# Falsifying causal hypotheses in time series models with conditional-independence tests

Running header: falsifying time-series causal models

James T. Thorson[1*], Jennifer S. Bigman[2], Cole C. Monnahan[1], Lauren A. Rogers[3]

[1] Resource Ecology and Fisheries Management, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA

[2] Office of Science and Technology, NMFS, Seattle, WA, USA

[3] Recruitment Process Program, Alaska Fisheries Science Center, NOAA, NMFS, Seattle, Seattle, WA, USA

* Corresponding author: James.Thorson@noaa.gov

Keywords: structural causal model, time-series model, directional separation, d-sep, conditional independence, dynamic structural equation model

**Data availability:**

Data for the pollock spawning phenology case study are from Rogers et al. (2025), available online at https://github.com/larogers123/spawn_timing_catchability. Data for the Isle Royale are

from https://www.isleroyalewolf.org/, and we use the copy available in package *dsem*. Code to reproduce case studies and the simulation experiment are available as an anonymized GitHub (https://anonymous.4open.science/r/dsep_in_dsem-61FF/plot_histogram.R) and will be available as a public GitHub repo with Zenodo for DOI upon acceptance. The d-separation test is available in *dsem* as function *test_dsep*.

**Acknowledgements**

We thank J. Sullivan, B. Chasco, and two anonymous reviewers for helpful comments on an earlier draft. We also thank G. Marchetti (the developer of R-package *ggm*) for allowing us to copy necessary functions under open-source license, and W. van der Bijl for inspiration regarding how to automate the d-sep test.

**Conflict of Interest**

None to declare

**Author contributions**

J. Thorson derived the statistical method, developed code, conducted the simulation and case study analyses, and lead writing. C. Monnahan reviewed causal papers in ecology, wrote portions of the Introduction and Discussion, and reviewed code. J. Bigman and L. Rogers wrote portions of the Discussion, and L. Rogers curated data for the pollock case study. All authors revised the manuscript.

**Abstract:**

1. Ecologists often use time-series models to approximate dynamics arising from density dependence, species interactions, community synchrony, and other processes. Dynamic structural equation models (DSEM) can represent simultaneous and lagged interactions among variables with missing data, and therefore encompasses a wide family of analyses (linear regression, vector autoregressive models, and dynamic factor analysis). However, before interpreting a DSEM as a causal model, analysts should first test whether its assumptions about conditional independence are inconsistent with available data (i.e., attempt to falsify the model).
2. In site-replicated and phylogenetic contexts, ecologists seek to falsify causal assumptions by testing implied conditional-independence relationships using a directional-separation ("d-sep") test, but this has not been demonstrated using time-series analysis of ecological systems involving simultaneous and lagged interactions. Here, we propose a time-series d-sep test and use a simulation experiment and case studies to explore its performance.
3. The simulation confirms that this test results in a uniform p-value when using a correct causal model, and a low p-value (i.e., a decision to reject a model) when the causal model is incorrect. As expected, time-series that are short or have a large proportion of missing data have less power to reject an incorrect model. In a previously published analysis involving wolf-moose interactions in Isle Royale, the test supports top-down control but cannot distinguish whether bottom-up control is supported. In a novel application involving pollock in the Gulf of Alaska, the test supports a conceptual model where temperature drives spawning phenology, which subsequently affects availability to a spawning survey.

4. We conclude that d-sep is a useful test to falsify the conditional-independence assumptions of a time-series model. It is therefore complementary to other methods used to validate causal inference (i.e., controlled experiments, ecological theory, and system knowledge).

## Introduction

Ecologists study causality in natural systems using controlled experiments and the analysis of observational data (Grace, 2024; Siegel & Dee, 2025). Developing a well-formed hypothesis is a key first step, and causal analysis has been proposed as a useful scientific framework to confront hypotheses with data (Grace & Irvine, 2020). Generating hypotheses is an iterative process of building graphical causal networks (directed acyclical graphs; DAGs) of key variables in a system independent of the data and prior to modeling, and this requires eliciting and representing expert knowledge about ecological mechanisms (e.g., see Table 5 of Grace & Irvine, 2020). Structural causal models (SCM) can then be used to estimate causal relationships by fitting statistical models to graphical models (Pearl, 2009). This approach resolves well-known issues with bias when making causal statements from predictive statistical models (Arif & MacNeil, 2022a), where the SCM forces explicit consideration of confounding factors (Byrnes & Dee, 2025). SCMs are widely used outside of ecology, and controlled experiments can be interpreted as a variant of SCM where some variables (i.e., experimental treatments) are known *a priori* to be independent of other variables. However, ecologists also use observational data for systems that are not amenable to experimental manipulation, and these settings require validating causal hypotheses to ensure unbiased causal estimates (Arif & MacNeil, 2022b; Siegel & Dee, 2025). Thus, it is vital for analysts to be able to validate their causal models fitted to observational time-series data to advance understanding of ecological mechanisms.

Time-series dynamics pose particular challenges, because interactions among variables may be either simultaneous (e.g., occurring much faster than the time-step in available observations) or lagged (e.g., where a variable in one observed time-interval affects another variable at a later time). Lagged interactions result in temporal dependence, which violates a key statistical

assumption of the popular structural equation model (SEM; Pearl, 2012) framework for estimating causal relationships and limits the practical application of SEM for time-series analysis. Thorson et al. (2024) extended the SEM modeling framework to allow for correlated observations including linear interactions among variables that include simultaneous and lagged effects. This dynamic structural equation model (DSEM) framework is efficiently represented as a Gaussian Markov random field (GMRF) and fitted as a generalized linear mixed model (GLMM), as implemented in the 'dsem' package (Thorson et al., 2024) in the R statistical environment (R Core Team, 2023). DSEM is computationally efficient, can account for missing data, and encompasses a wide range of statistical analyses including linear models, errors-in-variables, ARIMA models, dynamic factor analysis, structural vector autoregressive models, and linear SCMs. However, it is not clear how an analyst could seek to determine whether a hypothesized DSEM is consistent with available data, and potentially falsify models that are not.

In general, the best way to validate causal assumptions is by using controlled experiments to confirm that variables are independent conditional upon fixed conditions. However, experiments often cannot be run at the scale of a system (due to logistical or legal constraints). In these cases, analysts might seek to determine whether hypothesized dynamics are inconsistent with available data (i.e., falsify one or more hypotheses). For example, consider a trophic cascade, where we might specify a DSEM in which predator $X$ has an approximately linear effect on consumer $Y$ and consumer $Y$ has a linear effect on producer $Z$. We write this as two causal paths: $X \rightarrow Y$ and $Y \rightarrow Z$. In this DSEM, variation in predators is assumed to be independent of producers, conditional upon a fixed value for consumers (i.e., $X \perp Z|Y$). We can therefore test this conditional independence relationship as a regression ($Z = \beta_X X + \beta_Y Y + \epsilon$), and if the slope $\beta_X$ significantly departs from zero, then we can "reject" this component of DSEM as invalid. This

111 insight is formalized by the Shipley directional-separation ("d-sep") test (Shipley, 2000), where
112 all conditional-independence (CI) relationships implied by a given DSEM are sequentially tested
113 and results are then combined in a single "omnibus" test. This Shipley d-sep test is widely used
114 in the ecological analysis of controlled experiments (Meziane & Shipley, 2001) and phylogenetic
115 comparative analysis (von Hardenberg & Gonzalez-Voyer, 2013), and has been extended to
116 multi-level models (Shipley, 2009). However, we are not aware of studies using the Shipley d-
117 sep test to falsify causal assumptions when analyzing time-series in ecology.
118 We therefore demonstrate using an extension of the Shipley d-sep test for ecological time-
119 series. We first summarize the d-sep test for structural equation models, and then discuss
120 modifications that are necessary for application to time-series models that include simultaneous
121 and lagged effects or when dealing with missing data. We then provide a simulation experiment
122 to determine whether the proposed test has good statistical performance (i.e., results in a uniform
123 distribution for p-values) when the model is correctly specified, and also how often it can reject
124 an incorrectly specified model given a mis-specified causal structure, varied time-series lengths,
125 and varied proportions of missing data. Finally, we use two real-world case studies to illustrate
126 the types of ecological inference that can be drawn from the time-series d-sep test. Results
127 suggest that the method performs well for simple (2-4 variable) models incorporating
128 simultaneous and lagged effects given the range of time-series that are common in population
129 dynamics (25-100 time points), and the method is freely available as function `test_dsep(.)` in
130 the R package *dsem* for future use.

131 **Methods**

132 The Shipley (or d-sep) test can be applied to a directed acyclic graph (DAG) representing a
133 structural causal model. It proceeds by:

1. identifying the set of conditional independence (a.k.a. directional separation or "d-sep") relationships that are implied by the DAG. This set depends upon an *a priori* ordering of variables, but the number of relationships is invariant to ordering. To identify this set, the algorithm identifies whether every pair of variables is directly linked by the DAG. If that pair is not directly linked, the algorithm identifies the set of "conditioning variables" that (if held constant) would result in that pair then being independent. That pair of variables and the set of conditioning variables is then recorded as a "conditional independence relationship". We automate this step using code extracted from the R package $ggm$ (Marchetti, 2006);
2. fitting each d-separation relationship as a regression model, and extracting the p-value $p_i$ associated with rejecting the null hypothesis for each conditional independence relationship from Step 1;
3. combining these p-values using Fisher's formula, $C = -2\log(\sum_{i=1}^{N} p_i)$, and calculating an overall (a.k.a. "omnibus") p-value representing the strength of evidence that the model is incorrectly specified, under the assumption that $C$ follows a chi-squared distribution with $2N$ degrees of freedom.

This d-sep test is specifically designed to identify whether a hypothesized causal structure is more inconsistent with available data than would be expected by chance alone (i.e., falsify the causal hypothesis). It is distinct from standard diagnostic tests (e.g., omnibus tests or visual inspection of model residuals), which are designed to falsify the statistical assumptions of the fitted model (e.g., the assumed distribution for residual or process errors, linearity, homoskedasticity, etc.). To see this distinction, we note that standard diagnostic tests inspect the goodness-of-fit for the included direct effects in the model, whereas the Shipley $d$-sep test

evaluates whether the effects that are constrained to zero (that is, the causal relationships the model assumes are absent) are indeed justifiably absent. Here, we focus on developing and exploring performance for the time-series extension of the d-sep test in isolation. Future research could explore the performance of a workflow that combines standard diagnostics and d-sep tests.

*Simultaneous and lagged effects in time-series structural equation models*

We seek to generalize the d-sep test for application in time-series models that can include both simultaneous and lagged interactions among variables. Next, we briefly summarize dynamic structural equation models (DSEM). For a set of $j \in \{1,2,\dots J\}$ variables over $t \in \{1,2,\dots,T\}$ time intervals, we define a matrix of latent variables $\mathbf{X}$ with dimension $T \times J$. DSEM then defines a structural vector-autoregressive (SVAR) process for row-vector $\mathbf{x}_t$ containing $x_{tj}$ for all variables in time $t$:

$$\mathbf{x}_t = \underbrace{\mathbf{B}_0\mathbf{x}_t}_{\text{Simultaneous}} + \underbrace{\mathbf{B}_1\mathbf{x}_{t-1}}_{\text{Lag}-1} + \underbrace{\dots}_{\text{Higher}-\text{order}} + \boldsymbol{\epsilon}_t \quad (1)$$

where $\mathbf{B}_0$ are simultaneous interactions among variables, $\mathbf{B}_1$ is lag-1 interactions, and the model can include any arbitrary lag up to $T-1$ (indicated by ... in Eq. 1). We can then re-write this as a simultaneous equation model by defining a lower-triangle joint path matrix $\mathbf{P}_{\text{joint}}$ with dimension $JT \times JT$. For illustration when $T = 4$, this results in a joint path matrix:

$$\mathbf{P}_{\text{joint}} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_1 & \mathbf{B}_0 & \mathbf{0} & \mathbf{0} \\ \dots & \mathbf{B}_1 & \mathbf{B}_0 & \mathbf{0} \\ \dots & \dots & \mathbf{B}_1 & \mathbf{B}_0 \end{bmatrix} \quad (2)$$

where ... again indicates the potential inclusion of higher-order lag matrices. This defines a simultaneous equation:

$$\text{vec}(\mathbf{X}) = \mathbf{P}_{\text{joint}}\text{vec}(\mathbf{X}) + \text{vec}(\mathbf{E}) \quad (3)$$

$$\text{vec}(\mathbf{E}) \sim \text{MVN}(\mathbf{0}, \mathbf{V}_{\text{joint}})$$

where $\mathbf{E}$ is the $J \times T$ matrix of exogenous errors, $\mathbf{V}_{\text{joint}}$ is the $JT \times JT$ covariance for these errors (which is assumed to be block-diagonal, i.e., zero for any errors occurring in different times), and $\text{vec}(\mathbf{X})$ is the operator that stacks the $J$ columns into a single vector of length $JT$. Conveniently, this simultaneous equation can be re-arranged as a Gaussian Markov random field where $\mathbf{Q} = (\mathbf{I} - \mathbf{P}_{\text{joint}}^{t})\mathbf{V}_{\text{joint}}^{-1}(\mathbf{I} - \mathbf{P}_{\text{joint}})$ is the sparse precision (inverse-covariance) matrix. The probability density of this GMRF can then be rapidly evaluated using the sparse precision $\mathbf{Q}$, and it can be fitted efficiently using the Laplace approximation as a GLMM. The model is completed by defining a distribution for data matrix $\mathbf{Y}$ with dimensions $T \times J$. For each column $\mathbf{y}_j$, the user can specify that measurements are without error (i.e., $\mathbf{y}_j = \mathbf{x}_j$) or can specify a link function and distribution, i.e., $y_{tj} \sim f_j(g_j^{-1}(x_{tj}), \theta_j)$ where $g_j^{-1}(x_{tj})$ is the inverse-link function, $\theta_j$ is the estimated variance for measurement errors, and $f_j$ is the distribution for errors. In the following, we focus upon the case of no measurement errors (i.e., $\mathbf{y}_j = \mathbf{x}_j$), which then collapses to a "process error" model. Importantly, this process-error model can include missing values where $y_{tj} = \text{NA}$.

DSEM is specified using "arrow-and-lag" notation. For example, a one-headed arrow, $A \rightarrow B, 1$ indicates that variable $A$ in time $t$ affects $B$ in time $t + 1$ and corresponds to parameter in the lag-1 interaction matrix $\mathbf{B}_1$. In addition to restricting dynamics to a DSEM (i.e., linear interactions), in the following we make the following restrictions: (1) that exogenous covariance is diagonal; (2) that variables can be re-ordered such that simultaneous interaction matrix $\mathbf{B}_0$ is lower-triangle (i.e., a recursive graph); and (3) that there are no "latent variables" that are entirely missing observations. Future research could relax these restrictions using insights from ongoing research in causal discovery methods, e.g., using SVAR-FCI as developed for SVAR

models such as DSEM (Malinsky & Spirtes, 2018). In particular, assumption-3 ("no latent variables") is a key assumption of our present work, and SVAR-FCI replaces this by applying "m-separation" to incorporate two-headed arrows that arise from marginalization across latent variables. M-separation was introduced by Richardson & Spirtes (2002) and applied to VARs by Eichler (2007). However, m-separation has not been discussed in recent ecological reviews of causal falsification (Arif & MacNeil, 2023; Grace, 2024), so we leave it as a topic for future extensions (but see preprint: Correia et al., 2025). Similarly, PCMCI+ allows nonlinear relationships among variables (Runge, 2022). We recommend that future research introduce both topics for ecological time-series analysis.

*Conditional independence in time-series modelling*

Using a DSEM with maximum lag $M = 1$ implies that the $J$ variables $\mathbf{x}_t$ in time $t$ might depend upon $\mathbf{x}_t$ but also $\mathbf{x}_{t-1}$. Therefore, the d-sep test involves testing conditional independence relationships among a set of $J(M+1)$ pairwise relationships, representing each variable $j \in \{1,2,\dots,J\}$ at each potential lag $m \in \{0,\dots,M\}$ where $M$ is the maximum lag included in the model (see Table 1 for an overview of the time-series d-sep algorithm). This insight yields a further complication. Say for a maximum lag of $M = 1$, variable $x_{t,j}$ and $x_{t+1,j^*}$ might be independent only when conditioning upon preceding states $x_{t-1,j^*}$. To see this, consider a bivariate time-series model with maximum lag $M = 1$:

$$\begin{aligned} A &= \beta_1 \mathrm{lag}_1(A) + \epsilon_A \\ B &= \beta_2 A + \beta_3 \mathrm{lag}_1(B) + \epsilon_B \end{aligned} \tag{4A}$$

where $\mathrm{lag}_1(A)$ indicates the lag-1 operator for variable $A$ such that $A$ has a simultaneous (lag-0) impact on B, and both $A$ and $B$ exhibit first-order autocorrelation (e.g., Gompertz density dependence). This is specified in arrow-and-lag notation as:

$$A \rightarrow A, 1$$
$$A \rightarrow B, 0 \quad (4B)$$
$$B \rightarrow B, 1$$

As our later algorithm shows, this model implies two CI relationships (Fig. 1). The first implies that $B_{t+1}$ is independent of the preceding $A_t$ conditional upon fixed values for:

1. $A_{t-1}$, because $A_t \leftarrow A_{t-1} \rightarrow B_{t-1} \rightarrow B_t \rightarrow B_{t+1}$ such that variation in $A_{t-1}$ causes a correlation between $A_t$ and $B_{t+1}$; and
2. $B_t$, because $A_t \rightarrow B_t \rightarrow B_{t+1}$, such that a fixed value for $B_t$ blocks (a.k.a. controls for) the correlation between $A_t$ and $B_{t+1}$;
3. $A_{t+1}$, because $A_t \rightarrow A_{t+1} \rightarrow B_{t+1}$, such that a fixed value for $A_{t+1}$ blocks the correlation between $A_t$ and $B_{t+1}$.

We can therefore test for this CI relationship by fitting an alternative time-series model:

$$A = \epsilon_A$$
$$B = \beta_0 \text{lag}_1(A) + \beta_1 \text{lag}_2(A) + \beta_2 \text{lag}_1(B) + \beta_3 A + \epsilon_B \quad (5A)$$

and testing whether $\beta_0$ is significantly different from zero. This CI relationship is then specified in arrow-and-lag notation as:

$$A \rightarrow B, 1$$
$$A \rightarrow B, 2$$
$$B \rightarrow B, 1 \quad (5B)$$
$$A \rightarrow B, 0$$

where the parameter in the first line corresponds to $\beta_0$.

This example therefore illustrates that we need to test for conditioning variables at lag-2 when fitting a maximum lag $M = 1$, and in general we need to include conditioning variables for

$M$ prior times given a maximum lag of $M$. In the case of $M = 0$ (i.e., no lagged effects), then we can again ignore conditioning variables prior to the time of interest, and the protocol collapses to the three steps in the standard d-sep test (see beginning of the Methods section).

To define conditional independence relationships in time-series models involving maximum lag $M$ and $J$ variables, we therefore define a conditioning matrix $\mathbf{A}$ with dimension $J(M+1) \times J(M+1)$. For the case of maximum lag $M = 1$, we have:

$$\mathbf{A} = \begin{bmatrix} \mathbf{B}_0 & 0 & 0 \\ \mathbf{B}_1 & \mathbf{B}_0 & 0 \\ 0 & \mathbf{B}_1 & \mathbf{B}_0 \end{bmatrix} \quad (6)$$

where the first row and column are the conditioning (or "burn-in") interval where conditioning variables might arise, and we only test for CI relationships among the 2nd and 3rd rows and columns. To do so, we first define all conditional independence relationships within that conditioning matrix $\mathbf{A}$, in this case by copying functions from the R package *ggm*. However, we only keep those that define an independent relationship between two variables that are both after the $M = 1$ "burn-in" intervals, while still allowing conditioning variables to occur anywhere in the matrix $\mathbf{A}$. We then iterate sequentially through each conditional independence relationship, where we sequentially fit DSEM with that specified relationship, calculate the p-value for a two-sided Wald test, and combine these using Fisher's formula.

As further complication, we reiterate that DSEM can account for missing data (i.e., $y_{tj} =$ NA). In these instances, we impute missing data from the predictive distribution of random effects (i.e., the precision matrix $\mathbf{H}$ given available data and fixed effects), and then use these imputed data as "fixed" for each CI test. We explored alternative options where we re-simulate missing data independently for each CI relationships, or used a single imputed data set across all CI relationships for a given d-sep test. This exploration suggested relatively little difference in

253 performance, and we show the former in the following. We note that imputing a single replicate
254 of missing data and using that in multiple CI tests will likely lead to correlated p-values, and
255 therefore a less sensitive omnibus test. Future studies could explore alternative strategies for
256 data-imputation to improve statistical efficiency.

257 *Simulation experiment*

258 To explore the likely performance of this proposed application of omnibus d-separation testing,
259 we first conduct a factorial simulation experiment. This involves 500 replicates of each
260 combination of the following levels:

261 1. *Three simulation models*: We simulate data from three different dynamic structural equation
262 models. The simplest ("sem") has four variables and only simultaneous effects, where $A \rightarrow$
263 $B$, $A \rightarrow C$, $B \rightarrow D$, and $C \rightarrow D$. The intermediate ("dsem_simple") involves two variables
264 with simultaneous and lagged effects, where $A \rightarrow B$, and an autoregressive process for both $A$
265 and $B$. The most complicated ("dsem_complex") involves four variables, combining the
266 same simultaneous effects as the "sem" scenario, but also including first-order
267 autocorrelation for each variable. The intermediate "dsem_simple" corresponds to the
268 example discussed in the *Conditional independence in time-series modelling* section (Eq. 4);
269 2. *Three sample sizes*: We simulate time-series of length $T = \{25, 50, 100\}$, representing short,
270 medium, and long ecological data sets;
271 3. *Five levels of missing data*: We randomly exclude data for each combination of variable and
272 year, with probability $p_{\text{missing}} = \{0, 0.1, 0.2, 0.35, 0.5\}$;
273 4. *Two estimation models*: For each combination of simulation model, sample size, and missing
274 data, we fit DSEM either using the true model structure ("right"), or using a mis-specified
275 DSEM ("wrong"; see Fig. 2);

This design therefore involves $3 \times 3 \times 5 \times 2 \times 500 = 45{,}000$ applications of the time-series d-sep test.

We assess two characteristics for the d-sep test in this experiment:

1. *Calibration:* A well-calibrated test will result in a uniform $U(0,1)$ distribution for p-values when the simulation model matches the estimation model;
2. *Efficiency*: An efficient test will result in a large proportion of p-values that are close to zero when the estimation model does not match the simulation model. Ideally, this p-value will remain close to zero even when time-series are short, the simulation model is complicated, and a large proportion of data are missing.

*Case study applications*

We also demonstrate the potential use of time-series d-sep via application to two real-world data sets:

1. *Wolf-moose interactions on Isle Royale*: Building upon an analysis from Thorson et al. (2024), we re-analyze a population census of wolves and moose on Isle Royale from 1959-2019 (Vucetich & Peterson, 2012), where $W$ and $M$ are log-abundance of wolves and moose, respectively. We fit a model with just Gompertz density dependence ($W \rightarrow W, 1$ and $M \rightarrow M, 1$), adding bottom up interactions ($M \rightarrow W, 1$), adding top-down interactions ($W \rightarrow M, 1$), or adding both;
2. *Spawning phenology and environment*: In a new example of DSEM, we use published data representing spawning phenology for walleye pollock in the Gulf of Alaska from 1992-2021 and its relationship to survey availability (Rogers et al., 2025). This includes four variables, representing sea surface temperature $T$, the average number of days between mean date of spawning (as estimated from larval-derived hatch dates) and the mean date of a survey $A$, the

logit-transformed proportion of females >30cm in a spawning or spent stage during the spawning-grounds survey $P$, and the survey availability $Q$ measured as log-ratio between the surveyed biomass and predicted biomass where the latter is taken from a population dynamics model fitted to the survey data without accounting for timing or temperature (Monnahan et al., 2021). We explore three alternative models for these data. The first ("temperature as driver") views temperature as the driver of all other variables (i.e., $T \rightarrow A$, $T \rightarrow P$, and $T \rightarrow Q$). The second ("availability regression") views variables as independent predictors of survey availability (i.e., $T \rightarrow Q$, $P \rightarrow Q$, and $A \rightarrow Q$). The third ("timing as mediator", described in Rogers et al. 2025) claims that temperature affects survey availability via its mediating effect on spawning phenology (i.e., $T \rightarrow A$, $A \rightarrow P$, and $A \rightarrow Q$). Across all three models, we also estimate first-order autoregression for each variable (i.e., $T \rightarrow T, 1$, $A \rightarrow A, 1$, $P \rightarrow P, 1$, and $Q \rightarrow Q, 1$) and assume that variables are measured without error (i.e., a process-error model)

In each case study, we record the p-value from the time-series d-sep test as well as the marginal Akaike Information Criterion (AIC) for the fitted model. In the following, we use AIC as additional information to compare among models that are not falsified using the proposed test.

## Results

### *Simulation experiment*

We first illustrate the performance (i.e., calibration and efficiency) of the proposed test across simulation models and time-series lengths when data are complete (Fig. 3). In the simulation model without lagged effects (Fig. 3 top row), the correct model has an approximately uniform $U(0,1)$ distribution for p-values across all sample sizes indicating that the test is well calibrated. Similarly, the incorrect model results in a p-value $< 0.1$ in nearly all replicates, indicating that

the test is statistically efficient across sample sizes. Moving to the two-variable model with lags (Fig. 3 middle row), we see that the test is well calibrated across time-series lengths (i.e., the correct model results in an approximately uniform distribution of p-values). However, it only detects the mis-specification of an incorrect model (i.e., a p-value $< 0.1$) in 60% of the replicates at low sample sizes ($T = 25$) and 80% of replicates at intermediate sizes ($T = 50$), before attaining good performance for long time-series ($T = 100$). Finally, for the four-variable model with lags (Fig. 3 bottom row), we see that the test is poorly calibrated (i.e., departs from a $U(0,1)$ distribution) for short time-series and incorrectly identifies the model as mis-specified in nearly 40% of replicates. It then becomes well calibrated as the time-series length increases. Expanding this experiment across different levels of missing data (Fig. 4), we see that the simple estimation model remains well calibrated across the level of missing data (Fig. 4 top row), but that the efficiency drops as $p_{\text{missing}}$ increases from 0 to 50%. A similar pattern holds for the other simulation models (Fig. 4 middle and bottom rows). However, the decline in efficiency is notable at a lower value of $p_{\text{missing}}$ in the intermediate-complexity simulation model (Fig. 4 middle row), and the complex simulation model remains poorly calibrated across levels of missing data for short sample sizes (Fig. 4 bottom-left panel, red bullets).

*Case studies*

We also use two real-world case studies to illustrate the types of ecological inference that are feasible when using the proposed test to falsify hypotheses using time-series models. In the case study involving predator-prey interactions of moose and wolves in Isle Royale (Fig. 5), we explored four models corresponding to single-species (Gompertz) density dependence, adding bottom-up or top-down interactions individually, and adding both interactions jointly. The test then provides strong evidence ($p < 0.01$) that the "bottom-up" model is incorrect, weak

345 evidence ($p = 0.15$) that the model with only density dependence is incorrect, and no evidence
346 ($p > 0.9$) to reject the remaining two models. We therefore use AIC to conclude that the model
347 with top-down interactions is parsimonious ($\Delta AIC = 0$) and not falsifiable relative to the model
348 with both interactions ($\Delta AIC = 1.1$). In the case study involving spawning phenology and
349 survey availability for pollock in the Gulf of Alaska (Fig. 6), we explored three models
350 representing "temperature as driver", "availability regression" or "timing as mediating effect"
351 hypotheses. The test provides strong evidence ($p < 0.01$) to falsify the first two models, but
352 fails to reject the phenology model ($p = 0.7$). We therefore conclude that this is the most
353 appropriate interpretation of those data given the proposed causal hypotheses.

354 **Discussion**

355 Conditional independence testing is an established practice in structural equation models and
356 phylogenetic path analysis. Here, we demonstrate its application to falsify causal hypothesis
357 regarding simultaneous and lagged interactions among ecological time using structural vector
358 autoregressive models like DSEM. Our simulation experiment confirms that the algorithm
359 proposed here is well calibrated, and that short time series ($T = 25$) can be sufficient for simple
360 structural models with complete data, but that longer time series ($T = 100$) are required as
361 model complexity increases. Similarly, the test efficiency drops as the proportion of missing
362 data increases towards $p_{\text{missing}} = 0.5$. Finally, the case studies illustrate that the test will retain
363 several candidate models in some cases (i.e., for the Isle Royale data set), such that assessing
364 model parsimony and multi-model averaging might be appropriate in these cases. In other cases
365 (e.g., involving pollock spawning phenology), the test provides quantitative support for the
366 ecological interpretation of observational data.

Here, we have restricted ourselves to small systems (scenarios involving 2-4 variables) and few lags (simultaneous and first-order cross-lags). We do this because the limits of the test are already evident at this small model size. For example, using 4-variables with first-order lags and using short time series ($T = 25$), we already see poor calibration (i.e., rejecting the true model above intended rates). To understand this, consider that $J = 4$ variables and one lag involves up to $\frac{2J(2J+1)}{2} = 36$ conditional independence relationships to test. The number of CI relationships therefore grows as the square of the number of variables, and the test seems to lose power rapidly for sample sizes that are common when analyzing annualized dynamics. Presumably this loss of statistical power is why previous simulation tests of d-sep in ecology (e.g., in phylogenetic path analysis) have involved systems with $< 5$ variables (von Hardenberg & Gonzalez-Voyer, 2013). In summary, the time-series d-separation test explored here was unreliable when applied to models with many variables, particularly when time-series were relatively short or had missing values.

Others have advocated that ecologists adopt a causal analysis framework, which is a workflow for developing and quantifying DAGs and understanding causal linkages from observational data (Arif & MacNeil, 2023; Grace & Irvine, 2020). Adopting this framework could help mitigate biases associated with traditional statistical models (e.g., linear regression) and understand causality. D-sep is one step in this workflow and broadly tests consistency between DAGs and data, or whether the data support the DAG structure (i.e., configuration of linkages) ('Step 2' in Figure 2 of Arif & MacNeil (2023), part of 'Step 3' in Figure 2 of Grace & Irvine (2020)). The backdoor and frontdoor criteria are other steps in the workflow that identify whether DAGs are susceptible to confounding variables, which introduce bias into the estimation of parameters and misrepresent causal linkages (Arif & MacNeil, 2023; Byrnes & Dee, 2025;

Pearl, 2009). The backdoor and frontdoor criteria are unavailable for DSEM models but are needed to advance our understanding of using DSEM to identify causal linkages among variables from models fitted to time-series data with simultaneous and lagged interactions, as well as missing data. Consequently, it will not be possible to follow all recommended steps in the causal analysis framework, such as those in Arif and MacNeil (2023). Further, many correct DAGs will fail a d-sep test for reasons including DAG complexity, time series length, and the presence of missing data, as our simulation showed. When communicating the results from models where we expect d-sep to be less reliable, analysts should take care to acknowledge the potential for biases in parameter estimates due to model mis-specification, explain model assumptions, and be explicit about the limits of causal inference (Grace & Irvine, 2020; Siegel & Dee, 2025).

We also note that d-sep is only testing for significant linear relationships among variables, and therefore cannot detect nonlinear or state-dependent relationships (unless they can be expressed using lagged linear relationships). We therefore recommend further cross-comparison with nonlinear causal analysis, e.g., using "empirical dynamic modelling" EDM (Munch et al., 2023). EDM has proven to be powerful in detecting nonlinear causal systems, as validated via microcosm experiments and methods comparisons (Chang et al., 2022; Sugihara et al., 2012). However, EDM also appears to be more informative with longer time series. We therefore envision a workflow using linear models (e.g., d-sep tests for a DSEM) when time-series are relatively short, and comparison with a nonlinear method for longer time-series. We also encourage further work estimating a linear "skeleton" within EDM models, so that EDM collapses to linear interactions when data are limited, but can express a wide range of nonlinear systems when data are abundant. Both DSEM and EDM involve fitting a Gaussian process model, so it seems like their statistical integration would be feasible in future statistical research.

Recent studies have pursued a rich vein of parallel line of research for "causal discovery," i.e., using observational data to identify what combination of one- and two-headed arrows can be identified from available data. Starting with the FCI algorithm (Spirtes et al., 2000), these causal-discovery algorithms typically start with a fully-connected causal model and then proceed backwards by either (A) identifying pairs of variables that are conditionally independent, or (B) triplets that have a specific structure. In particular, the SVAR-FCI algorithm is applicable to the DSEM explored here (Malinsky & Spirtes, 2018), and provides many insights (e.g., identifying two-headed arrows arising from latent variables) relative to the algorithm tested here. Similarly, PCMCI+ incorporates nonlinear linkages among larger numbers of variables (Runge, 2022). However, we believe that causal discovery involves a different goal than the one addressed here: we instead start with one (or a small number of) causal hypotheses that are derived from ecological knowledge, and then seek to falsify that specific hypothesis. For scientists who have already developed hypotheses about system dynamics, we think that this "falsification" step remains important and separate from parallel research regarding causal discovery.

In summary, we recommend that analysts seek to falsify causal assumptions for time-series models when they are intended for causal analysis. When developing an DSEM, we recommend that only models with a priori ecological support that also pass the d-sep test be considered, and that model parsimony or averaging then be considered for those models that are consistent with data (i.e., pass the d-sep test). However, in models with 5+ variables and lagged dynamics, we caution that d-sep appears to be poorly calibrated such that models may be erroneously rejected. We therefore recommend ongoing research to integrate causal falsification and discovery into ecological workflows.

**Works cited:**


Arif, S., & MacNeil, M. A. (2022a). Predictive models aren't for causal inference. *Ecology Letters*, *25*(8), 1741–1745. https://doi.org/10.1111/ele.14033

Arif, S., & MacNeil, M. A. (2022b). Utilizing causal diagrams across quasi-experimental approaches. *Ecosphere*, *13*(4), e4009. https://doi.org/10.1002/ecs2.4009

Arif, S., & MacNeil, M. A. (2023). Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*, *93*(1), e1554. https://doi.org/10.1002/ecm.1554

Byrnes, J. E. K., & Dee, L. E. (2025). Causal Inference With Observational Data and Unobserved Confounding Variables. *Ecology Letters*, *28*(1), e70023. https://doi.org/10.1111/ele.70023

Chang, C.-W., Munch, S. B., & Hsieh, C. (2022). Comments on identifying causal relationships in nonlinear dynamical systems via empirical mode decomposition. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-30359-8

Correia, H. E., Dee, L. E., Byrnes, J. E. K., Fieberg, J., Fortin, M.-J., Glymour, C., Runge, J., Shipley, B., Shpitser, I., Siegel, K. J., Sugihara, G., Holle, B. von, & Ferraro, P. J. (2025). *Best practices for moving from correlation to causation in ecological research*. https://ecoevorxiv.org/repository/view/9361/

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, *137*(2), 334–353. https://doi.org/10.1016/j.jeconom.2005.06.032

Grace, J. B. (2024). An integrative paradigm for building causal knowledge. *Ecological Monographs*, *94*(4), e1628. https://doi.org/10.1002/ecm.1628

Grace, J. B., & Irvine, K. M. (2020). Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology*, *101*(4), e02962. https://doi.org/10.1002/ecy.2962

Malinsky, D., & Spirtes, P. (2018). Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, 23–47. https://proceedings.mlr.press/v92/malinsky18a.html

Marchetti, G. M. (2006). Independencies induced from a graphical Markov model after marginalization and conditioning: The R package ggm. *Journal of Statistical Software*, *15*, 1–15.

Meziane, D., & Shipley, B. (2001). Direct and Indirect Relationships Between Specific Leaf Area, Leaf Nitrogen and Leaf Gas Exchange. Effects of Irradiance and Nutrient Supply. *Annals of Botany*, *88*(5), 915–927. https://doi.org/10.1006/anbo.2001.1536

Monnahan, C. C., Dorn, M. W., Deary, A. L., Ferriss, B. E., Fissel, B. E., Honkalehto, T., Jones, D. T., Levine, M., Rogers, L., & Shotwell, S. K. (2021). *Assessment of the Walleye Pollock Stock in the Gulf of Alaska*.

Munch, S. B., Rogers, T. L., & Sugihara, G. (2023). Recent developments in empirical dynamic modelling. *Methods in Ecology and Evolution*, *14*(3), 732–745. https://doi.org/10.1111/2041-210X.13983

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. https://doi.org/10.1214/09-SS057

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle, *Handbook of structural equation modeling* (pp. 68–91). Guilford press.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Richardson, T., & Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, *30*(4), 962–1030. https://doi.org/10.1214/aos/1031689015

Rogers, L. A., Monnahan, C. C., Williams, K., Jones, D. T., & Dorn, M. W. (2025). Climate-driven changes in the timing of spawning and the availability of walleye pollock (Gadus chalcogrammus) to assessment surveys in the Gulf of Alaska. *ICES Journal of Marine Science*, *82*(1), fsae005. https://doi.org/10.1093/icesjms/fsae005

Runge, J. (2022). *Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets* (No. arXiv:2003.03685). arXiv. https://doi.org/10.48550/arXiv.2003.03685

Shipley, B. (2000). A New Inferential Test for Path Models Based on Directed Acyclic Graphs. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(2), 206–218. https://doi.org/10.1207/S15328007SEM0702_4

Shipley, B. (2009). Confirmatory path analysis in a generalized multilevel context. *Ecology*, *90*(2), 363–368. https://doi.org/10.1890/08-1034.1

Siegel, K., & Dee, L. E. (2025). Foundations and Future Directions for Causal Inference in Ecological Research. *Ecology Letters*, *28*(1), e70053. https://doi.org/10.1111/ele.70053

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT press. https://books.google.com/books?hl=en&lr=&id=vV-U09kCdRwC&oi=fnd&pg=PR9&dq=Causation,+Prediction,+and+Search&ots=DZ_WpqwMqf&sig=5rMntPAlTR0eWEhQMkWx1Y9OCr8

Sugihara, G., May, R., Ye, H., Hsieh, C., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting Causality in Complex Ecosystems. *Science*, *338*(6106), 496–500.

Thorson, J. T., Andrews III, A. G., Essington, T. E., & Large, S. I. (2024). Dynamic structural equation models synthesize ecosystem dynamics constrained by ecological mechanisms. *Methods in Ecology and Evolution*, *15*(4), 744–755. https://doi.org/10.1111/2041-210X.14289

von Hardenberg, A., & Gonzalez-Voyer, A. (2013). Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution; International Journal of Organic Evolution*, *67*(2), 378–387. https://doi.org/10.1111/j.1558-5646.2012.01790.x

Vucetich, J. A., & Peterson, R. O. (2012). *The population biology of Isle Royale wolves and moose: An overview*. www.isleroyalewolf.org

Table 1: Summarizing the steps required when extending the d-separation test for use in time-series models that include both simultaneous and lagged relationships among variables.

| Number | Title | Description |
|---|---|---|
| 1 | Extract path matrix | Extract path matrix, including conditioning interval for maximum number of lags to define conditioning matrix **A** |
| 2 | Define conditional independence (CI) relationships | Use directional separation ("d-sep") to define the set of CI relationships |
| 3 | Eliminate relationships | Eliminate duplicative CI relationships, and restrict target and predictor variables outside the initialization buffer, while allowing conditioning variables within the "burn-in" interval |
| 4 | Simulate missing data from predictive distribution | Simulate any missing data, either once across all CI tests or separately for each CI test |
| 5 | Fit CI relationships and combine p-values | Fit each CI relationship, record the p-value for each individual CI test, and combine them using Fisher's formula |

Fig 1: A visual depiction of the two conditional-independence relationships implied by the "dsem_simple" structural causal model DSEM, as calculated using conditioning matrix **A** (Eq. 6). The CI relationship is shown with a solid line, while the conditioning variables are shown as dashed lines. Given a DSEM with maximum lag $M = 1$, the CI must condition upon a maximum of lag-2 relationships; e.g., the top CI relationship can be fitted as $b = \beta_0 \text{lag}_1(a) + \beta_1 a + \beta_2 \text{lag}_2(a) + \epsilon$ where we then test for the significance of the $\beta_0$ coefficient.

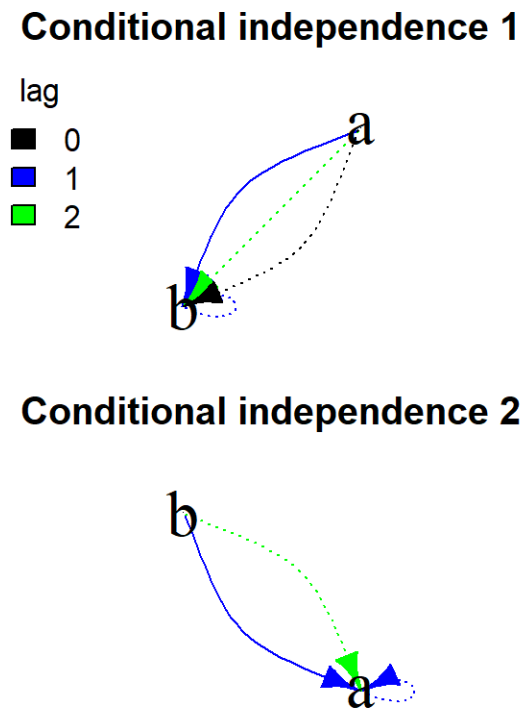


Fig. 2: The dynamic structural equation model (DSEM) used to simulate data (left column) in three simulation scenarios (rows), and the DSEM that is specified when intentionally fitting with a mismatched SCM (right column). In each DSEM, we show 2-4 time-series variables (labeled "a" through "d"), and causal paths showing either simultaneous effects (black arrows) or lag-1 effects (blue arrows), where a blue arrow from a variable to itself (e.g., in the 2nd row) shows a first-order autoregressive effect.

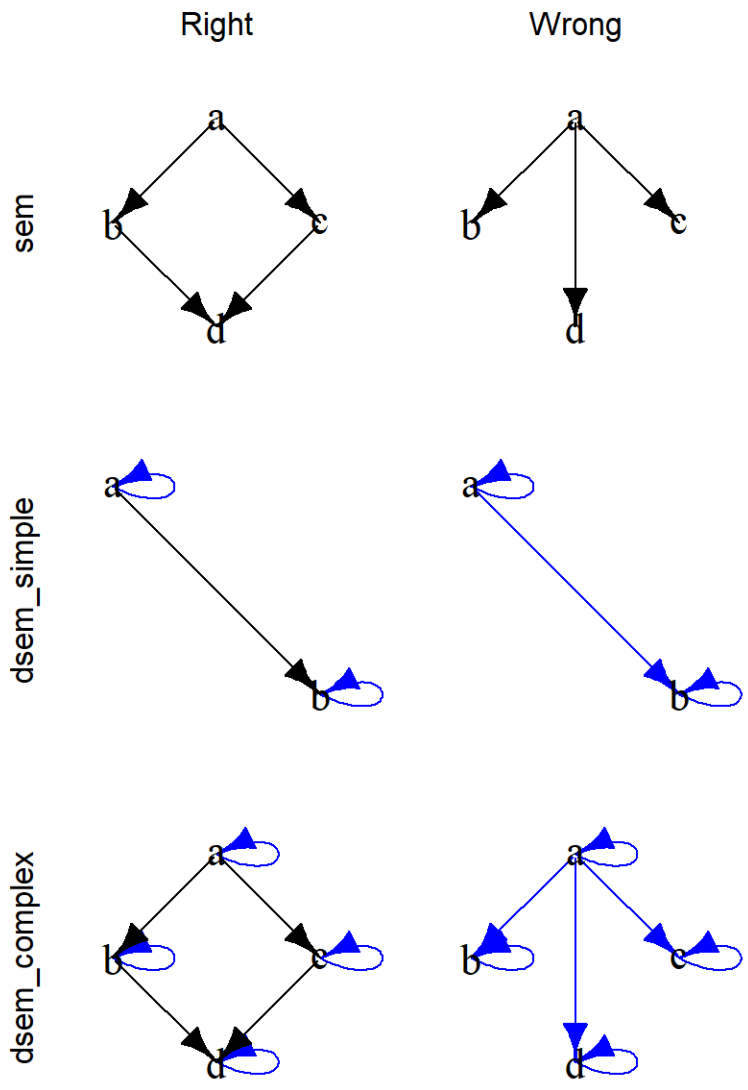


Fig. 3: Results from the simulation experiment showing the frequency of 500 replicates (y-axis) with a given p-value (x-axis) for a time-series d-separation test, while simulating time-series of length $T = \{25, 50, 100\}$ (columns) from three dynamic structural equation models DSEM (rows, see Fig. 1 left column). Simulated data were either fitted with the correct DSEM (red histogram, Fig. 1 left column) or wrong DSEM (blue histogram, Fig. 1 right column). A well-calibrated d-separation test will result in a p-value that follows a uniform $U(0,1)$ distribution (i.e., horizontal dashed line) when fitting the correct model, and an efficient test will result in a p-value that is close to zero when fitting a mis-specified model.

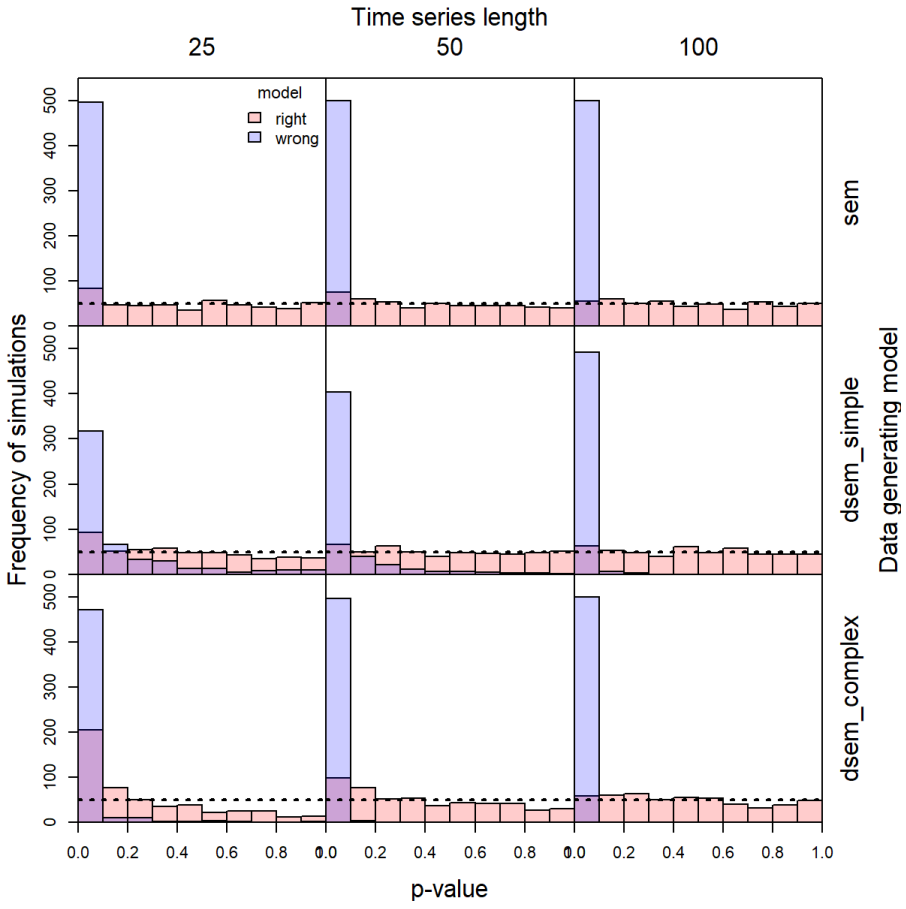
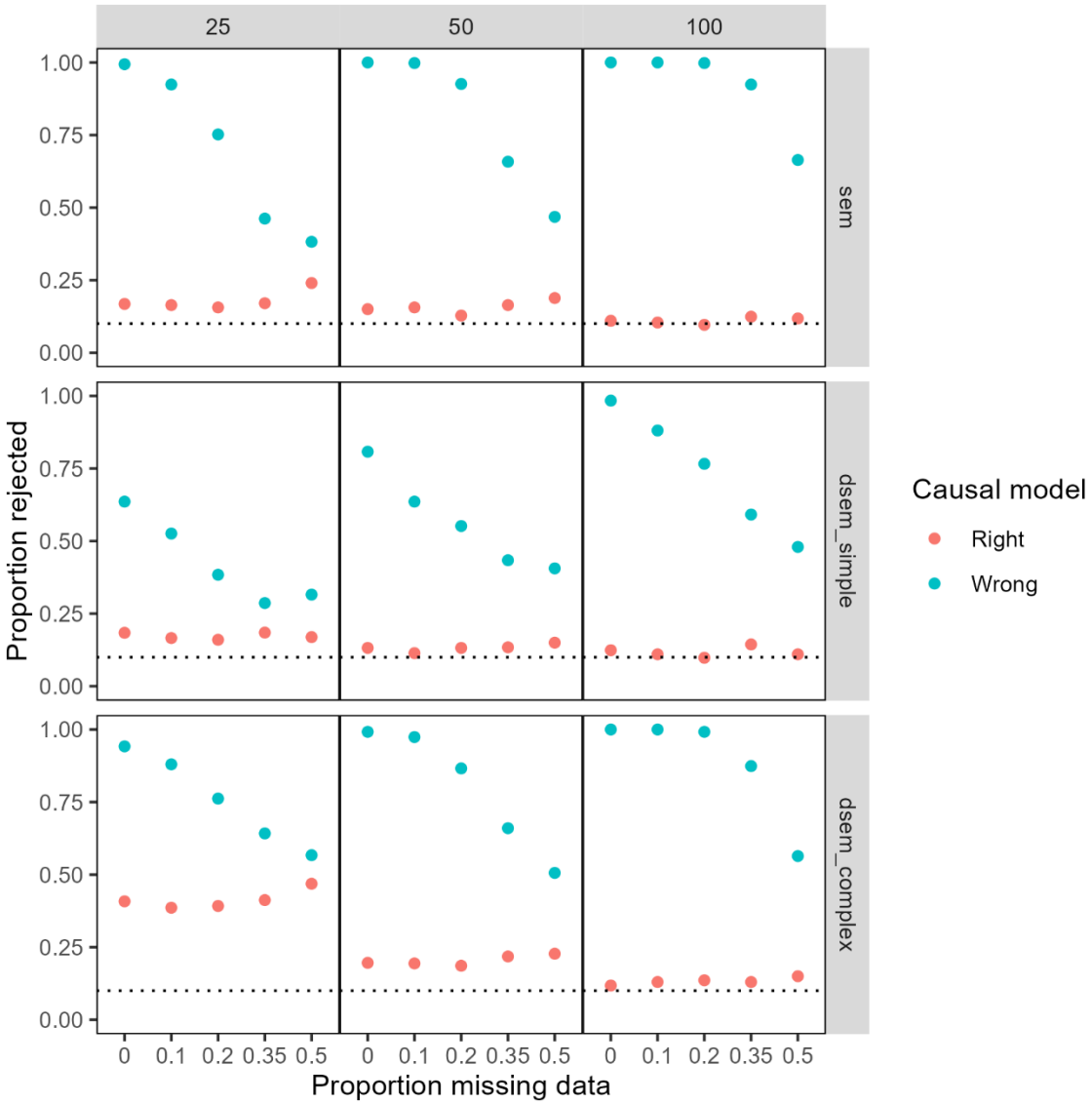
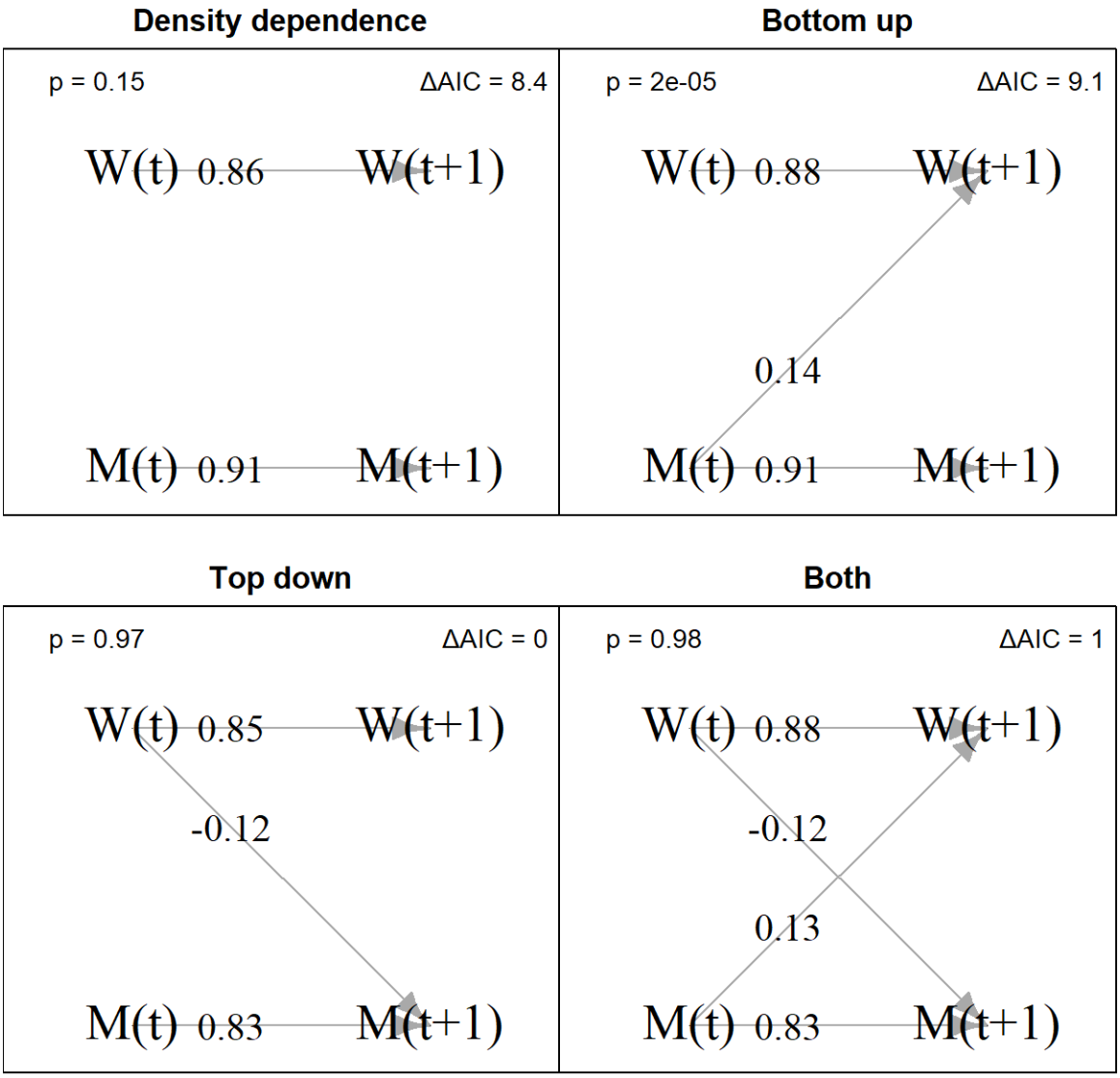


Fig. 4: Results from the simulation experiment when showing the proportion of simulation replicates with d-separation test resulting in $p < 0.1$ (y-axis) across five proportions of missing data $p_{\text{missing}} = \{0, 0.1, 0.2, 0.35, 0.5\}$ (x-axis), and across different time-series lengths (columns) and dynamic structural equation models DSEM (rows, see Fig. 3 caption for more details). A well-calibrated model will reject the test at a nominal 0.1 rate (horizontal dotted lines) when the DSEM causal assumptions are correct, and ideally will reject it at close to 1.0 rate when the DSEM assumptions are mis-specified.



553 Fig. 5 – Estimated dynamic structural equation model showing a vector-autoregressive model
554 fitting to data for wolf ($W$) and moose ($M$) log-abundance in Isle Royale 1959-2019 (Vucetich &
555 Peterson, 2012). We compare a model assuming Gompertz density dependence (i.e., $W \rightarrow W, 1$
556 and $M \rightarrow M, 1$), adding either bottom-up or top-down controls, or adding both jointly. For each
557 model, we show the time-series d-sep test p-value (p, top-left corner) and the delta-marginal
558 Akaike Information Criterion (top-right corner), where the most parsimonious model has $\Delta\text{AIC} =$
559 0.




560

561

Fig. 6: Estimated dynamic structural equation model (DSEM) showing the estimated path coefficient between temperature $T$, the average number of days between mean date of spawning and the mean date of a survey on spawning grounds $A$, the logit-transformed proportion of females >30cm in a spawning or spent stage during the spawning-grounds survey $P$, and the log-ratio between the surveyed biomass and predicted biomass given other data $Q$. We show three DSEMs (columns), either using temperature as an explanatory variable for all processes ("Temperature as driver"), using all variables to explain availability ("Availability regression"), or using survey timing as a mediating variable linking temperature to survey availability ("Timing as mediator"). We also show the time-series d-sep p-value (top left) and delta-marginal AIC (top-right) for each model.