Received: 7 February 2022 | Revised: 21 March 2022 | Accepted: 8 May 2022

DOI: 10.1111/ele.14033

ECOLOGY LETTERS cnrs WILEY

VIEWPOINT

Predictive models aren't for causal inference

Suchinta Arif | **M. Aaron MacNeil**

Ocean Frontier Institute, Dalhousie University, Department of Biology, Halifax, Nova Scotia, Canada

Correspondence
Suchinta Arif, Ocean Frontier Institute, Dalhousie University, Department of Biology, Life Sciences Building, 1355 Oxford St., B3H 3Z1, Halifax, Nova Scotia, Canada.
Email: suchinta.arif@dal.ca

Editor: Marti Anderson

Abstract

Ecologists often rely on observational data to understand causal relationships. Although observational causal inference methodologies exist, predictive techniques such as model selection based on information criterion (e.g. AIC) remains a common approach used to understand ecological relationships. However, predictive approaches are not appropriate for drawing causal conclusions. Here, we highlight the distinction between predictive and causal inference and show how predictive techniques can lead to biased causal estimates. Instead, we encourage ecologists to valid causal inference methods such as the backdoor criterion, a graphical rule that can be used to determine causal relationships across observational studies.

KEYWORDS
back-door criterion, causal inference, directed acyclic graphs (DAGs), model selection, prediction

As ecologists, we are often interested in answering causal questions, such as the effect of climate-induced bleaching events on coral reef ecosystems (e.g. Graham et al., 2015), the impact of deforestation on biodiversity (e.g. Brook et al., 2003), or the effect of conservation and management responses on restoring ecosystem services (e.g. Sala & Giakoumi, 2018). Often, randomised controlled experiments are unfeasible, and ecologists instead rely on observational data to answer fundamental causal questions in ecology (MacNeil, 2008). New advances in technology such as remote-sensing and animal-borne sensors, as well as increased availability of citizen science and electronic data have further increased opportunities to answer causal questions from observational data (Sagarin & Pauchard, 2010).

In recent years, some ecologists have advocated for the increased application of developed causal inference methodologies for answering cause and effect relationships from observational data. For example Cronin and Schoolmaster Jr. (2018), Schoolmaster Jr. et al. (2020), Laubach et al. (2021) and Arif and MacNeil (2022) utilise causal models based on the Structural Causal Model framework (SCM; Pearl, 2009); Butsic et al. (2017), Larsen et al. (2019) and Arif and MacNeil (2022) discuss quasi-experimental approaches for observational causal inference; and Ferraro et al. (2019) and Kimmel et al. (2021) discuss causal assumptions required for both observational and experimental set-ups. However, causal inference methodologies have yet to be widely adopted in observational ecology. Instead, drawing causal conclusions from observational data is typically taboo, with Pearson's oft-cited 'correlation doesn't equal causation' used to block attempts to do so (Glymour, 2009). This misconception—that causality cannot be inferred using observational data—has resulted in a culture where ecologists dependent on observational data for understanding causal relationships avoid explicitly acknowledging the causal goal of research projects and instead use coded language that implies causality without explicitly saying so (Arif & MacNeil, 2022; Hernan, 2018).

A common strategy used to quantify ecological relationships is to apply model selection, using information metrics such as Akaike's information criterion (AIC; Akaike, 1973). These approaches select the 'best' model among a candidate set and subsequently make inferences from parameters that are of ecological interest within the top-ranked model. Often, these inferences are tied up with causal language, implying that having selected

[Corrections added on 21 July 2022, after first online publication: In Figure 1, an arrow has been added between "Forestry" and "Invasive Species Z"; the reference to backdoor paths has been corrected on page 3; the author name "Aaron MacNeil" has been corrected to "M. Aaron MacNeil" in this version.]

TABLE 1 A sample of recent observational ecological studies that have used model selection techniques to answer causal questions, using causal language (e.g. effect, driver, influence) to communicate results

Paper	Causal question	Model selection	Causal language
Millard et al., 2021	What are the global effects of land-use intensity on local pollination biodiversity?	AIC	Yes
Lu et al., 2021	What is the impact of land surface temperature on urban net primary productivity?	AIC	Yes
Safaie et al. (2018)	How does high-frequency temperature variability effect the risk of coral bleaching?	AICc	Yes
Morton et al., 2021	What is the impact of wildlife trade on terrestrial biodiversity?	BIC	Yes
Chinn et al., 2021	What is the influence of intrinsic and extrinsic attributes on neonate survival in wild pigs?	WAIC	Yes
Montano-Centellas et al. (2020)	What are the ecological drivers of avian community assembly along a tropical elevation gradient?	AIC	Yes
Rode et al., 2021	What are the combined effects of sea ice, seal body condition and atmospheric circulation patterns on polar bears in the Shukchi Sea?	AIC	Yes
Sinnot-Armstrong et al. (2021)	What are the biotic and abiotic drivers of fruit colour syndrome?	AICc	Yes
Walker et al., 2021	What factors influence scavenger guilds and scavenging efficacy in Southwestern Montana?	AICc	Yes
Teixeira et al., 2021	How did past environmental changes (prior to human impact) effect lemur population dynamics in Madagascar?	AIC	Yes

the best model, one can proceed to using causal language in reference to it (Table 1). However, model selection is not a valid method for inferring causal relationships—rather, these techniques aim to select the best model for predicting a response variable of interest. For example AIC approximates a model's out of sample predictive accuracy, using only within-sample data (Akaike, 1973). Although numerous model selection criteria exist (e.g. BIC, Schwarz, 1978; DIC; Spiegelhalter et al., 2002; WAIC, Watanabe, 2013; LOO-CV; Vehtari et al., 2017), they are all used to compare models based on predictive accuracy (Laubach et al., 2021; McElreath, 2020; Tredennick et al., 2021). Thus, model selection is appropriate for predictive inference (i.e. which model best predicts Y?), which is fundamentally distinct from causal inference (i.e. what is the effect of X on Y?).

To demonstrate this distinction, the directed acyclic graph (DAG) in Figure 1 shows the causal structure of a hypothetical ecological system. DAGs can be used to visualise causal relationships, where variables (nodes) are connected to each other via directed arrows, pointing from cause to effect (Elwert, 2013). For example forestry affects species Y both directly (there is a directed arrow between them) and indirectly, via the directed arrow from forestry to species A and from species A to species Y (Figure 1). We created a simulated dataset that matches the linear causal structure of this DAG, setting the total (i.e. direct and indirect) causal effect of forestry on species Y to −0.75 (Appendix S1). We further specified candidate linear regression models that included

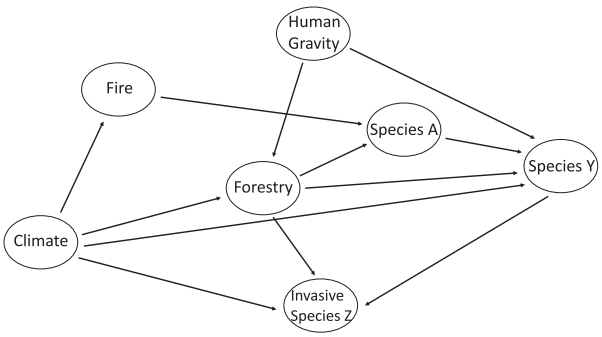

FIGURE 1 A directed acyclic graph (DAG) representing the causal structure of a hypothetical ecological system.

all possible covariate combinations where species Y is a response. Using our simulated data and our candidate models, both AIC and BIC selected a 'best' model where forestry, species A, human gravity, climate and invasive species Z were included as covariates (Appendix S1). However, interpreting the coefficients of this model provided biased causal estimates, with the effect of forestry on species richness estimated to be −0.36 [−0.38, −0.33], instead of −0.75 (Appendix S1).

In this scenario, there are two statistical biases at play (see Appendix S1 for a breakdown of common statistical biases). The first is overcontrol bias, which occurs when the inclusion of an intermediate variable along a causal pathway removes the indirect causal effect between

predictor and response (Cinelli et al., 2022; Appendix S1). Here, the inclusion of the intermediate variable species A removes the indirect effect between forestry and species Y. Second, the inclusion of invasive species Z as a covariate leads to collider bias, which can result from adjusting for a variable that is caused by both predictor and response (Cinelli et al., 2022; Appendix S1). Here, the inclusion of invasive species Z induces an additional, but non-causal, association between forestry and species Y. Past studies have noted that other forms of statistical biases (i.e. non-causal associations) such as collider bias (Appendix S1) can also increase predictive accuracy (e.g. Griffith et al., 2020). Thus, a model selected based on predictive accuracy should not be assumed to be causally accurate and ecologists need to understand these dimensions to ensure the tools used are fit for purpose.

Despite the distinction between predictive and causal inference, many ecological studies continue to misuse predictive techniques for causal interpretation. In addition to model selection, machine learning (ML) techniques meant for predicting outcomes have also been misused for causal interpretation. ML is a rapidly growing approach concerned with generating accurate predictive models from often large and complex datasets (Greener et al., 2022). However, despite its relatively recent emergence, predictive ML techniques have already been conflated with causality in the ecological literature. For example predictive machine learning techniques have been used to understand the drivers of extinction risk in marine mammals (Davidson et al., 2012), geographical drivers of the Mediterranean fruit fly (Bekker et al., 2019), drivers of viral density in bats (Guy et al., 2020) and mechanisms driving arbovirus outbreaks (Alkhamis et al., 2021). Ecological reviews of predictive ML approaches have further suggested that such techniques can be used to both predict and causally explain ecological processes (Olden et al., 2008; Yu et al., 2021). The increased application of predictive ML techniques combined with the growing availability of big data may lead to their further misuse for causal interpretations in ecology.

Ultimately, the widespread use of predictive techniques for causal inference across ecological studies signal that developed observational causal inference methodologies have not been properly introduced to many ecologists. To help remedy this, we outline a widely applicable graphical rule, the backdoor criterion, that can be applied to determine causal effects from observational data. In contrast to predictive techniques, causal models based the backdoor criterion are built based on the specific causal question at hand, as well as through the careful consideration of the overall causal structure of a system, including how different predictor variables may be related to one another. It provides a formal means for isolating causal effects from observational data, and eliminating common statistical biases including confounding, overcontrol and collider bias (Appendix S1).

THE BACKDOOR CRITERION: COVARIATE SELECTION FOR CAUSAL INFERENCE

A causal inference methodology that has recently emerged in ecology is Judea Pearl's Structural Causal Model framework (SCM; Pearl, 2009). This framework uses DAGs to visualise researchers' assumptions about the causal structure of a system or process under study. DAGs should include all relevant variables required to depict a system under study and be carefully constructed and sufficiently justified based on domain ecological knowledge. We refer readers to Cronin and Schoolmaster Jr. (2018), Schoolmaster Jr. et al. (2020) and Arif and MacNeil (2022) to gain a deeper understanding of how to construct DAGs to represent a system or process under study. Once a DAG has been created, the backdoor criterion can be applied to determine the covariates required to answer a specified causal question from observational data.

Conceptually, the backdoor criterion instructs us to block all non-causal paths between a predictor and response variable of interest, while leaving all causal pathways open. Graphically, this translates to blocking all backdoor paths between a predictor and response variable. Backdoor paths are sequences of nodes and arrows with an arrow pointing into both the predictor and response variable of interest; if left open, they can induce non-causal associations between variables of interest. To block a backdoor path, we can either (1) adjust for an intermediate arrow-emitting variable or (2) not adjust for a variable with two incoming arrows (i.e. a collider variable: ➔ X ←).

For example given our DAG in Figure 1, to determine the total effect of forestry on species Y, there are three backdoor paths that must be blocked:

1. Species Y ← Climate ➔ Forestry
2. Species Y ← Species A ← Fire ← Climate ➔ Forestry
3. Species Y ← Human Gravity ➔ Forestry

The first two backdoor paths can each be blocked by adjusting for the intermediate arrow-emitting variable climate. The third backdoor path can be blocked by adjusting for the intermediate arrow-emitting variable human gravity. Therefore, to determine the total effect of forestry on species Y, we must adjust for climate and human gravity. Following covariate selection, researchers can determine the appropriate statistical analysis, given their data. It is important to note that DAGs and the backdoor criterion are compatible with both linear and non-parametric approaches (Elwert, 2013; Pearl, 2009). As our simulated data were created using linear relationships, we have chosen a linear regression model, setting species Y as our response, forestry as our predictor and including climate and human gravity as controls. This model returned an accurate total causal estimate of −0.75[−0.77, −0.73] (Appendix S1). The application of the backdoor criterion

can become increasingly complex with larger DAGs and as such, tools such as 'dagitty' (www.dagitty.net; instructions within site) can help in composing DAGs and specifying causal questions, which will subsequently identify required backdoor adjustment sets. We further recommend that ecologists create their DAG and determine potential backdoor adjustment sets before collecting observational data, to ensure that variables needed for causal analysis are measured.

DISTINCTION FROM PREDICTIVE INFERENCE

Covariate selection using the backdoor criterion is fundamentally distinct from commonly applied predictive techniques. The backdoor criterion is based on counterfactual reasoning, equating observational distributions to what would be expected under a randomised control experiment (Pearl, 2009). Unlike predictive approaches, the backdoor criterion was specifically created to answer cause and effect relationships from observational data. Furthermore, whereas predictive approaches often rely on the data to determine the best model, the backdoor criterion uses domain knowledge, above all else, to determine the best causal model for a given causal query. The use of DAGs and the subsequent application of the backdoor criterion allows ecologists to move away from an automated model selection and other predictive techniques to one that empowers ecologists to think critically about the cause-and-effect relationships in their study system. The use of DAGs also facilitates open critique of causal assumptions therefore their causal conclusions, which in turn can lead to a productive scientific debate that deepens our understanding of ecological phenomena (e.g. see Schoolmaster Jr. et al., 2020; rebuttal by Grace et al., 2021; and reply by Schoolmaster Jr. et al., 2021).

Currently, DAGs and the backdoor criterion are underutilised relative to the dominant predictive model selection techniques for understanding causal relationships in ecology. Thus far, the backdoor criterion has been applied to understand the causes of species level trait covariation (Cronin & Schoolmaster Jr., 2018), biodiversity-ecosystem function correlations (Schoolmaster Jr. et al., 2020), and causal drivers of coral-algal regime shifts (Arif et al., 2021). As these varied examples demonstrate, the backdoor criterion can be widely applicable for understanding ecological causal relationships.

While we have highlighted the backdoor criterion as a widely applicable observational causal inference tool, we note that numerous other tools and frameworks exist. Readers are also encouraged to learn about the front-door criterion, a second graphical rule under the SCM framework that can be used to determine causal effects in the presence of an unobserved confounding variable (Paul, 2011; Pearl, 1995). Other causal inference frameworks, such as quasi-experimental approaches (Arif & MacNeil, 2022; Butsic et al., 2017; Larsen et al., 2019), and time-series causal analysis (Runge et al., 2019) may also be useful, given the causal questions and data at hand. Ultimately, the uptake of valid causal inference methods across observational ecological research will lead to better statistical analysis and causal understanding of ecological phenomena.

AUTHORSHIP

SA conceived the idea for this viewpoint and led the drafting of the manuscript. All authors edited the manuscript and collaboratively discussed their perspectives throughout writing the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ele.14033.

DATA AVAILABILITY STATEMENT

No new data were used in this manuscript (https://doi.org/10.6084/m9.figshare.19775242).

ORCID

Suchinta Arif https://orcid.org/0000-0001-8381-3071

REFERENCES

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N. & Csaki, B.F. (Eds.) *Second international symposium on information theory*. Budapest: Academiai Kiado, pp. 267–281.

Alkhamis, M., Fountain-Jones, N., Aguilar-Vega, C. & Sanchez-Vizcaino, J. (2021) Environment, vector, or host? Using machine learning to untangle the mechanisms driving arbovirus outbreaks. *Ecological Applications*, 31(7), e02407.

Arif, S., Graham, N., Wilson, S. & MacNeil, A. (2021) Causal drivers of climate-mediated coral reef regime shifts. *Ecosphere*, 13(3), e3956.

Arif, S. & MacNeil, A. (2022) Utilizing causal diagrams across quasi-experimental approaches. *Ecosphere*, 13(4), e4009.

Bekker, G.F.H.V.G., Addision, M., Addison, M., Addison, P. & Niekerk, A. (2019) Using machine learning to identify the geographical drivers of Ceratitis capitata trap catch in an agricultural landscape. *Computers and Electronics in Agriculture*, 162, 582–592.

Brook, B., Sodhi, N. & Ng, P. (2003) Catastrophic extinctions follow deforestation in Singapore. *Nature*, 424, 420–423.

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

Chinn, S.M., Kilgo, J.C., Vukovich, M.A. & Beasley, J.C. (2021) Influence of intrinsic and extrinsic attributes on neonate survival in an invasive large mammal. *Scientific Reports*, 11, 11033.

Cinelli, C., Forney, A. & Pearl, J. (2022) Crash course in good and bad controls. *Sociological Methods and Research*. https://doi.org/10.1177/00491241221099552

Cronin, J. & Schoolmaster, D., Jr. (2018) A causal partition of trait correlations: using graphical models to derive statistical models from theoretical language. *Ecosphere*, 9(9), e02422.

Davidson, A., Boyer, A., Kim, H., Pompa-Mansilla, S., Hamilton, M., Costa, D. et al. (2012) Drivers and hotspots of extinction risk in marine mammals. *PNAS*, 109(9), 3395–3400.

Elwert, F. (2013) Graphical causal models. In: Morgan, S.L. (Ed.) *Handbook of causal analysis for social research*. Dordrecht: Springer, pp. 245–273.
Ferraro, P.J., Sanchirico, J.N. & Smith, M.D. (2019) Causal inference in coupled human and natural systems. *PNAS*, 116(12), 5311–5318.
Glymour, C. (2009) Causation and statistical inference. In: Beebee, H., Hitchcock, C. & Menzies, P. (Eds.) *The Oxford handbook of causation*. New York, USA: Oxford University Press.
Grace, J.B., Loreau, M. & Schmid, B. (2021) A graphical causal model for resolving species identity effects and biodiversity–ecosystem function correlation: comment. *Ecology*, 0, e03378.
Graham, N., Jennings, S., MacNeil, A., Mouillot, D. & Wilson, S. (2015) Predicting climate-driven regime shifts versus rebound potential in coral reefs. *Nature*, 518, 94–97.
Greener, J., Kandathil, S., Moffat, L. & Jones, D. (2022) A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23, 40–55.
Griffith, G., Morris, T., Tudball, M., Herbert, A., Mancano, G., Pike, L. et al. (2020) Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications*, 11, 5749.
Guy, C., Ratcliffe, J. & Mideo, N. (2020) The influence of bat ecology on viral diversity and reservoir status. *Ecology and Evolution*, 10(12), 5748–5758.
Hernan, M. (2018) The C-word: scientific euphemisms do not improve causal inference from observational data. *AJPH*, 108(5), 616–619.
Kimmel, K., Dee, L.E., Avolio, M.L. & Ferraro, P.J. (2021) Causal assumptions and causal inference in ecological experiments. *Trends in Ecology & Evolution*, 36(12), 1141–1152.
Larsen, A., Meng, K. & Kendall, B. (2019) Causal analysis in control-impact ecological studies with observational data. *Methods in Ecology and Evolution*, 10, 924–934.
Laubach, Z., Murray, E., Hoke, K., Safran, R. & Perng, W. (2021) A biologist's guide to model selection and causal inference. *Proceedings of the Royal Society B: Biological Sciences*, 288(1943), 20202815.
Lu, X., Chen, X., Zhao, X., Lv, D. & Zhang, Y. (2021) Assessing the impact of land surface temperature on urban net primary productivity increment based on geographically weighted regression model. *Scientific Reports*, 11, 22282.
MacNeil, A. (2008) Making empirical progress in observational ecology. *Environmental Conservation*, 35(3), 193–196.
McElreath, R. (2020) *Statistical rethinking: a Bayesian course with examples in R and Stan*. Boca Raton, USA: CRC Press.
Millard, J., Outhwaite, C.L., Kinnersley, R., Freeman, R., Gregory, R., Adedoja, O. et al. (2021) Global effects of land-use intensity on local pollinator biodiversity. *Nature Communications*, 12, 2902.
Montano-Centellas, F., Loiselle, B. & Tingley, M. (2020) Ecological drivers of avian community assembly along a tropical elevation gradient. *Ecography*, 44(4), 574–588.
Morton, O., Scheffers, B.R., Haugaasen, T. & Edwards, D. (2021) Impacts of wildlife trade on terrestrial biodiversity. *Nature Ecology & Evolution*, 5, 540–548.
Olden, J., Lawler, J. & Poff, N. (2008) Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology*, 83(2), 171–193.
Paul, W.L. (2011) A causal modelling approach to spatial and temporal confounding in environmental impact studies. *Envirometrics*, 22, 626–638.
Pearl, J. (1995) Causal diagrams for empirical research. *Biometrika*, 82(4), 669–710.
Pearl, J. (2009) *Causality: models, reasoning and inference*, 2nd edition. UK: Cambridge University Press; Cambridge.
Rode, K., Regehr, E., Bromaghin, J., Wilson, R., Martin, M., Crawford, J. et al. (2021) Seal body condition and atmospheric circulation patterns influence polar bear body condition, recruitment, and feeding ecology in the Chukchi Sea. *Global Change Biology*, 27(12), 2684–2701.
Runge, J., Bathiany, S., Bolt, E., Camps-Valls, G., Coumou, D., Deyle, E. et al. (2019) Inferring causation from time series in earth system sciences. *Nature Communications*, 10, 2553.
Safaie, A., Silbiger, N.J., McClanahan, T.R., Pawlak, G., Barshis, D.K., Hench, J.L. et al. (2018) High frequency temperature variability reduces the risk of coral bleaching. *Nature Communications*, 9, 1671.
Sagarin, R. & Pauchard, A. (2010) Observational approaches in ecology open new ground in a changing world. *Frontiers in Ecology and the Environment*, 8(7), 379–386.
Sala, E. & Giakoumi, S. (2018) No-take marine reserves are the most effective protected areas in the ocean. *ICES Journal of Marine Science*, 75(3), 1166–1168.
Schoolmaster, D., Jr., Zirbel, C. & Cronin, P. (2020) A graphical causal model for resolving species identity effects and biodiversity–ecosystem function correlations. *Ecology*, 101(8), e03070.
Schoolmaster, D., Jr., Zirbel, C. & Cronin, P. (2021) A graphical causal model for resolving species identity effects and biodiversity-ecosystem function correlations: reply. *Ecology*, 0, e03593.
Schwarz, G.E. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
Sinnot-Armstrong, M., Donoghue, M. & Jetz, W. (2021) Dispersers and environment drive global variation in fruit colour syndromes. *Ecology Letters*, 24(7), 1387–1399.
Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, 64(4), 583–639.
Teixeira, H., Montade, V., Salmona, J., Metzger, J., Bremond, L. & Kasper, T. (2021) Past environmental changes affected lemur population dynamics prior to human impact in Madagascar. *Communications Biology*, 4, 1084.
Tredennick, A., Hooker, G., Ellner, S. & Adler, P. (2021) A practical guide to selecting models for exploration, inference and prediction in ecology. *Ecology*, 102(6), e03336.
Vehtari, A., Gelman, A. & Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.
Walker, M., Uribasterra, M., Asher, V., Getz, W., Ryan, S., Ponciano, J. et al. (2021) Factors influencing scavenger guilds and scavenging efficiency in southwestern Montana. *Scientific Reports*, 11, 4254.
Watanabe, S. (2013) A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867–897.
Yu, Q., Ji, W., Prihodko, L., Ross, C., Anchang, J. & Hanan, N. (2021) Study becomes insight: ecological learning from machine learning. *Methods in Ecology and Evolution*, 12(11), 2117–2128.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Arif, S. & MacNeil, M.A. (2022) Predictive models aren't for causal inference. *Ecology Letters*, 25, 1741–1745. Available from: https://doi.org/10.1111/ele.14033