Validating causal inference in time series models with conditional-independence tests

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

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

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

* Corresponding author: James.Thorson@noaa.gov

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

Abstract:

Ecologists often use time-series models to approximate dynamics arising from density dependence, species interactions, community synchrony, and other processes. Dynamic structural equation models 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, their interpretation as structural causal models (i.e., counterfactual analysis) requires validating that the assumed dynamics are consistent with available data. In site-replicated and phylogenetic contexts, ecologists validate causal assumptions by testing implied conditional-independence relationships (a directional-separation or “d-sep” test), but this has not been extended to include simultaneous and lagged effects in time-series contexts. Here, we propose a time-series d-sep test and use a simulation experiment and case studies to explore its performance. 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 then have less power to reject an incorrect model. 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 survey availability for a spawning survey. In a previously published analysis involving wolf-moose interactions in Isla Royale, the test supports top-down control but cannot distinguish whether bottom-up control is supported. We conclude that d-sep is a useful test to evaluate the structural validity of a time-series model, allowing ecologists to make better causal inference about dynamical systems from correlated time series data.

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 have been proposed as a useful scientific framework to achieve this (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 DAGs (Pearl, 2009), and resolves well-known issues with bias when making causal statements from predictive statistical models (Arif & MacNeil, 2022a) in particular when there are unobserved confounding variables (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 to be independent of others. 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, to advance understanding of ecological mechanisms it is vital for analysts to be able to validate their causal models fitted to observational time-series data.

Time-series dynamics pose particular challenges, because interactions among variables may be either simultaneous (i.e., occurring much faster than the time-step in available observations) or lagged. Lagged interactions result in temporal dependence, and dependence violates a key statistical assumption of the popular structural equation model (SEM; Pearl, 2012) statistical framework for estimating causal relationships. This then limits practical application of SEM to time-series analysis. Thorson et al. (2024) extended the SEM modeling framework to allow for correlated observations and lagged effects. This dynamic structural equation model (DSEM) framework is efficiently represented as a Gaussian Markov random field and fitted as a generalized linear mixed model, as implemented in the 'dsem' package (Thorson et al., 2024) in the R statistical environment (R Core Team, 2023). DSEM encompasses a wide range of statistical analyses including linear models, errors-in-variables, ARIMA models, dynamic factor analysis, vector autoregressive models, and structural causal models. The DSEM framework allows analysts to test novel causal hypotheses due to its decreased restrictions on data necessary in a standard SEM, but it remains unclear how to validate them when fitted to (correlated) observational data.

In general, the best way to validate a SCM 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). Validating a SCM using observational data generally involves testing whether the specified causal model is consistent with available data. For example, consider a trophic cascade, where we might specify a SCM where predator X affects consumer Y and consumer Y affects producer Z. We write this as two causal paths: $X \rightarrow Y$ and $Y \rightarrow Z$. In this SCM, variation in predators is assumed to be independent of producers, conditional upon a fixed value for consumers (i.e., $Z \perp X|Y$). We can therefore test this conditional independence relationship as a regression ($Z = \beta_X X + \beta_Y Y + \epsilon$), and if the slope β_X significantly departs from zero, then we can "reject" this component of SCM as invalid. This insight is formalized by the directional-separation ("d-sep") test (Shipley, 2000), where all conditional-independence relationships implied by a given SCM are sequentially tested and results are then combined in a single "omnibus" test. This d-sep test is widely used in the analysis of controlled experiments (Meziane & Shipley, 2001) and phylogenetic comparative analysis (von Hardenberg & Gonzalez-Voyer, 2013). The test has previously been extended to multi-level models (Shipley, 2009), but has not to our knowledge been extended to time-series analysis involving a combination of simultaneous and lagged interactions among variables.

We therefore address this by extending d-sep tests to measure whether a proposed time-series structural model is consistent with available data. To address this, we first summarize the d-sep test for path analysis, and then introduce modifications that are necessary for application to time-series models that include simultaneous and lagged effects, or when dealing with missing data. We then provide a simulation experiment to determine whether the proposed test performs correctly (i.e., results in a uniform distribution for p-values) when the model is correctly specified, and also how often it can reject an incorrectly specified model given different simulation model structures, time-series lengths, and amounts of missing data. Finally, we use two real-world case studies to illustrate the types of ecological inference that can be drawn from the time-series d-sep test. Results suggest that the method performs well for simple (2-4 variable) models incorporating simultaneous and lagged effects given the range of time-series that are common in population dynamics (25-100 time points), and the method is freely available as function `test_dsep(.)` in the R package *dsem* for future use.

Methods

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

1. identifying the set of conditional independence (or "d-separation") relationships that are implied by the DAG. This set depends upon an *a priori* ordering of variables. Then for each

unique pair of variables, it identifies whether those variables are 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”. This step can be automated, and we use 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 omnibus p-value under the assumption that C follows a chi-squared distribution with $2N$ degrees of freedom.

We seek to generalize this method for application in time-series models that can include both simultaneous and lagged interactions among variables.

Conditional independence in time-series modelling

Next, we briefly summarize dynamic structural equation models (DSEM). For a set of $j \in \{1,2, \dots J\}$ variables over $t \in \{1,2, \dots, T\}$ times, we define a matrix of latent variables $\mathbf{X}$ with dimension $T \times J$. We can represent any set of simultaneous and lagged interactions by defining a path matrix $\mathbf{P}_{\text{joint}}$ with dimension $JT \times JT$, and defining a simultaneous equation:

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

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

where $\mathbf{E}$ is the $J \times T$ matrix of exogenous errors, and $\mathbf{V}_{\text{joint}}$ is the $JT \times JT$ covariance for these errors. Usefully, this simultaneous equation can be re-arranged as a Gaussian Markov random field:

$$\mathrm{vec}(\mathbf{X}) \sim \mathrm{GMRF}(\mathbf{0}, \mathbf{Q})$$

where $\mathbf{Q} = (\mathbf{I} - \mathbf{P}_{\mathrm{joint}}^{t})\mathbf{V}_{\mathrm{joint}}^{-1}(\mathbf{I} - \mathbf{P}_{\mathrm{joint}})$ 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 Generalized Linear Mixed Model (GLMM).

The joint path matrix $\mathbf{P}_{\mathrm{joint}}$ is formed by summing across simultaneous and lagged effects:

$$\mathbf{P}_{\mathrm{joint}} = \underbrace{\mathbf{G}_0 \otimes \mathbf{P}_0}_{\mathrm{Lag-0}} + \underbrace{\mathbf{G}_1 \otimes \mathbf{P}_1}_{\mathrm{Lag-1}} + \cdots$$

Where $\mathbf{P}_0$ is the $J \times J$ matrix of simultaneous (lag-0) interactions, $\mathbf{P}_1$ is the $J \times J$ matrix of lag-1 interactions, $\mathbf{G}_0$ is a $T \times T$ matrix representing the lag-0 operator (i.e., an identity matrix), $\mathbf{G}_1$ is a $T \times T$ matrix representing the lag-1 operator (i.e., a matrix of 0s with a band of 1s one below the diagonal), $\otimes$ is the Kronecker product, and we only show lag-0 and lag-1 interactions for simplicity of presentation. Similarly, we define the exogenous variance to only include simultaneous cross-correlations:

$$\mathbf{V}_{\mathrm{joint}} = \mathbf{G}_0 \otimes (\mathbf{L}^t\mathbf{L})$$

Where $\mathbf{L}$ can include variances (diagonal elements) and covariances (off-diagonal elements), and represents the Cholesky (i.e., square root) of simultaneous exogenous covariance $\mathbf{L}^t\mathbf{L}$. 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, θ_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.

This model then implies that the J variables $\mathbf{x}_t$ in time t might depend upon $\mathbf{x}_t$ but also $\mathbf{x}_{t-1}$ in a model with a maximum lag of $M = 1$, where we use an "arrow-and-lag" notation e.g., $A \rightarrow B, 1$ to indicate that variable A in time t affects B in time $t + 1$. Therefore, the test for conditional dependence involves testing conditional independence relationships among a set of $J(M + 1)$ artificial variables, representing each variable j at each potential lag $m \in \{0, \dots, M\}$ where M is the maximum lag included in the model. 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 where A has a simultaneous (lag-0) impact on B, and both A and B exhibit first-order autocorrelation (e.g., Gompertz density dependence):

$$A \rightarrow B, 0$$

$$A \rightarrow A, 1$$

$$B \rightarrow B, 1$$

This implies that B is independent of the preceding A (i.e., path $A \rightarrow B, 1$ is zero) conditional upon lag-2 A (i.e., path $A \rightarrow B, 2$), lag-0 A (i.e., path $A \rightarrow B, 0$), and autoregressive effects of B (i.e., path $B \rightarrow B, 1$). This example therefore illustrates that for a maximum lag of M, we need to include conditioning variables for M times prior to the window of interest. 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 lags, we therefore define a "conditioning interval" with conditioning matrix $\mathbf{A}$. For the case of maximum lag $M = 1$, we have:

$$\mathbf{A} = \begin{bmatrix} \mathbf{P}_0 & 0 & 0 \\ \mathbf{P}_1 & \mathbf{P}_0 & 0 \\ 0 & \mathbf{P}_1 & \mathbf{P}_0 \end{bmatrix}$$

We then define all conditional-independent relationships within that conditioning matrix **A**, in this case using package *ggm*. However, we only keep those that define an independence 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. We then fit DSEM to each conditional independence relationship independently, calculate the p-value for a two-sided Wald test, and combine these using Fisher's formula.

As further complication, we note that DSEM can account for missing data (i.e., $y_{tj} = \text{NA}$). In these instances, we impute missing data from the predictive distribution of random effects (i.e., their precision matrix **H** given available data and fixed effects), and then use these imputed data as "fixed" for each conditional-independence (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 performance, and we show the former in the following (see Table 1 for overview).

Simulation experiment

To explore the likely performance of this time-series d-sep test, we first conduct a factorial simulation experiment. This involves 500 replicates of each combination of the following levels:

1. *Three simulation models*: We define three structural causal models. The simplest "sem" has four variables and only simultaneous effects, where $A \rightarrow B$, $A \rightarrow C$, $B \rightarrow D$, and $C \rightarrow D$. The intermediate involves two variables with simultaneous and lagged effects, where $A \rightarrow B$, and an autoregressive process for both A and B. The most complicated involves four variables,

combining the same simultaneous effects as the "sem" scenario, but also including first-order autocorrelation for each variable;

2. *Three sample sizes*: We simulate time-series of length $T = \{25,50,100\}$, representing short, medium, and long ecological data sets;
3. *Five levels of missing data*: We randomly exclude data for each combination of variable and year, with probability $p_{\text{missing}} = \{0,0.1,0.2,0.35,0.5\}$;
4. *Two estimation models*: For each combination of simulation model, sample size, and missing data, we fit DSEM either using the true model structure, or using a mis-specified structural model (see Fig. 1);

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 d-sep 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 d-sep test will result in a large proportion of p-values that are close to zero when the estimation model does not match the estimation 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. 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 climate*: In a new example of DSEM, we use published data representing spawning phenology for walleye pollock in the Gulf of Alaska from 1983-2023 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 log-ratio Q 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., 2023). 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 ("regression for availability") views variables as independent predictors of survey availability (i.e., $T \rightarrow Q$, $P \rightarrow Q$, and $A \rightarrow Q$). The third ("phenology as mediating effect", 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.

Results

Simulation experiment

We first illustrate the performance (i.e., calibration and efficiency) of the d-sep 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 d-sep 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 correct model remains well calibrated across sample sizes, but that the incorrect model only detects the mis-specification (i.e., a p-value < 0.1) in about 60% of the replicates at low sample sizes ($T = 25$), about 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_{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_{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 d-sep to validate 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 d-sep test then provides strong evidence ($p < 0.01$) that the "bottom-up" model is incorrect, provides weak evidence ($p = 0.15$) that the model with only density dependence is incorrect, and similar weight-of-evidence for the remaining models. We therefore use AIC to conclude that the model with top-down interactions is both validated and parsimonious relative to the model with both interactions ($\Delta AIC = 1.1$). In the case study involving spawning phenology and survey availability for pollock in the Gulf of Alaska (Fig. 6), we explored three models representing "temperature as driver", "regression for availability" or "phenology as mediating effect" hypotheses. The d-sep test provides strong evidence ($p < 0.01$) against the validity of the first two models, but fails to reject the third model ($p = 0.7$). We therefore conclude that this is the most appropriate interpretation of those data given proposed hypotheses.

Discussion

Conditional independence (d-sep) testing is an established practice in structural equation models and phylogenetic path analysis, and we provide a novel extension to time-series models that include simultaneous and lagged interactions among variables. Our simulation experiment confirms that the test is well calibrated, and that short time series ($T = 25$) can be sufficient for simple structural models but that longer time series ($T = 100$) are required as model complexity increases. Similarly, the model is statistically efficient, but this efficiency drops as the proportion of missing data increases towards $p_{\text{missing}} = 0.5$. Finally, the case studies illustrate that d-sep will in some cases retain several candidate models (i.e., for the Isle Royale data set), such that model parsimony and multi-model averaging might be appropriate in these cases. In other cases (e.g., involving pollock spawning phenology), the d-sep test provides quantitative support for the ecological interpretation observational data.

Despite this progress in developing d-sep for time-series models, we have restricted ourselves to scenarios involving 2-4 variables with simultaneous and first-order lags. We do this because the limits of d-sep are already evident at this small model size. For example, using 4-variables with 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 d-sep test seems to lose power rapidly for the sample sizes that are common when analyzing annualized dynamics. Presumably this loss of statistical power is why previous simulation tests of d-sep (e.g., in phylogenetic path analysis) have involved systems with < 5 variables (von Hardenberg & Gonzalez-Voyer, 2013). To address this limit, we therefore envision that analysts may choose to do some form of dimension reduction (e.g., dynamic factor analysis) on sets of variables to identify a reduced set of composite variables, and testing the SCM validity for that reduced set of variables. This procedure ultimately “masks” any concern about causal inter-relationships among variables that are being combined in a single composite variable, and therefore focuses statistical power on the remaining relationships of scientific interest.

We also note that d-sep is only test 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.

In summary, we recommend that d-separation be routinely tested for time-series models when they are intended as structural causal models. When developing an SCM, 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 refine methods for causal validation in ecological systems, including methods to integrate experimental and observation studies. We hope that these validation methods will help to unleash the potential for SCM in ecological systems.

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 via GitHub (https://github.com/James-Thorson-NOAA/dsep_in_dsem).

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

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

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

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., Adams, Grant D., Ferriss, B. E., Shotwell, S. Kalei, McKelvey, D.R., & McGowan, David W. (2023). *Assessment of the walleye pollock stock in the Gulf of Alaska* (Stock Assessment and Fishery Evaluation Report for Groundfish Resources of the Gulf of Alaska). North Pacific Fishery Management Council. https://apps-afsc.fisheries.noaa.gov/Plan_Team/2021/GOApollock.pdf

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/

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

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

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. https://doi.org/10.1126/science.1227079

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 initialization buffer for maximum number of lags
2	Define conditional independence relationships	Use “d-separation” to define the set of conditional independence (CI) relationships
3	Eliminate relationships	Filter CI relationships to eliminate duplicates, and restricting target and predictor variable to outside the initialization buffer (while allowing conditioning variables within the buffer)
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, recording the p-value for each individual CI test, and combining them using Fisher’s formula

Fig. 1: The structural causal model (SCM) used to simulate data (left column) in three simulation scenarios (rows), and the SCM that is specified when intentionally fitting with a mismatched SCM (right column). In each SCM, 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.

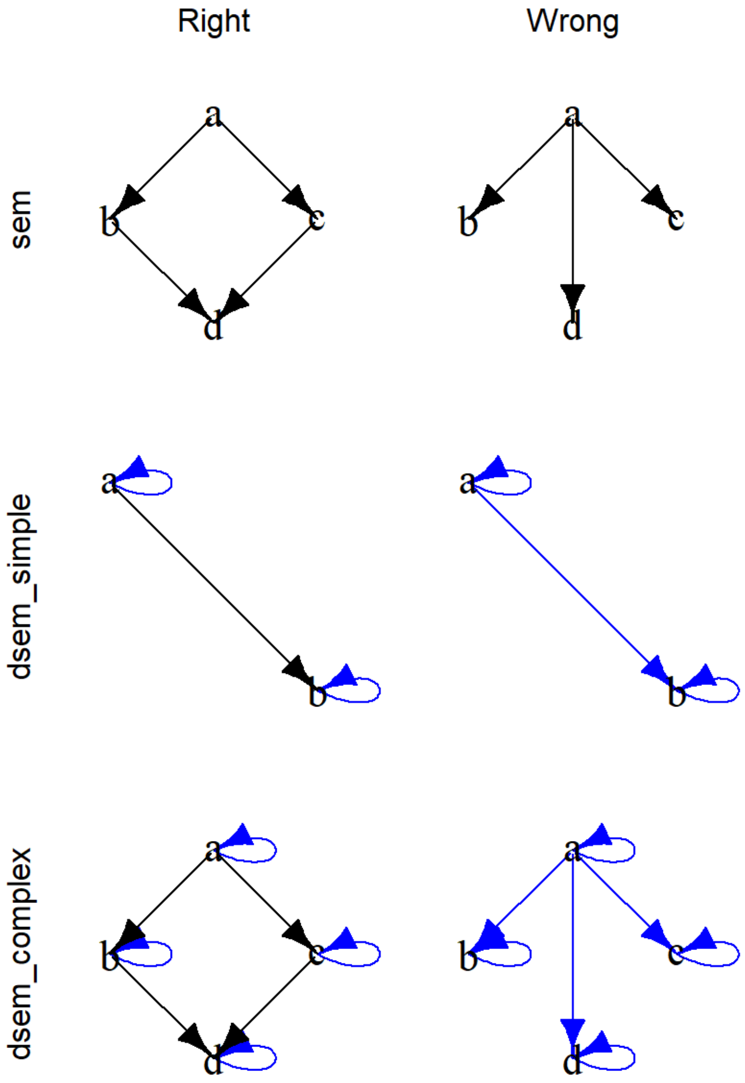
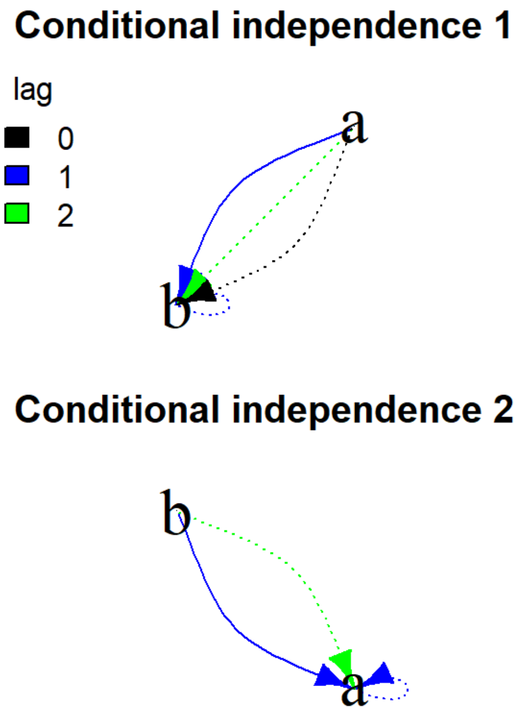


Fig 2: A visual depiction of the two conditional-independence (CI) relationships implied by the "dsem_simple" structural causal model SCM (e.g., 2nd row left column of Fig. 1), as calculated using conditioning matrix **A** (see Methods for structure). The CI relationship is shown with a solid line, while the conditioning variables are shown as dashed lines. Given a time-series SCM 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 $b = \beta_0 \text{lag}(a, 1) + \beta_1 a + \beta_2 \text{lag}(a, 2) + \epsilon$ where we then test for the significance of the β_0 coefficient.


Conditional independence 1

Conditional independence 2

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) and using three structural causal models SCM (rows, see Fig. 1 left column), and then refitting those simulated data with either the correct SCM (red histogram, Fig. 1 left column) or wrong SCM (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.

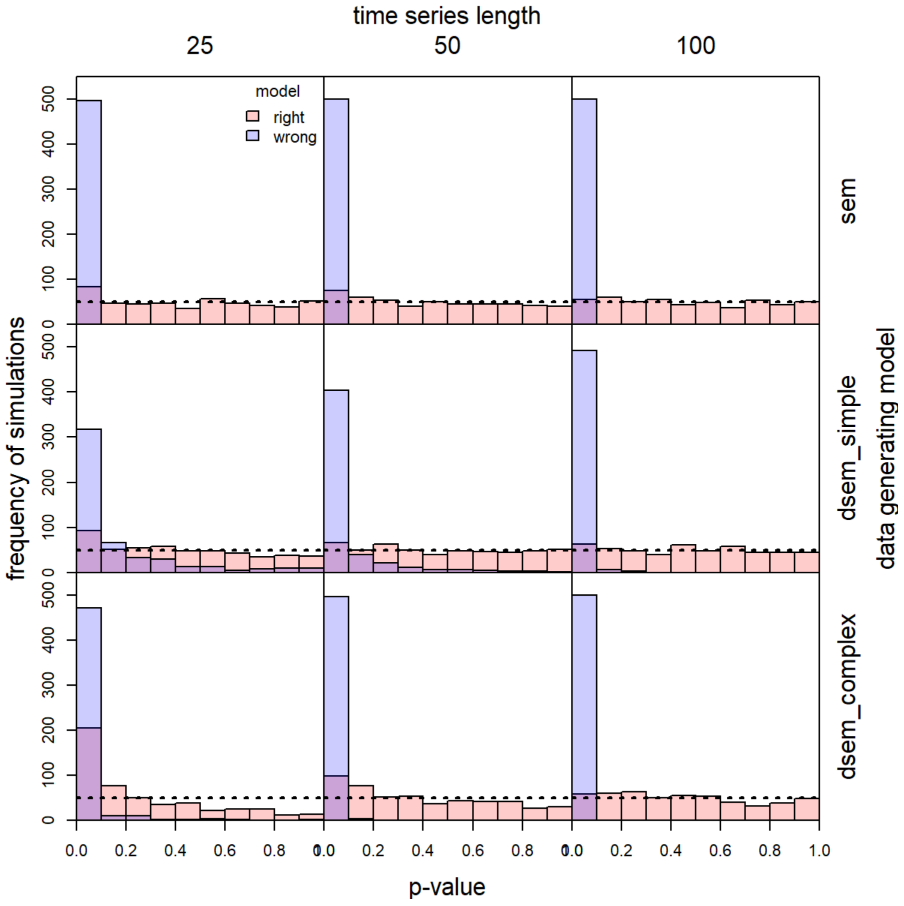


Fig. 4: Results from the simulation experiment when showing the proportion of simulation replicates with d-separation test resulting in $p < 0.1$ (y-axis) across five proportions of missing data $p_{\text{missing}} = \{0, 0.1, 0.2, 0.35, 0.5\}$ (x-axis), and across different time-series lengths (columns)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

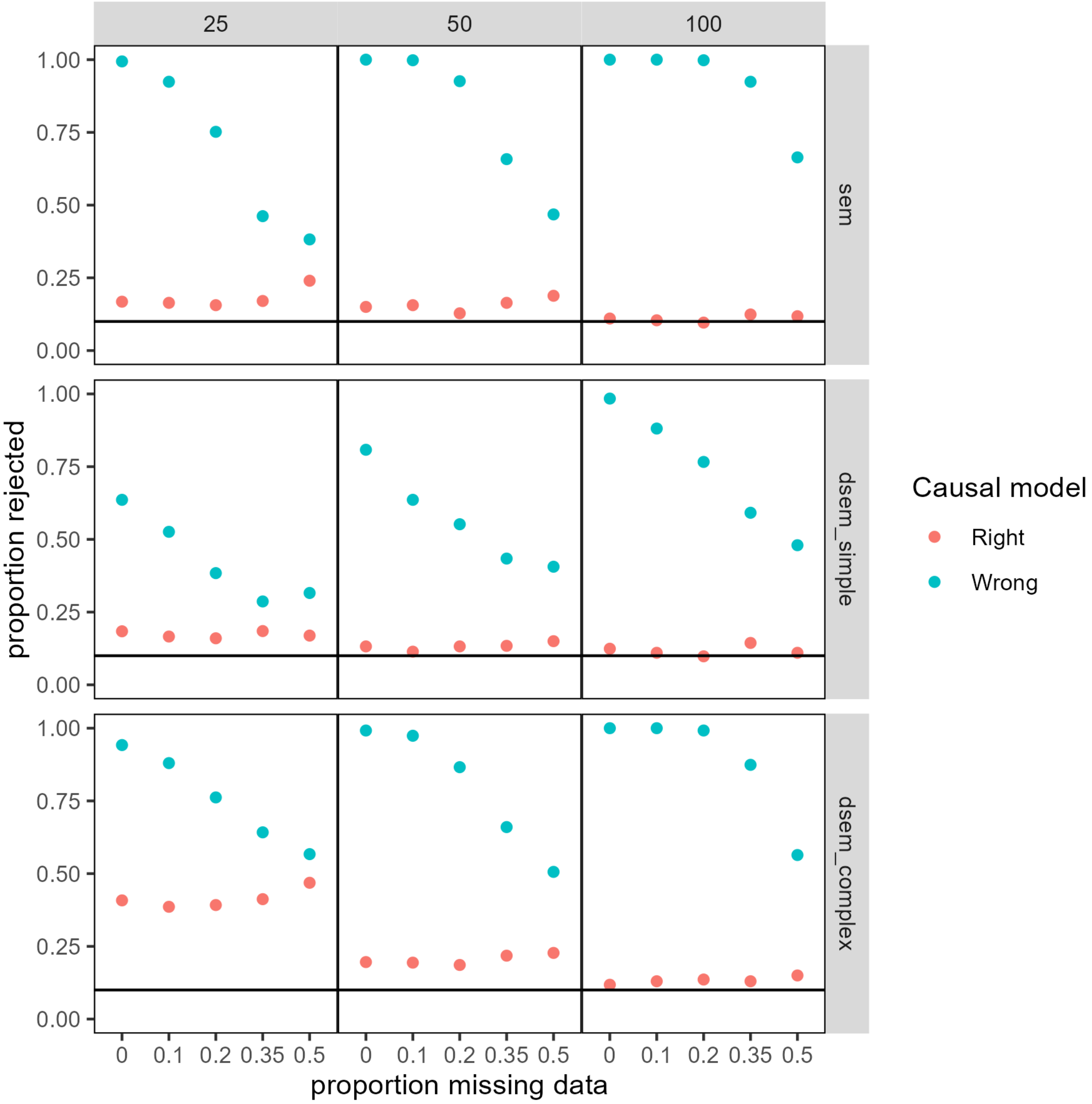


Fig. 5 – Estimated SCM showing a vector-autoregressive model fitting to data for wolf (W) and moose (M) log-abundance in Isle Royale 1959-2019 (Vucetich & Peterson, 2012). We compare a model with just Gompertz density dependence (i.e., $W \to W, 1$ and $M \to M, 1$), adding either bottom-up or top-down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

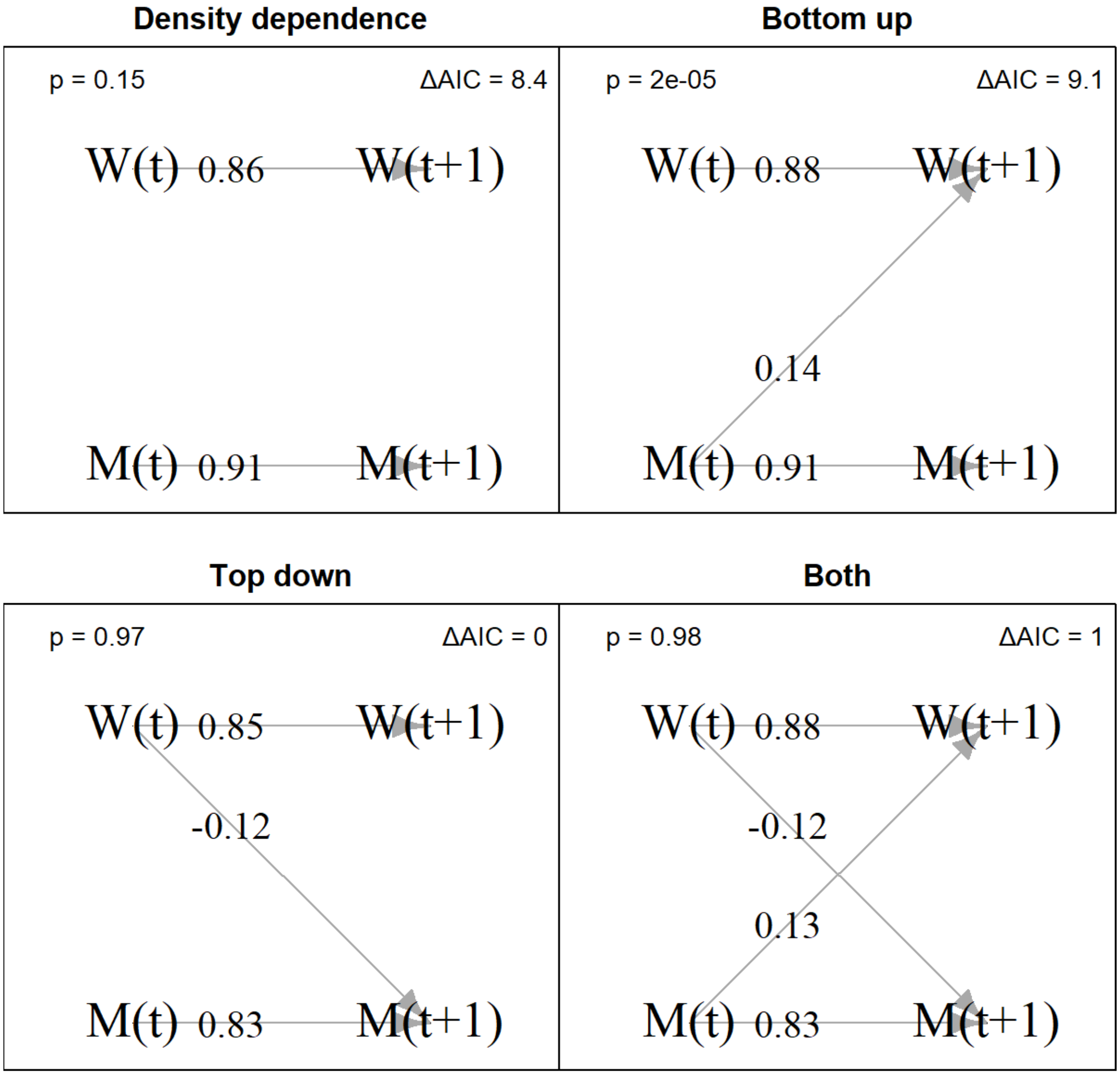
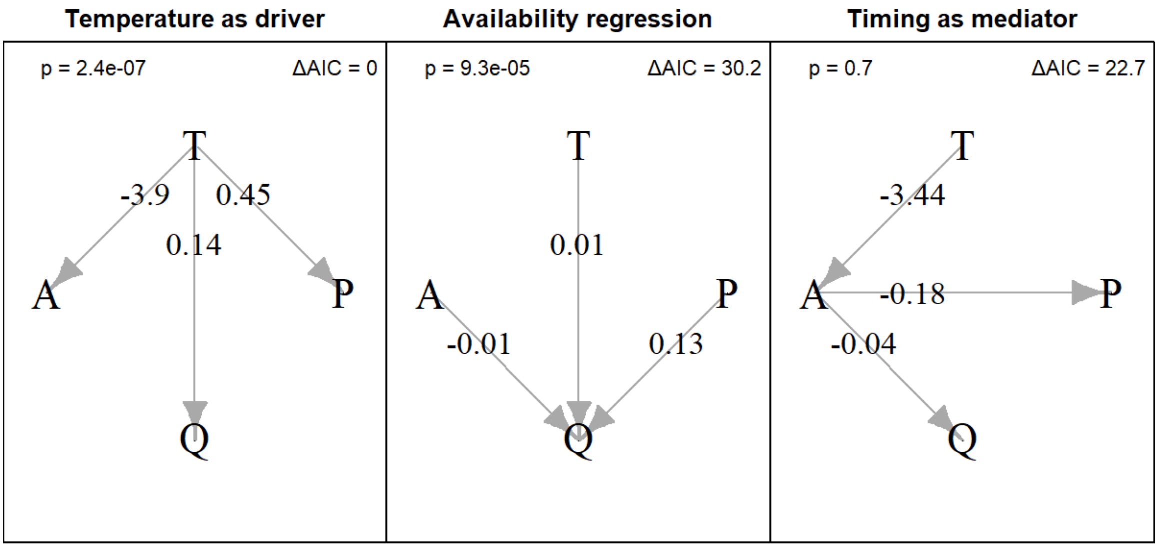


Fig. 6: Estimated SCM showing the estimated path coefficient between temperature T, the average number of days between mean date of spawning and the mean date of a survey on spawning grounds A, the logit-transformed proportion of females >30cm in a spawning or spent stage during the spawning-grounds survey P, and the log-ratio between the surveyed biomass and predicted biomass given other data Q. We show three SCMs (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.


Temperature as driver
p = 2.4e-07
ΔAIC = 0
T
-3.9
0.45
0.14
A
P
Q
Availability regression
p = 9.3e-05
ΔAIC = 30.2
T
0.01
A
P
-0.01
0.13
Q
Timing as mediator
p = 0.7
ΔAIC = 22.7
T
-3.44
A
-0.18
P
-0.04
Q