated across multiple farms in a growing region. Gosset termed these
253 representative conditions 'large scale', employing study designs that matched weather, soil and
254 farming conditions to those of commercial-scale production. Gosset viewed such evidence to be
255 "*…necessary as a final demonstration…*" because "*…large scale conditions cannot be accurately*

produced in a wire cage…"[60]. By 'wire cage' he was referring to small experimental plots. The inability to recreate 'large scale conditions' in small plots meant that the Guinness company relied on practice-scale scientific evidence – integrated with mechanistic knowledge derived from controlled work – to make strategic operational decisions, such as those that determined which dominant barley varieties were grown in Ireland[57].

The Guinness company's return on investment in research was many times higher than their expenditure[61], likely because they developed, designed and practiced archetypal causal research. It required framing questions in quantitative ways that were specific to the applied context, asking for example which barley variety grew best on average for making stout (a combination of higher yield but low nitrogen grain) at the scale of the Irish growing region. They recognized that variety data for each individual field were too 'noisy' to make effective management decisions. Therefore, they focused on estimating the average treatment effect for the sample of fields they selected from the farms in the region.

The focus on the sample average was because the Guinness scientists understood that the influence of barley variety depended on other causes that strongly affected and interacted with variety to influence grain yield and chemistry. Specifically, within- and between-field differences in soil fertility, and regional and interannual variability in weather. To account for spatial variation in soil fertility they balanced the growing of new versus established varieties at each farm, with fields selected to be representative of conditions across the region[60]. To account for temporal variation in weather, they repeated their work across multiple years to identify which varieties consistently performed best[60]. Such careful attention to study design and repetition under different conditions allowed them to tease out the treatment effect of

barley variety from the effects of soil fertility, weather and chance alone[24,25]. That is, they focused on estimating the treatment effect of the cause they could manipulate (i.e. barley variety) – balanced across the causes they could not (e.g. weather) – in a way that led to high external validity in the estimated treatment effect of the management practice at the population-scale of its implementation (i.e. farms across the Irish barley-growing region).

It was perhaps the focus on balanced, as opposed to random, study designs that contributed to Gosset's and Fisher's departure in views about valid ways of knowing[57]. Both scientists sought experimental designs that accounted for non-target causes that strongly affect outcomes. Fisher believed that treatment randomization sufficed to solve this problem and developed statistical approaches such as ANOVA around this belief. Gosset, in contrast, favoured manually selecting treatment and control units across the range of variation of non-target causes. Such 'deterministic balancing' reflected the costs of large-scale research and the inherent 'noise' of working at that scale. Specifically, because costs necessitated fewer replicates (<30), Gosset argued that the scientist should use all knowledge of non-target causes available to them when designing experiments because 'balancing' is more precise, powerful and efficient[51,57,64]. Pearson, another influential 20th century statistician, commented on the later emergence in modern statistics of a Gosset-Fisher compromise (i.e. stratified random designs). He emphasized the value of Gosset's practical knowledge and how it is, "…*too often forgotten that mathematical models using probability theory* [i.e. statistical significance tests] *are there to provide an aid to human judgement.*"[65]

Whether Fisher agreed with such a sentiment is unknown, but he centered statistical significance tests in his book, "*The Design of Experiments*", published in 1935[53]. The book is

300 considered foundational to how scientists conduct experiments to this day by introducing ideas
301 such as null hypothesis significance testing and a *P* value threshold of <0.05. Fisher's definition
302 of the validity of evidence is made clear in this text, where he states that, "…*randomization will*
303 *suffice to guarantee the validity of the test of significance, by which the result of the experiment*
304 *is to be judged*" (p. 24). Fisher's focus was on designing experiments to randomize and minimize
305 the error (i.e. maximize precision) of the average outcomes under treatment and control
306 conditions. High precision is more likely to permit rejection of the null hypothesis that the
307 treatment has an effect not statistically distinguishable from zero. In the book, Fisher does not
308 consider the accuracy of the average treatment effect at real-world scales, thereby separating
309 experimental research from quantification of intervention effectiveness.
310 Indeed, through agricultural examples, the book guided experimentalists toward
311 reliance on statistical significance – whose shortcomings for reliable knowledge generation are
312 well understood[52,66–69]. Less appreciated, however, was Fisher's related guidance to use
313 smaller, more homogenous areas for pairing experimental treatments and controls. Such
314 groups (or 'blocks') then had more similar values of non-target causes, increasing precision (but
315 not necessarily at real-world scales; Fig. 3) of the effect of the agricultural intervention. In
316 Fisher's words, "…*within so large an area* [1 acre] *considerably greater soil heterogeneity will be*
317 *found, than would be the case if the blocks could be reduced in size to a quarter of an acre or*
318 *less*" (p.114). This small-plot Fisherian approach has dominated much of ecological
319 experimentation for at least the last 50 years[14], extending beyond agricultural experimentation
320 to topics such as whether biodiversity begets ecosystem function[70]. With that dominance, the
321 validity of evidence *sensu* Fisher's definition is detached from Gosset's (and modern causal

inference's) philosophy about internal and external validity and thereby leads to large uncertainty about ecology's current ability to inform intervention effectiveness at real-world scales.

RE-CONNECTING ECOLOGY WITH CAUSAL INFERENCE

The preconditions for re-connecting ecology with the causal philosophy of Gosset exist. First, ecologists are generally motivated to develop an understanding that directly informs policies and practices for ecosystem management that benefit nature and people[71–75]. There is deep appreciation that quantitative ecological evidence is increasingly demanded for policy efforts that require consideration of ecosystem services when evaluating potential costs and benefits of regulatory action[76–78]. Second, there is a growing ecological literature aimed at increasing awareness of causal inference[25,26,34–36]. Yet challenges to capitalising on this groundwork to build applied causal ecological knowledge remain. Specifically, different definitions of validity and generalisability unique to ecology versus causal disciplines are difficult to overcome. This barrier stems from ingrained differences in philosophies about the value and generation of knowledge. Overcoming this barrier requires ecologists to wrestle with 1) the value of applied ecology in producing knowledge, and 2) how to realise the PICO framework in terms of a) defining Populations (P) and individuals, along with how to measure Outcomes (O), and b) what constitutes a valid causal Intervention (I) and counterfactual Control (or Comparison: C).

342 *Value of applied ecology*

343 Ecology has a rich tradition of basic (or fundamental) knowledge generation about causes and
344 their mechanisms (Fig. 1). This knowledge is critical for building confidence that policy and
345 practice interventions are based on a sound understanding of the underlying science[18,21,79]. But
346 appropriately translating mechanistic understanding to policy and practice needs further
347 attention. For example, in the health sciences, poor characterization of the aetiology underlying
348 Alzheimer's disease and depression led to development of medications (the intervention)
349 targeting mechanisms that were not necessarily the primary causal pathways underlying the
350 conditions[80–82]. Causal science helped to reveal these shortcomings in presumed strong
351 fundamental understanding because the expected effectiveness of the medications at real-
352 world scales of practice was not realised[80,81]. When applying ecological science to policy and
353 practice, it then seems necessary to question whether knowledge is truly 'fundamental' when it
354 identifies causes and their mechanistic pathways, but not their quantitative impact under real-
355 world conditions.

356 Without embracing a truly applied causal ecology, we then suggest that fundamental
357 knowledge in ecology will remain woefully incomplete for understanding how causal change
358 affects real-world outcomes (Fig. 2). This synergy between foundational and causal research
359 belies the claim that stems from a worldview that the generation of new foundational
360 knowledge will cease if applied research is prioritised over basic[83]. The claim is rooted in the
361 ecological definition of generalisability, with its focus on generalising mechanistic principles[13].
362 This worldview holds that efforts to evaluate the effectiveness of a causal (mechanistic)

363 intervention at real world scales has limited value for developing general scientific
364 understanding[83,84]. The perception of limited value is false.
365 Indeed, Joshua Angrist, one of three recipients of the Nobel Prize in Economic Sciences
366 in 2021 for his work on 'causality' and natural experiments, argued that "*The process of*
367 *accumulating empirical evidence is rarely sexy in the unfolding, but accumulation is the*
368 *necessary road along which results become more general*"[84]. Moreover, the reduction in basic
369 knowledge generation has not been observed in health sciences. Instead, the combination of
370 basic and applied research has driven knowledge generation that produces foundational
371 mechanistic knowledge that is inherently needed to develop quantitative evidence of
372 intervention effectiveness at scale[80,81]. The importance of the two together – the 'weight of
373 evidence' approach – is instrumental today, and has been historically, for informing major
374 public health interventions[29,85,86] and was similarly employed by Guinness to meaningfully
375 improve agricultural management[60]. Ecology must wrestle with how it values and conducts
376 applied versus basic research if it is to incentivize uptake of its scientific knowledge into policy
377 and practice to produce effective solutions.
378 For example, most syntheses in ecology now take the form of meta-analyses[87] of
379 Fisherian-style small-scale experiments that are suited to identifying causes and understanding
380 mechanisms (Fig. 1). In such analyses, the average meta-analytical effect does not provide an
381 externally valid estimate of the average intervention effect. Yet the ecological estimates are
382 intended to inform intervention effects for such things as national-to-market level greenhouse
383 gas accounting[2,88–92]. In contrast, meta-analyses in causal fields commonly synthesize studies
384 conducted under real-world conditions that follow the PICO framework (Box 1), generating

385 average intervention effects for a comparable set of samples or populations[87]. Given the policy
386 stakes, ecology needs to place greater attention on when and where average effects can be
387 considered accurate at real-world scales of interventions. Doing so requires that ecology
388 develop evaluative criteria for the external validity of evidence[26,42,87,93,94], so that the value of
389 evidence can be appropriately considered for informing application.
390
391 *Populations and outcomes*
392 In the statistical design of experiments in ecology, the definition of a population and individual
393 is commonly mismatched with the scale at which interventions are applied. For example, in a
394 biodiversity-ecosystem function experiment, the population may consist of 120 individual plots
395 at a research station, with each individual constituting a plot of 2 square metres[70]. This
396 definition of a population and individuals does not lend itself well to quantifying the
397 effectiveness of an intervention, such as row crop diversification, applied at real-world scales[78].
398 In healthcare research, for example, the individual unit of observation is commonly a person or
399 group of individuals (e.g. a neighbourhood) – sampled from the target population – that will, at
400 least theoretically, be able to receive and benefit from the intervention (Box 1). In ecology,
401 then, the designation of the focal population and individual unit should be contingent upon the
402 organism or the area of land and/or water that could receive an intervention (Table 1). For
403 agroecosystems, the population may consist of individual agricultural fields within a growing
404 region, where the field is the scale at which management decisions are commonly applied. For
405 species of concern, the unit could be an individual organism, a group, or sub population,
406 sampled from the wider meta-population. Mixing of individuals, such as if a wolf moves from

one pack to another, or if two fields are combined into a larger one, will complicate the concept of an individual unit, requiring flexibility in study designs and analyses to deal with such situations. Hence, just like the modern definition of an ecosystem[95], a population and individual in a causal study need not be of fixed dimensions, but nonetheless it cannot be ambiguous. Instead, the population and individual must be defined by the intended intervention (Table 1). That is, experimentally quantifying effects in ways that reliably inform action must be predicated on first knowing the scales at which practitioners and policymakers enact interventions.

Translating the spatial scale of management intervention to the individual will not, however, guarantee valid causal inferences. For example, a common assumption in causal inference is that an individual's potential outcome depends only on their treatment assignment and not assignments of other individuals[96]. Yet outcomes of some ecosystem interventions do 'spillover' from one individual to influence observed outcomes of others (Table 1). For example, marine protected areas (MPAs) research has designated protected areas and adjacent non-protected areas as treatment and control individuals, respectively[97]. The measurement challenge is that MPAs are intended to increase fish stocks, within MPAs and beyond their borders. Fish abundances in controls are not then independent of treated (i.e. MPA) individuals[36]. Various approaches have been proposed to overcome spillover issues[69]. For example, one approach compares fish species – in control and MPA locations – that are caught commercially with those that are not. Compared to pre-MPA conditions, 'catch' fish species should see gains in population abundance disproportionally greater than 'non-catch' species in both control and MPA locations given that the 'non-catch' species were presumably not being

429 over-fished and so stand to gain less from an MPA[99]. For valid inferences, ecologists must
430 define, measure and/or model phenomena, like spillover effects, that result in inaccurate
431 quantification of intervention effects on outcomes (Table 1).

432 Outcome (i.e. measurement) challenges also emerge from assumptions about construct
433 validity[100]. That is, our ability to translate the concept of the variable we would like to evaluate
434 into a valid measurement (Table 1). For example, human health is a multidimensional construct
435 that cannot be directly observed and instead is inferred by measuring indicators such as blood
436 pressure. Even when the indicator metric is valid, how it is measured may influence construct
437 validity. Blood pressure, for example, provides different information about a person's health
438 when taken in a standing or supine position[101]. Such considerations are core to the broader
439 field of 'relational measurement theory', which focuses on the consistency and coherence of
440 what is measured and the theoretical concepts those observations are intended to represent. It
441 is an area of enquiry little explored in ecology[102,103] but with large implications for informing
442 interventions.

443 For example, attention to measurement theory in evolution has led to questioning of
444 the validity of how the concept of fitness is customarily proxied. If observations of fitness
445 outcomes are mismatched with actual fitness, then scientific evidence gathered to inform
446 which factors affect viability of animal and plant populations may be flawed[100,104]. Accordingly,
447 the flawed estimate will have a strong bearing on whether an intervention based on the
448 observed 'fitness outcome' will reliably translate to recovery of wild populations.
449

Interventions and controls

When quantifying intervention effectiveness, Holland[23] defined a causal intervention as something that could, in principle, be a treatment in an experiment, and hence manipulated. Following Rubin[21], this definition of cause applies to experimental and observational studies, where in the case of observational studies, manipulation of a causal intervention is not under the investigator's control. The inclusion of observational studies—that do not necessarily meet the 'gold standard' of manipulative experiment—expands availability of evidence. Such evidence may be more immediately available and less costly than generating experimental data at real-world scales. To quote Rubin[21], "*…it seems more reasonable to try to estimate the effects of the treatments from nonrandomized studies than to ignore these data and dream of the ideal experiment or make "armchair" decisions without the benefit of data analysis*."

Conceiving causes as potentially manipulable – in an externally valid manner – means that not every cause of interest to ecologists may be amenable to direct quantification of its 'intervention effect' under real-world conditions. For example, millions of dollars were spent on free-air CO_2 enrichment (i.e. FACE) experiments to query forest responses to elevated atmospheric CO_2[105]. These studies produced valuable knowledge by identifying mechanisms of plant and microbial response[106,107]. However, treatment plots received a large step change in CO_2 concentration, emulating projected levels in 2050, rather than the gradual change that ecosystems are experiencing. Step changes produce different outcomes, in part because physiological responses to step versus gradual changes are distinct[108]. Yet implementing a gradual increase in CO_2, along with associated changes (e.g. warming), would not generate knowledge at a rate faster than what is occurring anyway and a real-world counterfactual

scenario is not immediately apparent. For projecting ecosystem outcomes under changing atmospheric CO_2, contemporary ecological approaches – such as integrating controlled experimental knowledge through process-based models (Figs. 1, 2) – then seems necessary. But in many other policy- and practice-relevant instances, such as in the implementation of wildlife corridors or regenerative agriculture, ecology is presented with causes that are amenable to externally valid manipulation and hence opportunities for direct quantification of intervention effectiveness.

When a valid causal intervention is identified, its effectiveness is estimated by comparing the treatment with the control condition. The accuracy of this estimate depends on decisions about what constitutes an appropriate comparison scenario for non-treated individuals[109,110]. The scenario provides information on the counterfactual, i.e. what is presumed to have occurred in the absence of the intervention. The counterfactual need not come from a simultaneous control group and could be an historical baseline, or a weighted average of individuals approximating the non-intervention condition (e.g. a synthetic control[111]). Confidence in the accuracy of the intervention effect depends on the suitability of the counterfactual (Table 1). For example, much of the debate about how conservation efforts affect forest carbon stocks revolves around choice of 'control' individuals[112]. Specifically, representativeness of the comparison is greatest when the treatment and control individuals each span the same range in values of non-target causes and are equally likely to have been assigned the intervention. For example, forest harvest effects on soil carbon stocks are commonly estimated using forest stands already protected from timber harvest as the comparison group[113]. Such stands may be a poor baseline because they have confounding

494 characteristics, such as proximity to wetland or access restricted by topography[113], that make
495 them unlikely to be harvested and distinct from treated stands in ways that bias the estimated
496 effects of the practice intervention.

497 The comparison group should also receive the policy or practice that would be most
498 likely to occur absent the intervention. In forest carbon accounting, debate about which non-
499 intervention scenarios are valid determine whether interventions are viewed to create sources
500 or sinks for greenhouse gases[114,115]. For instance, if rates of assumed deforestation in the
501 comparison group are higher than what would have truly occurred absent the intervention,
502 carbon gains are overestimated[112]. Such a situation typifies what Holland[23] referred to as 'the
503 fundamental problem of causal inference'. That is, an individual cannot both receive the
504 treatment and control, meaning that the true causal effect of the intervention cannot be
505 determined but instead must be estimated using a valid counterfactual. Confidence in the
506 validity of the inference emerges when similar treatment effects are estimated for different
507 sets of reasonable assumptions (Fig. 3) about, for example, what is a suitable baseline for
508 comparison[116–118]. Ecology's rich mechanistic knowledge, when combined with causal
509 understanding, will be critical for determining these baselines. Consequently, the definition of a
510 causal intervention (and its comparator), like a population and individual, must emerge from
511 ecological knowledge about how a system works and how human decisions about interventions
512 translate to how they are implemented versus when they are not.
513

Conclusions

Adopting a causal framework is not a panacea for generating evidence about all causes of interest to ecology. Some causes, such as elevated greenhouse gas concentrations and climate warming, may not be amenable to the PICO framework in ways that generate high external validity and timely estimates of treatment impact. Yet for those causes that are amenable – such as natural climate solutions – adoption of a PICO framework is essential if ecologists are to accurately and reliably estimate the effectiveness of policies and practices centred on ecosystem management to realise production, biodiversity and climate goals. Connecting this new knowledge with contemporary ecological understanding, of potential causes and their mechanistic pathways, will support efforts to achieve expert consensus on which interventions are likely most beneficial. Guinness scientists showed, at the beginning of the last century, how such an integrated approach – an 'applied causal ecology' – could serve as the basis for impactful decisions around agricultural practices. If this integrated causal approach can be resurrected in ecology, we expect it to be instrumental in providing evidence to set policies and practices that most effectively restore the health of nature.

Acknowledgements

MAB thanks Eibhlin Colgan and Leanne Harrington of Diageo plc for making available, through the Guinness Archive in Dublin, Ireland, internal and external reports and correspondence from 1896-1945 for Guinness employees, and a wide range of external government and academic scientists, relating broadly to the large-scale, co-operative, experimental agricultural work in Ireland that was primarily funded by the then "Arthur Guinness, Son and Co., Ltd". Interpretation of published papers by Gosset (aka. Student) was strongly informed by the archival material. MAB thanks current and former members of his research group for many

causal inference discussions and related study designs, in particular Dan Maynard, Robert Warren, Fiona Jevon, Michael Culbertson, Steve Wood and Elisabeth Ward.

MAB and SEK conducted this work as a project of the Yale Applied Science Synthesis Program, which is an initiative of The Forest School (at the Yale School of the Environment) and the Yale Center for Natural Carbon Capture. Additional support was from the Environmental Defense Fund with awards from The Earth Fund, King Philanthropies, and Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin. The perspectives presented in this paper are those of the authors and not necessarily those of entities which funded the work.

Author contributions

All authors contributed to the ideas and content of this manuscript through collaborative work and discussions about robust quantification of the effects of manipulating a cause at the scales of practice and policy. MAB drafted the manuscript with input from SEK, AP, EF and LS, whose content was then shaped by comments and edits from all authors.

Declaration of interests

The authors declare no competing interests.

References and Notes

1. Millennium-Ecosystem-Assessment. *Ecosystems and Human Well-Being: Biodiversity Synthesis*. (2005).

2. IPCC. Summary for Policymakers. in *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* (eds Shukla, P. R. et al.) 48 (Intergovernmental Panel on Climate Change, Cambridge, UK and New York, NY, USA, 2022).

3. Powers, J. S., Corre, M. D., Twine, T. E. & Veldkamp, E. Geographic bias of field observations of soil carbon stocks with tropical land-use changes precludes spatial extrapolation. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 6318–6322 (2011).

4. Mitchell, E. *et al.* Making soil carbon credits work for climate change mitigation. *Carbon Manage* **15**, 2430780 (2024).

5. Griscom, B. W. *et al.* Natural climate solutions. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 11645–11650 (2017).

6. Pörtner, H.-O. *et al.* Overcoming the coupled climate and biodiversity crises and their societal impacts. *Science* **380**, eabl4881 (2023).

7. Griscom, B. W. *et al.* We need the largest experiments on Earth to achieve our climate targets. *Global Change Biology* **31**, e70555 (2025).

8. Zuidgeest, M. G. P. *et al.* Series: Pragmatic trials and real world evidence: Paper 1. Introduction. *Journal of Clinical Epidemiology* **88**, 7–13 (2017).

9. Dang, A. Real-World Evidence: A Primer. *Pharm Med* **37**, 25–36 (2023).

10. Leary, A., Besse, B. & André, F. The need for pragmatic, affordable, and practice-changing real-life clinical trials in oncology. *The Lancet* **403**, 406–408 (2024).

11. Lash, T. L., VanderWeele, T. J., Haneuse, S. & Rothman, K. J. *Modern Epidemiology*. (Wolters Kluwer, Philadelphia, USA).

12. Huntington-Klein, N. *The Effect: An Introduction to Research Design and Causality*. (CRC Press, Boca Raton, FL, 2022).

13. Hilborn, R. & Mangel, M. *The Ecological Detective: Confronting Models with Data*. (Princeton Uni. Press, Princeton, NJ, USA, 1997).

14. Bradford, M. A., Kuebbing, S. E., Polussa, A., Sanderman, J. & Oldfield, E. E. Upstream data need to prove soil carbon as a climate solution. *Nat. Clim. Chang.* **15**, 1013–1016 (2025).
15. Nature Food Editors. Rethinking field trials. *Nat. Plants* **11**, 935–936 (2025).
16. Khaipho-Burch, M. *et al.* Scale up trials to validate modified crops' benefits.
17. Kravchenko, A. N., Snapp, S. S. & Robertson, G. P. Field-scale experiments reveal persistent yield gaps in low-input and organic cropping systems. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 926–931 (2017).
18. Davey Smith, G. Post–modern epidemiology: when methods meet matter. *American Journal of Epidemiology* **188**, 1410–1419 (2019).
19. Grace, J. B. An integrative paradigm for building causal knowledge. *Ecological Monographs* e1628 (2024) doi:10.1002/ecm.1628.
20. Gelman, A. & Imbens, G. *Why Ask Why? Forward Causal Inference and Reverse Causal Questions*. w19614 http://www.nber.org/papers/w19614.pdf (2013) doi:10.3386/w19614.
21. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701 (1974).
22. Jones, J. P. G. & Shreedhar, G. The causal revolution in biodiversity conservation. *Nat Hum Behav* **8**, 1236–1239 (2024).
23. Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–960 (1986).
24. Gelman, A. Causality and statistical learning. *Am J Sociol* **117**, 955–966 (2011).
25. Siegel, K. & Dee, L. E. Foundations and Future Directions for Causal Inference in Ecological Research. *Ecology Letters* **28**, e70053 (2025).

26. Spake, R. *et al.* Improving quantitative synthesis to achieve generality in ecology. *Nat Ecol Evol* **6**, 1818–1828 (2022).
27. Rothwell, P. M. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet* **365**, 82–93 (2005).
28. Imbens, G. W. Causal inference in the social sciences. *Annual Review of Statistics and its Application* **11**, 123–152 (2024).
29. Goldman, G. T. & Dominici, F. Don't abandon evidence and process on air pollution policy. *Science* **363**, 1398–1400 (2019).
30. Bradford, M. A. & Reynolds, J. F. Scaling terrestrial biogeochemical processes: contrasting intact and model experimental systems. in *Scaling and uncertainty analysis in ecology: methods and applications* (eds Wu, J., Jones, B., Li, H. & Loucks, O. L.) 107–128 (Springer, Amsterdam, 2006).
31. Naeem, S. Experimental validity and ecological scale as criteria for evaluating research programs. in *Scaling relations in experimental ecology* (eds Gardner, R. H., Kemp, W. M., Kennedy, V. S. & Petersen, J. E.) 223–250 (Columbia University Press, New York, 2001).
32. Jones, A., Fernandes-Ugalde, O., Scarpa, S. & Eiselt. *LUCAS 2022*. (Publications Office of the European Union, Luxembourg, 2021).
33. Tóth, G., Jones, A. & Montanarella, L. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environ Monit Assess* **185**, 7409–7425 (2013).

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

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

36. Ferraro, P. J., Sanchirico, J. N. & Smith, M. D. Causal inference in coupled human and natural systems. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5311–5318 (2019).

37. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. (Chapman and Hall/CRC, 2020). doi:10.1201/9780429029608.

38. Breznau, N. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences* **119**, e2203150119 (2022).

39. Tredennick, A. T., Hooker, G., Ellner, S. P. & Adler, P. B. A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology* https://doi.org/10.1002/ecy.3336 (2021) doi:10.1002/ecy.3336.

40. Adamowicz, W. *et al.* Assessing ecological infrastructure investments. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5254–5261 (2019).

41. Christie, A. P. *et al.* Poor availability of context-specific evidence hampers decision-making in conservation. *Biological Conservation* **248**, 108666 (2020).

42. Martin, D. J., Kroll, A. J. & Knoth, J. L. An evidence-based review of the effectiveness of riparian buffers to maintain stream temperature and stream-associated amphibian

populations in the Pacific Northwest of Canada and the United States. *Forest Ecology and Management* **491**, 119190 (2021).

43. Wiik, E. *et al.* Experimental evaluation of the impact of a payment for environmental services program on deforestation. *Conservat Sci and Prac* **1**, e8 (2019).

44. Adams, V. M. *et al.* Multiple-use protected areas are critical to equitable and effective conservation. *One Earth* **6**, 1173–1189 (2023).

45. Adams, V. M., Barnes, M. & Pressey, R. L. Shortfalls in conservation evidence: Moving from ecological effects of interventions to policy evaluation. *One Earth* **1**, 62–75 (2019).

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

47. Palmer, M. A., Kramer, J. G., Boyd, J. & Hawthorne, D. Practices for facilitating interdisciplinary synthetic research: the National Socio-Environmental Synthesis Center (SESYNC). *Current Opinion in Environmental Sustainability* **19**, 111–122 (2016).

48. Meyfroidt, P. Emerging agricultural expansion in northern regions: Insights from land-use research. *One Earth* **4**, 1661–1664 (2021).

49. Eigenbrod, F. *et al.* Identifying Agricultural Frontiers for Modeling Global Cropland Expansion. *One Earth* **3**, 504–514 (2020).

50. Vansant, E. *et al.* Multipurpose trees on farms can improve nutrition in Malawi. *One Earth* **8**, 101165 (2025).

51. Ziliak, S. T. W.S. Gosset and some neglected concepts in experimental statistics: Guinnessometrics II. *J Wine Econ* **6**, 252–277 (2011).

52. Ziliak, S. T. How large are your *G* -Values? Try Gosset's Guinnessometrics when a little " *p* " is not enough. *Am Stat* **73**, 281–290 (2019).
53. Fisher, R. A. *The Design of Experiments*. (Oliver and Boyd, Edinburgh, 1935).
54. Christensen, R. Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician* **59**, 121–126 (2005).
55. Goodman, S. N. p values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology* **137**, 485–496 (1993).
56. Perezgonzalez, J. D. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front. Psychol.* **6**, 223 (2015).
57. Gosset, W. S. Co-operation in large-scale experiments. *Supplement to the Journal of the Royal Statistical Society* **3**, 115–136 (1936).
58. Fisher, R. A. The half-drill strip system agricultural experiments. *Nature* **138**, 1101 (1936).
59. Student. The half-drill strip system agricultural experiments. *Nature* **138**, 971–972 (1936).
60. Student. On testing varieties of cereals. *Biometrika* **15**, 271–293 (1923).
61. Dennison, S. R. & MacDonagh, O. *Guinness 1886-1939 From Incorporation to the Second World War*. (Cork University Press, Cork, Ireland, 1998).
62. Gelman, A. & Carlin, J. Beyond power calculations: Assessing type S (Sign) and yype M (Magnitude) errors. *Perspect Psychol Sci* **9**, 641–651 (2014).
63. Lemoine, N. P. *et al.* Underappreciated problems of low replication in ecological field studies. *Ecology* **97**, 2554–2561 (2016).
64. Student. Comparison between balanced and random arrangements of field plots. *Biometrika* **29**, 363–378 (1938).

65. Pearson, E. S. *'Student': A Statistical Biography of William Sealy Gosset*. (Clarendon Press, Oxford, 1990).
66. Wasserstein, R. L. & Lazar, N. A. The ASA statement on *p* -values: Context, process, and purpose. *The American Statistician* **70**, 129–133 (2016).
67. Wasserstein, R. L., Schirm, A. L. & Lazar, N. A. Moving to a world beyond "$p < 0.05$". *Am Stat* **73**, 1–19 (2019).
68. Greenland, S. *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* **31**, 337–350 (2016).
69. Goodman, S. N. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* **130**, 995 (1999).
70. Hooper, D. U. *et al.* Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological Monographs* **75**, 3–35 (2005).
71. Kuebbing, S. E. *et al.* Long-term research in ecology and evolution: a survey of challenges and opportunities. *Ecological Monographs* **88**, 245–258 (2018).
72. Kluger, D. M., Owen, A. B. & Lobell, D. B. Combining randomized field experiments with observational satellite data to assess the benefits of crop rotations on yields. *Environ. Res. Lett.* **17**, 044066 (2022).
73. Noack, F. *et al.* Environmental impacts of genetically modified crops. *Science* **385**, eado9340 (2024).
74. Sutherland, W. J. *et al.* Identification of 100 fundamental ecological questions. *Journal of Ecology* **101**, 58–67 (2013).

75. Yang, Y. *et al.* Climate change exacerbates the environmental impacts of agriculture. *Science* **385**, eadn3747 (2024).
76. HM Treasury. *The Green Book*. https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government/the-green-book-2020 (2022).
77. Office of Management and Budget. *Guidance for Assessing Changes in Environmental and Ecosystem Services in Benefit-Cost Analysis*. https://www.whitehouse.gov/wp-content/uploads/2024/02/ESGuidance.pdf (2024).
78. Fenichel, E. P., Dean, M. F. & Schmitz, O. J. The path to scientifically sound biodiversity valuation in the context of the Global Biodiversity Framework. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2319077121 (2024).
79. Lehman, J. T. The goal of understanding in limnology. *Limnology & Oceanography* **31**, 1160–1166 (1986).
80. Moncrieff, J. *Chemically Imbalanced: The Making and Unmaking of the Serotonin Myth*. (Flint, Cheltenham, UK, 2025).
81. Frisoni, G. B. *et al.* Alzheimer's disease outlook: controversies and future directions. *The Lancet* **406**, 1424–1442 (2025).
82. Moncrieff, J. *et al.* The serotonin theory of depression: a systematic umbrella review of the evidence. *Mol Psychiatry* **28**, 3243–3256 (2023).
83. Courchamp, F. *et al.* Fundamental ecology is fundamental. *Trends in Ecology & Evolution* **30**, 9–16 (2015).

84. Angrist, J. D. & Pischke, J.-S. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* **24**, 3–30 (2010).

85. Samet, J. M. *et al.* The IARC Monographs: Updated procedures for modern and transparent evidence synthesis in cancer hazard identification. *JNCI: Journal of the National Cancer Institute* **112**, 30–37 (2020).

86. Cornfield, J. *et al.* Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22**, 173–203 (1959).

87. Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. Meta-analysis and the science of research synthesis. *Nature* **555**, 175–182 (2018).

88. Eggleston, L., Buendia, K., Miwa, K., Ngara, T. & Tanabe, K. *2006 IPCC Guidelines for National Greenhouse Gas Inventories, Prepared by the National Greenhouse Gas Inventories Programme*. (2006).

89. Garsia, A., Moinet, A., Vazquez, C., Creamer, R. E. & Moinet, G. Y. K. The challenge of selecting an appropriate soil organic carbon simulation model: A comprehensive global review and validation assessment. *Global Change Biol* **29**, 5760–5774 (2023).

90. Le Noë, J. *et al.* Soil organic carbon models need independent time-series validation for reliable prediction. *Commun Earth Environ* **4**, 158 (2023).

91. Oldfield, E. E. *et al.* Crediting agricultural soil carbon sequestration. *Science* **375**, 1222–1225 (2022).

92. Yona, L., Cashore, B., Jackson, R. B., Ometto, J. & Bradford, M. A. Refining national greenhouse gas inventories. *Ambio* **49**, 1581–1586 (2020).

93. Eagle, A. J. *et al.* Meta-analysis constrained by data: recommendations to improve relevance of nutrient management research. *Agronomy Journal* **109**, 2441–2449 (2017).
94. Simonsohn, U., Simmons, J. & Nelson, L. D. Above averaging in literature reviews. *Nat Rev Psychol* **1**, 551–552 (2022).
95. Weathers, K. C., Strayer, D. L. & Likens, G. E. *Fundamentals of Ecosystem Science*. (Elsevier/AP, Waltham, MA, 2013).
96. Pearce, N. & Greenland, S. Confounding and interaction. in *Handbook of Epidemiology* 659–684 (Springer, New York, 2014).
97. Rodríguez-Rodríguez, D. & Martínez-Vega, J. Ecological effectiveness of marine protected areas across the globe in the scientific literature. in *Advances in Marine Biology* vol. 92 129–153 (Elsevier, 2022).
98. Viviano, D. Policy targeting under network interference. *Review of Economic Studies* rdae041 (2024) doi:10.1093/restud/rdae041.
99. Ovando, D. *et al.* Assessing the population-level conservation effects of marine protected areas. *Conservation Biology* **35**, 1861–1870 (2021).
100. Houle, D., Pélabon, C., Wagner, G. P. & Hansen, T. F. Measurement and meaning in biology. *The Quarterly Review of Biology* **86**, 3–34 (2011).
101. Sechrest, L. Validity of measures is no simple matter. *Health Services Research* **40**, 1584–1604 (2005).
102. Wolman, A. G. Measurement and meaningfulness in conservation science. *Conservation Biology* **20**, 1626–1634 (2006).

103. Romesburg, H. C. Wildlife science: Gaining reliable knowledge. *The Journal of Wildlife Management* **45**, 293 (1981).

104. Wagner, G. P. The measurement theory of fitness. *Evolution* **64**, 1358–1376 (2010).

105. U.S. DOE. *U.S. DOE. 2020. U.S. Department of Energy Free-Air CO2 Enrichment Experiments: FACE Results, Lessons, and Legacy.* 115 DOI:10.2172/1615612. (2020).

106. Norby, R. J. *et al.* Model–data synthesis for the next generation of forest free-air CO_2 enrichment (FACE) experiments. *New Phytologist* **209**, 17–28 (2016).

107. Maschler, J. *et al.* Links across ecological scales: Plant biomass responses to elevated CO_2. *Global Change Biology* **28**, 6115–6134 (2022).

108. Luo, Y. & Reynolds, J. F. Validity of extrapolating field CO_2 experiments to predict carbon sequestration in natural ecosystems. *Ecology* **80**, 1568–1583 (1999).

109. Greenstone, M. & Gayer, T. Quasi-experimental and experimental approaches to environmental economics. *Journal of Environmental Economics and Management* **57**, 21–44 (2009).

110. Rubin, D. B. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2**, (2008).

111. Ben-Michael, E., Feller, A. & Rothstein, J. Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84**, 351–381 (2022).

112. West, T. A. P. *et al.* Action needed to make carbon offsets from forest conservation work for climate change mitigation. *Science* **381**, 873–877 (2023).

113. Ward, E. B., Ashton, M. S., Wikle, J. L., Duguid, M. & Bradford, M. A. Local controls modify the effects of timber harvesting on surface soil carbon and nitrogen in a temperate hardwood forest. *Forest Ecology and Management* **572**, 122268 (2024).
114. Searchinger, T. D., Berry, S. & Peng, L. Reply to: Carbon implications of wood harvesting and forest management. *Nature* **646**, E20–E22 (2025).
115. Sohngen, B., Baker, J. S., Favero, A. & Daigneault, A. Carbon implications of wood harvesting and forest management. *Nature* **646**, E18–E19 (2025).
116. Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, 1866–1886 (2017).
117. Munafò, M. R. & Davey Smith, G. Robust research needs many lines of evidence. *Nature* **553**, 399–401 (2018).
118. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat Hum Behav* **4**, 1208–1214 (2020).
119. Auspurg, K. & Brüderl, J. Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "Many Analysts, One Data Set" project. *Socius* **7**, 237802312110244 (2021).
120. Huntington-Klein, N. *et al.* The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry* **59**, 944–960 (2021).
121. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspect Psychol Sci* **11**, 702–712 (2016).
122. Catford, J. A., Wilson, J. R. U., Pyšek, P., Hulme, P. E. & Duncan, R. P. Addressing context dependence in ecology. *Trends in Ecology & Evolution* **37**, 158–170 (2022).

123. Steenland, K. *et al.* Risk of bias assessments and evidence syntheses for observational epidemiologic studies of environmental and occupational exposures: strengths and limitations. *Environ Health Perspect* **128**, 095002 (2020).
124. Jean-Louis, G. & Seixas, A. A. The value of decentralized clinical trials: Inclusion, accessibility, and innovation. *Science* **385**, eadq4994 (2024).
125. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* **210**, 2–21 (2018).
126. Larsen, A. E., Quandt, A., Foxfoot, I., Parker, N. & Sousa, D. The effect of agricultural land retirement on pesticide use. *Science of The Total Environment* **896**, 165224 (2023).
127. Fierer, N., Wood, S. A. & Bueno De Mesquita, C. P. How microbes can, and cannot, be used to assess soil health. *Soil Biology and Biochemistry* **153**, 108111 (2021).
128. Köninger, J., Panagos, P., Jones, A., Briones, M. J. I. & Orgiazzi, A. In defence of soil biodiversity: Towards an inclusive protection in the European Union. *Biological Conservation* **268**, 109475 (2022).
129. Guerra, C. A. *et al.* Tracking, targeting, and conserving soil biodiversity. *Science* **371**, 239–241 (2021).
130. Wall, D. H., Nielsen, U. N. & Six, J. Soil biodiversity and human health. *Nature* **528**, 69–76 (2015).
131. Siles, J. A. *et al.* Land-use- and climate-mediated variations in soil bacterial and fungal biomass across Europe and their driving factors. *Geoderma* **434**, 116474 (2023).
132. Smith, L. C. *et al.* Large-scale drivers of relationships between soil microbial properties and organic carbon across Europe. *Global Ecol Biogeogr* **30**, 2070–2083 (2021).

133. VandenBygaart, A. J. & Angers, D. A. Towards accurate measurements of soil organic carbon stock change in agroecosystems. *Can. J. Soil. Sci.* **86**, 465–471 (2006).
134. Bradford, M. A. *et al.* Testing the feasibility of quantifying change in agricultural soil carbon stocks through empirical sampling. *Geoderma* **440**, 116719 (2023).
135. Chappell, A. & Viscarra Rossel, R. A. The importance of sampling support for explaining change in soil organic carbon. *Geoderma* **193–194**, 323–325 (2013).
136. Maillard, É., McConkey, B. G. & Angers, D. A. Increased uncertainty in soil carbon stock measurement with spatial scale and sampling profile depth in world grasslands: A systematic analysis. *Agriculture, Ecosystems & Environment* **236**, 268–276 (2017).
137. Robinson, S. V. J., Nguyen, L. H. & Galpern, P. Livin' on the edge: Precision yield data shows evidence of ecosystem services from field boundaries. *Agriculture, Ecosystems & Environment* **333**, 107956 (2022).
138. VanderWeele, T. J. Explanation in causal inference: developments in mediation and interaction. *Int. J. Epidemiol.* dyw277 (2016) doi:10.1093/ije/dyw277.

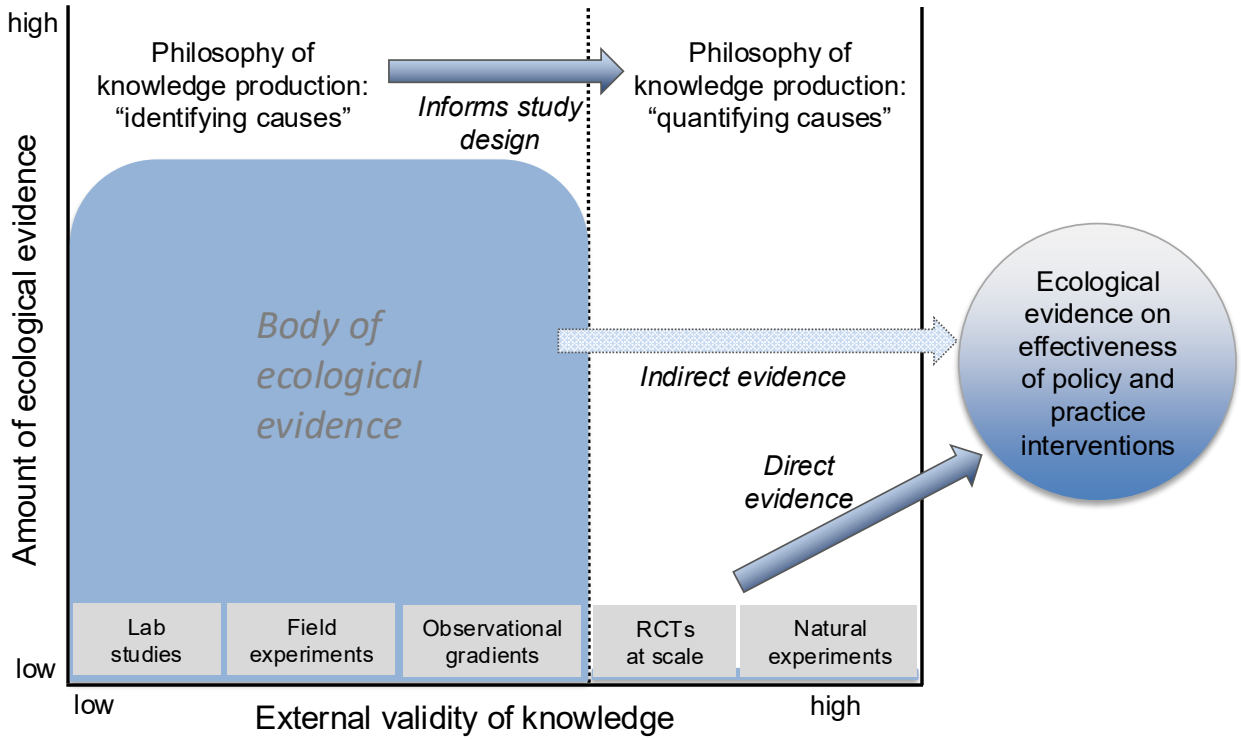
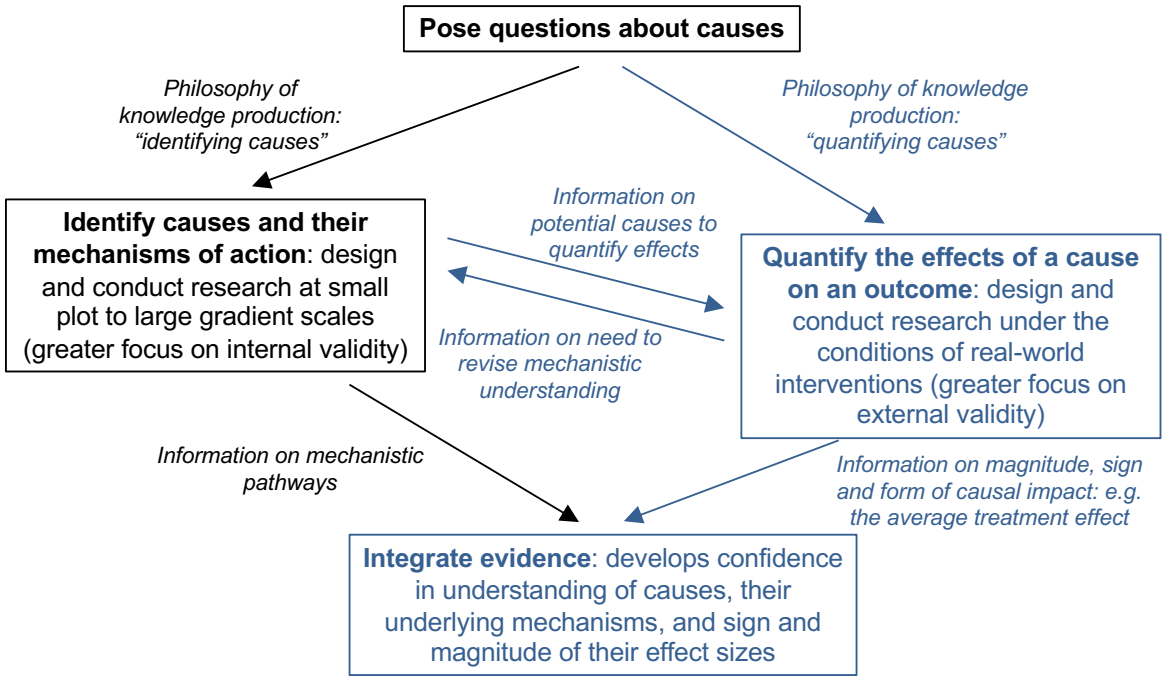


Fig. 1 Conceptualisation of knowledge imbalance in ecology generated by two, distinct philosophical approaches for developing evidence, with implications for science-based decision-making. Ecological knowledge is predominantly developed through an 'identifying causes' approach to research. Given that process-based and theoretical models (not shown) primarily make use of this evidence, they also fit within this body of knowledge (yet could incorporate knowledge from both philosophies). Applying 'identifying causes' evidence to inform policymakers and practitioners about the likely effectiveness of an intervention requires many assumptions about how mechanistic knowledge translates to intervention effectiveness (depicted as *indirect evidence*: light blue arrow). In contrast, 'quantifying causes' research is designed to robustly estimate the effectiveness of causal interventions under the real-world scales at which they are applied (*direct evidence*: dark blue arrow), giving evidence higher external validity. If ecology broadly adopts 'quantifying causes' approaches, the collective evidence base will comprise knowledge of why (i.e. mechanisms) and how (i.e. effectiveness) interventions are effective, facilitating more confident science-based decision making.



871

872 **Fig. 2 Flow diagram of two, distinct philosophical approaches for developing scientific**
873 **evidence.** Ecology is dominated by the 'identifying causes' philosophy (black text and lines),
874 which is typically used in ecology to identify causes by their ability to explain variation in
875 observed outcomes, and via controlled lab and field experiments that manipulate causal
876 variables while minimizing variation – through study design – in other causes. In contrast, the
877 'quantifying causes' approach (blue text and lines) focuses on estimating the treatment effect
878 of a cause (e.g. a policy or practice intervention) for a specified population and timescale and
879 subsamples individual units to orthogonally and representatively capture variation in other
880 influential causes on the population-level outcome. Both philosophies employ a range of
881 approaches to inference, including Bayesian and frequentist statistics, and inductive and
882 deductive reasoning. Yet the two philosophies are distinct in their questions, study designs and
883 analytical reasoning. Integration of the evidence (e.g. through process-based modelling) guides
884 further research under both philosophies (reverse arrows from 'integrate evidence' box not
885 shown). Were ecology to more broadly adopt 'quantifying causes' research – perhaps through a
886 new subfield of 'applied causal ecology' – it would dramatically expand the types of knowledge
887 generated and how they feedback upon one another (blue text boxes and arrows), thereby
888 leading to a more comprehensive basic and applied knowledge base.

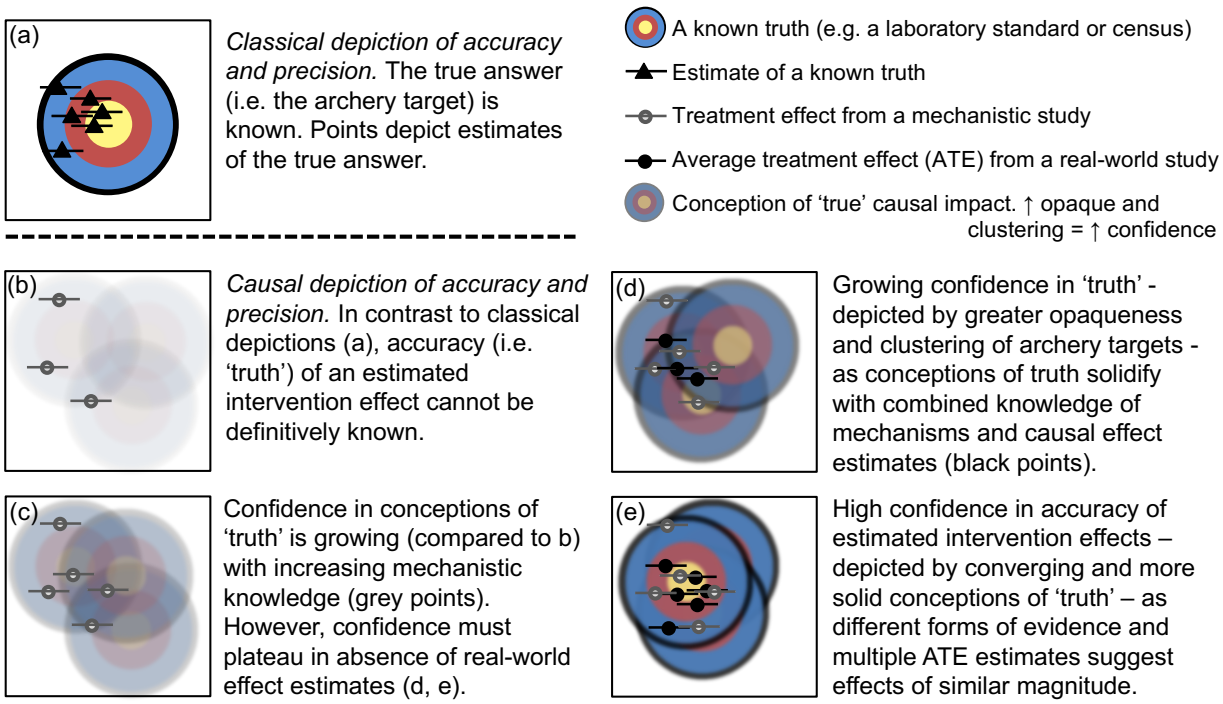


Fig. 3 Schematic of the classical (a) and causal (b-e) depiction of accuracy, where under the latter accuracy cannot be definitively known because conceptions of 'truth' emerge from inference based on evidence of mechanisms and estimates of causal effects. In the classical depiction (a), accuracy is determined against a known value (the archery target), such as a census (i.e. all individuals in the population are measured) or a laboratory standard, against which unknown values are compared. Higher accuracy is depicted by samples nearer to the centre of the target. Under causal inference, however, confidence in the accuracy of estimated intervention effectiveness emerges as mechanistic understanding (b,c) and real-world evidence of intervention effects grows (d,e). Multiple archery targets are depicted for causal conceptions (b-e) because equally plausible conceptions of 'truth' may be suggested by the available evidence. Confidence that estimated effects are likely accurate (transition from b to e) emerges as different types of knowledge, and the influence of researcher decisions[38,118–122], begin to constrain the range of plausible effect estimates. Note that confidence in the accuracy of estimated treatment effects is relatively low with only mechanistic knowledge (b,c), demanding study designs that yield externally-valid estimates (d,e) to improve confidence that estimated intervention effects will be realised under real-world conditions[123–125].

Box 1 Causal research designs to quantify the effectiveness of interventions

The Population-Intervention-Control-Outcome (PICO) framework establishes a framework for causal research designs intended to generate real-world evidence[8–10]. Such designs are common in causal fields, such as epidemiology. For example, consider a theoretical interventional study for a human population to estimate the efficacy of statins at reducing LDL (or 'bad') cholesterol. Such a study would only enrol individuals who would plausibly benefit from receiving statins as the intervention. Since adults, not children, generally experience health burdens attributed to high LDL, the target population from which study participants are selected would not include minors. Similarly, to robustly estimate the effect of statins, other factors associated with high LDL, such as exercise and diet, must be accounted for in study design and/or analysis, such as through inclusion of adult participants that span ranges in exercise and diets representative for the target population. With good baseline and subsequent follow-up data for high-LDL individuals opting to receive treatment or not, this study design will allow the investigator to identify variation due to the treatment from variation due to other causes, resulting in estimation of a plausibly accurate, sample average effect of statins[11]. Given the external validity of the design, this average treatment effect (i.e. ATE) could confidently be expected to approximate the average, population-level benefit of the intervention if it were adopted as a health intervention for the target population.

Study designs incorporating these principles for 'ecosystem management interventions' are rare in ecology (good examples include[43,99,126]) but common in causal fields[11,20,34,84]. Table 1 applies the study design principles from the statin-intervention study to an ecosystem example of agricultural land to forest conversion at a regional-scale. In contrast, current ecological studies conducted at this scale, such as the soil module of the European Union (EU) 'Land Use/Cover Area frame statistical Survey'[32,33], follow designs suited to identifying potential causes and effects versus quantifying causes (Fig. 2). Yet, the original intent of LUCAS Soil was to assess soil characteristics in relation to practices (e.g. land use) driven by the presence or absence of policy instruments[33], and hence the effectiveness of interventions that incentivize change in agricultural practices to improve soil health. For example, resulting studies to address questions about effects of practices on soil microbial biomass, a variable that is crucial to soil

health and hence food security[127–130], have focused on explaining variation across all measured fields due to multiple factors, including climate and soil texture[131,132]. Such knowledge is important for understanding mechanism but does not directly guide interventions because climate and soil texture are not causes that can be directly affected by agricultural policies focused on improving soil health. Furthermore, the inclusion of all fields irrespective of intended policy interventions means these data are not directly applicable to the population of interest. For example, for a program promoting reforestation to sequester soil carbon, high-yield fields may be ineligible given the need to produce food and prevent 'carbon leakage' through compensatory conversion of other lands to arable agriculture. Study designs tailored to quantify intervention effectiveness then require a focus on the subpopulation of fields that are eligible for the intervention, sampled to account for but not necessarily quantify the effects of non-target causal variables[40].

Ecology is largely depauperate in such causal intervention research designs (see Fig. 1) but by translating approaches from causal fields (Table 1), would be well positioned to address many questions that have high value for directly informing policies and practices to achieve production, biodiversity and climate goals.

Table 1. General components and related principles for studies designed to estimate the quantitative impacts of 'ecosystem management interventions' (see Box 1). The collective intent of employing these principles is to generate external validity in the estimated average, population-level benefit (and potential downsides) of an intervention. The principles are illustrated assuming their application to the Box 1 example of agricultural land to forest conversion impacts on soil carbon sequestration. Application of the principles centres study design and knowledge of the system as critical to making valid causal inferences. Such inferences provide robust quantitative evidence, that policymakers and practitioners can have high confidence in, for making decisions about the efficacy of interventions to protect and restore nature.

Component	Principle	Example of application
Individual/ Population/ Outcomes	- An 'individual' (or group – see main text) is the unit that receives the intervention	- For reforestation of arable land, an agricultural field might be considered the individual given that it is the unit of management.
	- The population is made up of the individuals, under a defined context, that are eligible to receive the intervention	- The population should consist only of treatment and control fields that are eligible to be reforested. Consideration of fields as controls that are ineligible for reforestation, perhaps because they have high fertility and relatedly crop yields, could confound understanding of intervention efficacy.
	- Measure the individual responses (the outcome) in a representative manner	- Agricultural soil studies often measure only a single small area within a field. Such designs are inefficient when contrasted with taking multiple samples across each field, and make estimated effects of interventions unreliable[133–136]. Similarly, the measured outcome variable for the individual should be a valid construct for the desired response. For example, soil carbon concentrations are not suitable proxies for assessing changes in soil carbon stocks. In addition, investigators should be aware of 'spillover effects' where, for example, trees grown immediately adjacent to a control field will influence crop yields in that field[137] and hence baseline carbon removals.
Causes (Interventions)	- Focus on quantifying a specific cause (i.e. the intervention) applied as in real-world settings	- The focus should be on quantification of the effects of reforestation, and not on 'causes' that cannot – at least in principle – be manipulated but which might influence the outcome (e.g. soil texture). Such a focus shifts study goals from explaining variation in the outcome, to quantification of the causal effect of interest[138]. Further, to inform policy and practice, the focus should be on evaluating the reforestation interventions under the real-world conditions under which they are applied.
	- The process of assignment to intervention treatment should satisfy – or	- Fields in the study to be reforested should have an approximately equal likelihood of being assigned to the 'control' group[21]. Such considerations help to avoid

	approximate – the exchangeability criterion	confounding, for example, where a farmer might choose to reforest their least fertile fields, raising questions about the suitability of unconverted fields to function as valid comparators.
	- Choose a suitable control/ comparison against which to estimate the intervention effect	- A causal effect is always estimated relative to a suitable comparison group. A suitable control or comparison group may, for example, allow consideration of dynamic baselines that serve as a plausible counterfactual scenario to estimate what would have happened had reforested fields not received that treatment.
External validity	- Select the sample of individuals to measure from the target population in a representative manner	- Population-level variation in non-target causal variables which influence the outcome (e.g. soil texture on soil carbon) should ideally be captured across the sampled individuals, and be orthogonal to the target cause, in this instance reforestation. For example, older farms might have more fertile soil (reflecting the original choice of location to establish a field), so treatment and control fields might be stratified by farm age. Capturing such variation through study design makes the sample treatment effect more representative of the likely population treatment effect.
	- The time and spatial scale should match with policy and practice needs	- Ecological evidence is often needed on accelerated timescales to inform decision makers. However, the impacts of those decisions, such as to reforest, are typically longer term and realised at broader spatial scales. Such realities require study designs that also measure impact once interventions are applied at scale to evaluate how nearer-term effects are realised at the real-world timescales of intervention impact.
	- The treatment effects should be robust to repetition and ideally transferable	- The estimated treatment effect should be approximately equal when different reforested and control fields are sampled from the same population. Such robustness to study design and analysis decisions builds confidence that the efficacy of the intervention will be realised when applied to the population (i.e. the estimate is generalisable). Testing the impact of the intervention under new contexts (e.g. in one region of France versus another physiographic province) and being able to explain how and why effect sizes compare (or don't) for the same intervention, builds confidence that the intervention will be applied only to populations where it is effective (i.e. transferability).

975