Inferring species interactions using Granger causality and convergent cross mapping

Frédéric Barraquand[1,2,*], Coralie Picoche[2], Matteo Detto[3], and Florian Hartig[4]

[1]CNRS, Institute of Mathematics of Bordeaux, France
[2]University of Bordeaux, Integrative and Theoretical Ecology, LabEx COTE
[3]Princeton University, Department of Ecology and Evolutionary Biology
[4]University of Regensburg, Theoretical Ecology

Abstract

1. Identifying directed interactions between species from ~~multiple time series of population sizes~~ is a statistical problem with numerous applications to ecology, ranging from the inference of trophic dependencies to evaluating coexistence mechanisms. Statistically, the problem is known as causal inference in time series ~~analysis.~~ The causality concept usually applied is that of Granger: a time series x Granger-causes y if adding x to a dynamical model ~~for y~~ helps predicting future values of y.

2. Standard approaches for detecting Granger causality (GC), notably Multivariate AutoRegressive models (MAR) of finite lag order p, rely on linearity assumptions. This has led to ~~a~~ skepticism in the ecological literature about the ability of (log)-linear MAR(p) models to infer causal links in nonlinear dynamical systems, and to the recommendation to use nonlinear approaches such as convergent cross-mapping (CCM) instead. Given that simple MAR(1) models have been shown to be robust to non-linearities with regard to interaction sign inference, however, it is unclear ~~when to prefer one approach over the other~~.

3. Here, we show that MAR(p) modelling, contrary to general belief, can be used to infer causal links between interacting populations with nonlinear dynamics, when appropriate log-density scaling and model selection based on information criteria are used. In our tests, Granger causality and CCM were able to uncover interactions with surprisingly similar performance, for highly nonlinear systems including real predator-prey cycles, deterministic chaos, 2-species stochastic competition models with and without forcing, as well 10- and 20-species interaction networks (including chaotic ones). Forcing species dynamics by an abiotic driver created some false positives for both methods, but not more than 25%.

4. Our results show that Granger causality, even in its simplest log-linear MAR(p) formulation, is a valid method for inferring interactions in ecological networks, including in highly nonlinear cases. We further explain why CCM and Granger causality can yield such similar results in spite of different mathematical formulations, and highlight nonlinear and high-dimensional extensions of Granger causality that could be of use to ecologists.

Keywords: time series, interaction network, interaction strength, causal inference, feedback, food web, community dynamics.

Introduction

~~Measuring~~ the ~~strength of linkages~~ between different species' population dynamics is a statistical endeavour with ~~multiple~~ implications for ~~ecological fields as varied as~~ coexistence ~~theory~~ (where the ratio between intra and interspecific competition acts as a measure of niche differentiation, Adler *et al.* 2010, 2018), ~~the study of~~ food webs (Berlow *et al.*, 2004; Wootton & Emmerson, 2005), ~~ecosystem-based fisheries management~~ (Link, 2002; Pikitch *et al.*, 2004), ~~and conservation.~~ ~~Statistically d~~etecting such ~~dependencies~~ using ~~correlative~~ approaches, however, can be extremely challenging ~~and fraught with biases~~ (Carr *et al.*, 2019; Coenen & Weitz, 2018), irrespective of sample sizes. ~~Indeed,~~ spatial or temporal co-occurrence (Cazelles *et al.*, 2016) or co-abundance patterns (Stone & Roberts, 1991; Loreau & de Mazancourt, 2008) ~~in fact~~ do not ~~directly map to~~ interactions between species. For instance, strongly competitive communities usually show a large amount of positive associations between species not only because abiotic forcing makes synchrony the general rule (Loreau & de Mazancourt, 2008), but also because indirect interactions make the enemy of my enemy a friend (Stone & Roberts, 1991). In order to discuss how to infer dependencies between species' population dynamics, we must therefore ground this question in a dynamic ecological and statistical theory.

~~More specifically, we must~~ first agree on a definition of an interaction (Berlow *et al.*, 2004). For the purpose of this paper, two species i and j are deemed to interact if species i's population growth rate is affected by the population density of species j or vice-versa. This definition maps well to theoretical ecology, where communities are modelled as variations of the generalized Lotka-Volterra equations (e.g., May, 1973; Yodzis, 1998; Coyte *et al.*, 2015; eq. 1):

$$\frac{dN_i}{dt} = r_i N_i + \sum_{j=1}^{S} g_{ij}(N_i, N_j) N_j \tag{1}$$

Interestingly, this definition ~~of interactions~~ also matches with ~~that of~~ statistical time series models (Ives *et al.*, 2003; Mutshinda *et al.*, 2009, 2011; Hampton *et al.*, 2013). ~~Embracing~~ that ~~all~~ ecological systems are inevitably stochastic, an interaction ~~between species~~ can therefore be defined as a link from species j density to species i per capita growth rate in a stochastic dynamical system ~~describing the whole community dynamics~~. This has also been referred to as local dependence (Schweder, 1970), dynamic causation (Aalen, 1987; Aalen *et al.*, 2012; Sugihara *et al.*, 2012), and Granger-Wiener causality (Granger, 1969; Geweke, 1982) in the more ~~technical,~~ statistically-orientated literature.

To ~~quantify~~ such dynamic causation ~~between species using~~ time series data, ecologists have used a ~~number~~ of statistical ~~models ranging from mechanistic to purely phenomenological,~~ most notably Multivariate AutoRegressive models of order one, or MAR(1) models (also called VAR(1) - vector autoregressive models in the statistical and neuroscience literatures). These are statistical multispecies generalisations of the Gompertz discrete-time single-species models (Ives *et al.*, 2003; Mutshinda *et al.*, 2009). MAR(1) can infer directional linkages between species population dynamics (Hampton *et al.*, 2013; Certain *et al.*, 2018). MAR(p) models, ~~with a maximum time lag of order $p \geq 1$,~~ generalise the MAR(1) framework familiar to ecologists and map more directly to the celebrated Granger-Wiener causality concept (Granger, 1969; Sims, 1980; Ding *et al.*, 2006; Chen *et al.*, 2006; Barnett *et al.*, 2009; Detto *et al.*, 2012; ~~Sugihara *et al.*, 2012;~~ Barnett & Seth, 2014). Granger-Wiener causality (usually referred to as Granger causality or GC for short) is strongly tied to the physical notion that the cause must precede the effect. ~~The fact that the temporal order of events is useful to infer the direction of causality corresponds to~~ the intuition of many biologists (Mayr, 1961) and especially ecologists, familiar with predators lagging behind their prey population dynamics (May, 1973). Granger causality uses both the ideas of temporal precedence and of prediction. If a dynamical model for time series y has its in-sample predictive ability of future y values improved by inclusion of time series x in the predictors, we say that x Granger-causes y. This ~~is a purely operational~~ definition of causality~~, yet it~~ is rather general and does not specify any particular model framework. It can in principle be applied to phenomenological and mechanistic frameworks alike, as well as in nonparametric and spectral settings (Detto *et al.*, 2012). However, parametric MAR(p) models, for which confidence intervals for coefficients, model selection, and other inferential tools are well understood are often preferred for Granger causality testing (Lütkepohl, 2005). Ecologists have in fact been using the Granger causality concept implicitly many times in the form of MAR(1) models (reviewed in Hampton *et al.*, 2013).

In the last decade, new methods such as Convergent Cross-Mapping (CCM; Sugihara *et al.*, 2012), using nonlinear dynamical systems theory and attractor reconstruction, have been introduced to infer interactions between species, and are believed to alleviate potential issues of linear autoregressive models. Although it can seem intuitive that linear autoregressive models will have difficulties with nonlinear time series, it should be noted that MAR(p) models are usually applied to log(population sizes). In an ecological setting, we use indeed mainly autoregressive model that are linear on the log-scale (Ives *et al.*, 2003), because this (i) allows to transform the log-normal distribution of abundance usually found in data into a normal one and (ii) transforms multiplicative growth processes into an additive model structure. MAR(p) models on the log-scale are therefore essentially power-law models when transformed back into the original scale, a flexible way to model monotonic nonlinearities. ~~This suggests that although nonlinear reconstruction methods may perform well for nonlinear dynamical systems, MAR(p) models, when applied with care, could still be used as an approximation~~ (Ives *et al.*, 2003).

In their introduction to convergent cross-mapping, Sugihara *et al.* (2012) strongly criticized instead the application of Granger causality concept to nonlinear dynamical systems. They deemed GC best suited for linear systems dominated by stochasticity, and unfit to model strongly nonlinear dynamical systems with a highly nonlinear (chaotic) deterministic skeleton. This viewpoint has been subsequently adopted by many ecology studies using CCM (e.g., Ye *et al.*, 2015; Ye & Sugihara, 2016; Deyle *et al.*, 2013; Mønster *et al.*, 2017; Harford *et al.*, 2017; Grziwotz *et al.*, 2018). However, while it is correct that the information contained in deterministic dynamical systems cannot be ascribed to a single component of the dynamical system (referred to as "nonseparability" by Sugihara *et al.*, 2012), which is true for nonlinear and linear dynamical systems alike (Granger, 1969; Runge, 2014), the addition of process noise may in fact allow separating the predictive abilities of y vs (x, y) (Runge, 2014, p. 19). Given that ~~there is almost always~~ both nonlinearity and process noise in ecology, Granger causality may in fact be useful for ecological time series.

Modelling has indeed shown that even the simplest MAR(1) models can be surprisingly robust to nonlinearities (Ives, 1995; Certain *et al.*, 2018), uncovering correctly the sign of interactions in the case of stochastic nonlinear competition with a fixed point and multiple predator-prey systems, including limit cycles. Another recent study found that CCM performed similarly to linear Granger causality for a large array of 2-species simulation models (Krakovská *et al.*, 2018), including some chaotic ones. Further evidence that linear GC can in fact be useful comes from studies that used linear GC, nonlinear GC, and CCM, and found consistent causal answers with all three (Hannisdal *et al.*, 2017; Hannisdal & Liow, 2018). These studies provide hints that Granger causality, even in its linear MAR(p) formulation, may therefore apply well to stochastic and nonlinear ecological dynamical systems.

Given the advantages stemming from the great conceptual and practical simplicity of MAR(p) models, there is a need to better understand in which ecological scenarios linear GC can be a good approximation for interaction inference, and in which case more sophisticated techniques are needed, such as CCM or entropy-based methods (e.g., Amblard & Michel, 2013; Hannisdal & Liow, 2018). With new monitoring tools like metabarcoding making community time series increasingly available, GC methods may become even more interesting for ecologists. For example, Granger-causality techniques are currently gaining traction in the rapidly evolving microbiome field that attempts at inferring interactions from metabarcoding data on microorganisms (Gibbons *et al.*, 2017; Mainali *et al.*, 2019; Carr *et al.*, 2019).

In this article, we evaluate the performance of linear MAR(p) models and compare it to CCM on a number of ecological examples for which CCM is currently thought to be more appropriate. We demonstrate that harsh criticism of the Granger causality concept by Sugihara *et al.* (2012) may have been induced by nonstandard model selection and evaluation techniques. Using simpler model selection techniques, routinely used by statisticians ~~(Lütkepohl, 2005)~~, ~~in order~~ to the infer lag order p of MAR(p) models as well as their parametric structure, we show that Granger causality techniques can infer interactions surprisingly well ~~in nonlinear cases~~. We also highlight intriguing parameter configurations and empirical case studies where Granger causality and CCM either both fail to some degree or both work, which suggests that seemingly different causality concepts might in fact share hidden similarities. Throughout our analysis, we take particular care to consider both effect sizes and statistical significance of causal inferences. We then demonstrate that MAR(p) modelling can be scaled up to large interaction networks using either appropriate model regularization techniques (based on a structured version of the LASSO) or pairwise inference with an appropriate false discovery rate correction. Comparison to CCM is provided in the latter case.

Methods and models

In the following, we recall the basics of Granger causality concepts and MAR(p) - Multivariate AutoRegressive of order p - modelling, which is the most common way to assess Granger causality (though by no means the only one, see e.g. Detto *et al.* 2012 for a nonparametric and spectral Granger approach, Barnett & Seth 2014 for parametric and spectral approaches). We describe shortly thereafter convergent cross-mapping (Sugihara *et al.*, 2012), which takes a different approach to causal inference, based on dynamical systems theory and state-space reconstruction. We then describe the real datasets and numerical simulations that will be used for evaluating causal inference methods.

Causality concepts

Granger causality and MAR(p) implementation

Formally, times series $\mathbf{x} = (x_t)_{t\in[|1:T|]}$ Granger-causes time series $\mathbf{y} = (y_t)_{t\in[|1:T|]} \iff$ including x in a time series model for y improves in-sample prediction of y. In the MAR(p) framework, this translates into performing two autoregressive model fits to explain time series y, one with only y and one with both y and x:

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \eta_t,\ \eta_t \sim \mathcal{N}(0,\sigma_\eta^2) \tag{2}$$

$$y_t = \sum_{i=1}^{p} a_{1i} x_{t-i} + \sum_{i=1}^{p} a_{2i} y_{t-i} + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0,\sigma_\epsilon^2) \tag{3}$$

Granger causality is inferred if $\sigma_\epsilon^2 < \sigma_\eta^2$. A simple measure of effect size is therefore the log ratio of the sum of squared residuals $G_{x\to y} = \ln\left(\frac{\sigma_\epsilon^2}{\sigma_\eta^2}\right)$ (Geweke, 1982; Detto *et al.*, 2012).

When more than two variables are considered, pairwise GC has to be differentiated from conditional GC (Geweke, 1984). Conditional GC occurs whenever a third variable z is considered and corrected for. When fitting a MAR(p) model to more than two species, we would typically be interested in conditional GC rather than pairwise GC, with conditional GC correcting for ~~potential confounders~~ (Geweke, 1984; Barnett & Seth, 2014). For instance, let us consider a MAR(1) model (eq. 4) with 3 species, which may be familiar to ecologists through the works of Ives *et al.* (2003); Hampton *et al.* (2013):

$$\mathbf{x}_t = \ln(\mathbf{N}_t),\ \ \mathbf{x}_{t+1} = \mathbf{a} + \mathbf{B}\mathbf{x}_t + \mathbf{C}\mathbf{u}_t + \mathbf{e}_t, \mathbf{e}_t \sim \mathcal{N}_3(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4}$$

~~so that its~~ **B** ~~(interaction) matrix is defined by~~

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \tag{5}$$

and $\mathbf{C}$ is a matrix representing the effect of environmental covariates (Ives *et al.*, 2003; Hampton *et al.*, 2013). Here, whenever b_{12} is significantly different from zero, we have a causal influence $x_2 \rightarrow x_1|(x_3, \mathbf{u})$, that is, an influence of x_2 on x_1 conditional to the population density x_3 of species 3 and all the control environmental variables in the vector $\mathbf{u}$. Using centered data so that the intercept disappears, the MAR(p) model is defined as

$$\mathbf{y}_{t+1} = \sum_{q=1}^{p} \mathbf{B}^{(q)}\mathbf{y}_t + \mathbf{e}_t,\ \mathbf{e}_t \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma}) \tag{6}$$

where d is the number of system components (individual time series).

For a general definition of causal effects, we drop $\mathbf{u}_t$ from eq. 6, as it corresponds to a special case where a subset $\mathbf{u}_t$ of the variables $\mathbf{y}_t = (\mathbf{x}_t, \mathbf{u}_t)'$ has a one-way causal impact (i.e., u_t affects x_{t+1} but not the other way around, which can be specified as well by forcing the $\mathbf{B}^{(q)}$ matrices to contain some zeroes). The condition for an interaction from system component j to system component i given all system component (either species densities or environmental variables) then becomes, in a general MAR(p) setting (according to eq. 6):

$$\exists b_{ij}^{(q)} \neq 0 \Leftrightarrow y_j \rightarrow y_i|(y_1, ..., y_{j-1}, y_{j+1}, ..., y_d) \tag{7}$$

where each time lag is indexed by q.

This formulation highlights the difficulties in assessing conditional GC using a large temporal lag order p and a large number of species: models become very high-dimensional, so that some model reduction must be implemented. Conversely, pairwise GC testing between y_i and y_j is assessed through a bivariate autoregressive model for each (i, j) pair, and therefore uses a considerably lower-dimensional model. Pairwise GC testing requires, however, a false discovery correction to attain meaningful statistical significance (Mukhopadhyay & Chatterjee, 2006).

To ~~implement~~ these concepts in practice, we fitted MAR(p) models using the package `vars` in `R` (version 3.4.4), which uses ordinary least squares for estimation. We mainly used the BIC as a default for lag order selection, although we also considered other information criteria for lag order selection (see below). The presence of Granger causality was assessed by the statistical significance and magnitude of the interaction matrix coefficients, and more directly using parametric significance tests for nested models. For pairwise Granger causality testing, we used the function `grangertest` in the R package `lmtest` which performs a Wald test for nested models, using Benjamini-Hochberg corrections for multiple testing to maintain a constant false discovery rate (Benjamini & Hochberg, 1995). For conditional Granger causality testing, we use the function `causality` in package `vars` which provides F-tests for the nested models. Both tests and implementations provided similar answers when compared.

Convergent-cross mapping

Convergent cross-mapping (CCM) was proposed by (Sugihara *et al.*, 2012) as an alternative nonparametric method to detect dependencies between time series. CCM relies on state-space reconstruction. We assume two time series $\mathbf{x} = (x_t)_{t\in[|1:T|]}$ and $\mathbf{y} = (y_t)_{t\in[|1:T|]}$ as previously. The attractor manifold M_X is constructed as a set of E-dimensional vectors $\tilde{\mathbf{x}}(t) = (x(t), x(t-\tau), x(t-2\tau), ..., x(t-(E-1)\tau), ...)$ for $t = 1+(E-1)\tau$ to $t = T$. E is the embedding dimension, denoting how many time lags one counts back in time. This set of vectors constitutes the reconstructed manifold. We now find the E+1 nearest neighbours of $\tilde{\mathbf{x}}(t)$ in M_X. Their time indices are denoted $t_1, ..., t_{E+1}$. The reconstruction of y_t from M_X proceeds as follows:

$$\hat{y}(t)|M_X = \sum_{i=1}^{E+1} w_i y(t_i)$$

with $w_i = u_i / \sum_{j=1}^{E+1} u_j$, and $u_j = \exp\left(\frac{-d(\tilde{\mathbf{x}}(t),\tilde{\mathbf{x}}(t_j))}{d(\tilde{\mathbf{x}}(t),\tilde{\mathbf{x}}(t_1))}\right)$ where $d(\tilde{\mathbf{x}}(t), \tilde{\mathbf{x}}(t_1))$ is the minimal distance between $\tilde{\mathbf{x}}(t)$ and all other embedded points.

The cross-map skill from X to Y is then measured by the correlation coefficient $\rho(\mathbf{y}, \hat{\mathbf{y}}|M_X) > 0$, which increases with the size L of the library of points used to reconstruct the manifold M_X if Y causes X. The surprising thing here is that predicting Y by M_X is equivalent to Y causing X and not the other way around (Sugihara *et al.*, 2012). Hence, to know if X causes Y, we look at $\rho(\mathbf{x}, \hat{\mathbf{x}}|M_Y)$.

Due to a lack of a parametric model, there is no direct way of computing p-values ~~related to the Convergent Cross-Mapping skill ρ~~, several p-value formulations have been proposed:

- Cobey & Baskerville (2016) suggested $p(X \nrightarrow Y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_i \left(\rho(\mathbf{x}, \hat{\mathbf{x}}|M_{Y,\mathrm{Lmax}}) < \rho(\mathbf{x}, \hat{\mathbf{x}}|M_{Y,\mathrm{Lmin}})\right)$

where n is the number of libraries of size L that were used to build M_Y. $M_{Y,\mathrm{Lmax}}$ (respectively, $M_{Y,\mathrm{Lmin}}$) is the manifold constructed with the maximum (respectively, minimum) library size. Two versions of this p-value can be computed depending on whether one samples with replacement (the bootstrap) for the librairies or without replacement (in which case $M_{Y,\mathrm{Lmin}}$ varies but not $M_{Y,\mathrm{Lmax}}$).

- When two species are forced by a shared forcing driver (e.g., seasonal temperature), spurious causality can emerge. This can be corrected by computing surrogate time series, which keeps the periodicity of the signal but shuffles its residuals, so that cross-correlations containing causal information are "erased". Cross-mapping is then computed on the surrogates and compared to the real value (Deyle *et al.*, 2016). In this case, $p(X \nrightarrow Y) = \frac{1}{n}\sum_{i=n}^{n} \mathbf{1}_i \left(\rho(\mathbf{x}_{\mathrm{real}}, \hat{\mathbf{x}}_{\mathrm{real}}|M_Y) < \rho(\mathbf{x}_{\mathrm{surr}}, \hat{\mathbf{x}}_{\mathrm{surr}}|M_Y)\right)$

- Given the good performance of surrogate-based p-values in the shared abiotic driver case (item above), we computed those for all simulations. In simpler cases where there was no confounding shared abiotic driver, surrogates were only computed by permutation of the time series. This was found to be simpler and more efficient than other techniques to provide significance for CCM (Appendix S2.1).

~~The analyses have been~~ performed using the package `rEDM` (version 0.7.1 Ye *et al.*, 2018). For each time series, we retrieved the best embedding dimension (which maximizes the forecast skill of the simplex) and used it in the cross-mapping function, with 100 different libraries for each library size and maximum library size depending on the length of the time series (300 timesteps if not mentioned otherwise). The libraries were obtained with random draws without replacement from the original time series.

Evaluating GC and CCM

For each case study, we compared the values of classical scores such as recall or sensitivity (fraction of true interactions TP found over the total number of true interactions, $\frac{TP}{TP+FN}$, where FN are false negatives) and the specificity (fraction of true negatives TN over the total number of negatives, $\frac{TN}{TN+FP}$, where FP are false positives).

Additionally, we measured similarity between GC and CCM detected causalities at the level of individual time series, within a single parameter set and model, to see if they detect matching causalities or have some degree of complementary (e.g., CCM detects causality in simulations where GC does not, and vice versa). The similarity measure that we used is the Sokal Michener index I_{SM}, i.e. the number of matches $(11 + 00)$ between GC and CCM over the total number of tests, ranging between 0 and 1.

Granger causality in high-dimensional models

If we have a large number of time series, corresponding to many species, fitting full MAR(p) models (i.e., models that account for all possible interactions without additional constraints) becomes impractical, unless those time series are extremely long (Michailidis & d'Alché Buc, 2013). For d species and p timelags, a $d \times d \times p$ dimensional model needs to be fitted to the data. For instance, 10 species with $p = 2$ yields $2 \times 10 \times 10 = 200$ parameters in the interaction $\mathbf{B}^{(q)}$ matrices only. Even a simpler MAR(1) model would be impossible to fit properly without a set of time series of length above 100 (or some added regularization). Preliminary simulations (Certain *et al.*, 2018; Barraquand *et al.*, 2018) suggest that a nonlinear, stochastic ecological system of dimension 10 or 12 requires approximately times series of length 500 to 800 to be fitted properly without implementing additional constraints.

To deal with high-dimensionality, we considered two solutions:

- Pairwise Granger causality testing with False Discovery Rate (FDR) correction (Benjamini-Hochberg), with a philosophy similar to Mukhopadhyay & Chatterjee (2006). This is done by fitting bivariate MAR(p) models, testing for Granger causality in both directions, and then re-adjusting the p-values obtained through the Benjamini-Hochberg correction.

- LASSO-penalized MAR(1) models with structured penalties, using the R package `SIMoNe` (Chiquet *et al.*, 2008; Charbonnier *et al.*, 2010). This allows to estimate (through non-zero interaction coefficients) conditional Granger causality. A naive idea would be to use the classic LASSO (Least Absolute Shrinkage and Selection Operator, Tibshirani *et al.*, 2015) to set some of the coefficients to zero. Unfortunately, this approach is known to yield substantial bias whenever there is an important structure (here, modular) in the network (Charbonnier *et al.*, 2010). The technique that we use explicitly accounts for network structure in addition to selecting coefficients with the LASSO, and is described in Appendix S1.1.

Simulated and real datasets of interacting species population dynamics

Real data: Veilleux's predator-prey cycles

The two first datasets that we consider are taken from Veilleux (1979) and have been analysed by other authors with mechanistic models that demonstrated two-way coupling (Jost & Ellner, 2000), with plausibly limit cycle behaviour. We additionally created 500 simulated times series from MAR(p) models that best fitted to this dataset, to provide an analogue 'linear' dynamical version of this empirical system.

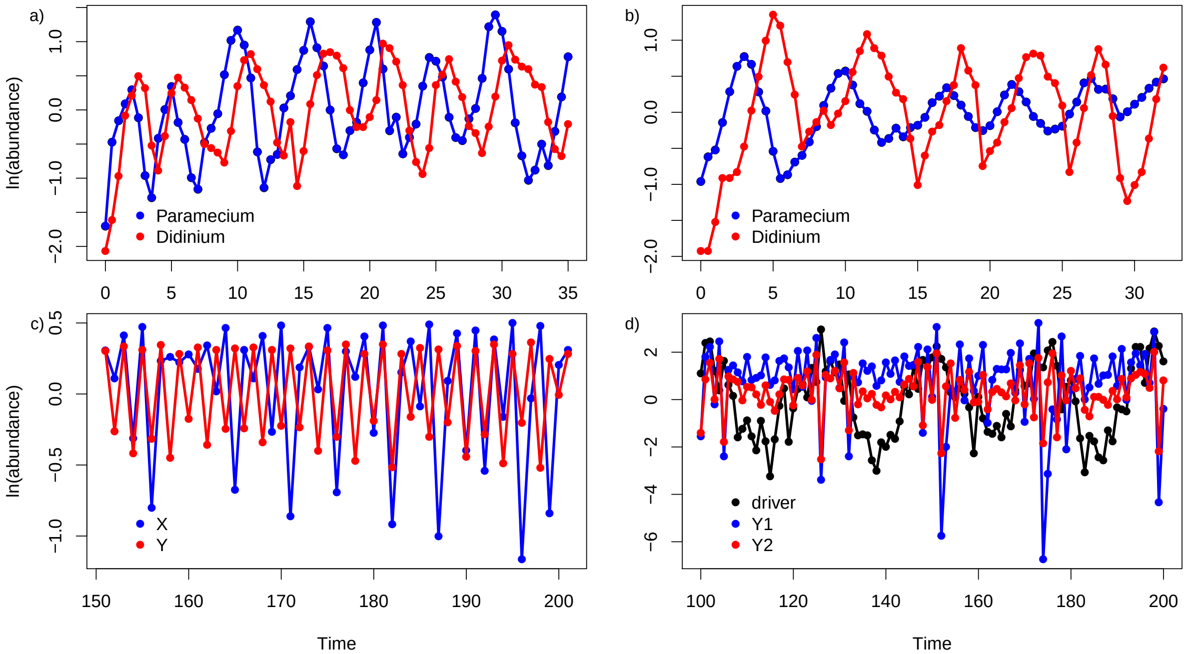

Figure 1: Time series of small-community models. Veilleux's predator-prey data are shown in (a) (dataset CC05) and (b) (dataset CC0375); an example simulation for the 2-species chaotic model is shown in panel (c) and a simulation of the competition model including an environmental driver is illustrated in panel (d).

Deterministic chaos in two-species competition models

Our second case study is the two-species discrete-time logistic competition model used in Sugihara *et al.* (2012) to evaluate the performance of CCM:

$$x_{t+1} = x_t(3.8 - 3.8x_t - 0.02y_t) \tag{8}$$

$$y_{t+1} = y_t(3.5 - 3.5y_t - 0.1x_t) \tag{9}$$

Model parameters are identical to Sugihara *et al.* (2012), which places this model in the chaotic regime (Lyapunov exponent LE = +0.41). This case study therefore constitutes a strong test of the log-linear MAR(p) framework. The only setting we varied compared to Sugihara *et al.* (2012) is the initial condition, which is randomly drawn from a Uniform(0,1) distribution 500 times. Although we acknowledge that "mirage correlations" can occur in some datasets, we aimed at reproducing the full distribution of what this model can provide, as there are no justifications to favour one specific set of initial conditions (outside of illustration purposes). The sample size is taken to be $T = 300$ as in Sugihara *et al.* (2012), with 500 time steps discarded ~~as burn-in.~~

Because a method that finds no interactions whenever absent (i.e., no false positives) is as important as one that finds interactions whenever they are present, we additionally created simulations without interactions:

$$x_{t+1} = x_t(3.8 - 3.8x_t - 0 \times y_t) \tag{10}$$

$$y_{t+1} = y_t(3.5 - 3.5y_t - 0 \times x_t) \tag{11}$$

We evaluate both GC and CCM's ability to find no interactions between these time series.

Two-species stochastic and nonlinear dynamics, including environmental drivers

We consider here a stochastic two-species competition model, with Lotka-Volterra interactions in discrete time and a Ricker type of multispecies density-dependence, which is most commonly used in ecology.

$$N_{1,t+1} = N_{1,t} \exp(3 - 4N_{1,t} - 2N_{2,t} + \epsilon_{1,t}) \tag{12}$$

$$N_{2,t+1} = N_{2,t} \exp(2.1 - 0.31N_{1,t} - 3.1N_{2,t} + \epsilon_{2,t}) \tag{13}$$

This case was already investigated in Certain *et al.* (2018), including as well an environmental driver on species 1 (but not species 2). The model of eqs. 12–13 has a stochastic Lyapunov exponent (SLE) of -0.18, and therefore exhibits a stable fixed point perturbed by noise (Ellner & Turchin, 2005). The stochastic Lyapunov exponent, as elsewhere in this manuscript, was computed following Dennis *et al.* (2001).

As a fourth case study, we create a scenario to investigate the effect of environmental drivers on the estimation of species interactions. This is done with a variant of eqs. 12–13 by adding an environmental driver u_t that has the same effect on both species, which constitutes a challenge for any causal method (u_t is a ~~confounding variable~~)

$$N_{1,t+1} = N_{1,t} \exp(3 + 0.5u_t - 4N_{1,t} - 2N_{2,t} + \epsilon_{1,t}) \tag{14}$$

$$N_{2,t+1} = N_{2,t} \exp(2.1 + 0.5u_t - 0.31N_{1,t} - 3.1N_{2,t} + \epsilon_{2,t}) \tag{15}$$

We consider, as in the deterministic case, the counterparts of the above models where the interspecific interactions are set to zero, i.e.

$$N_{1,t+1} = N_{1,t} \exp(3 + 0.5u_t - 4N_{1,t} - 0 \times N_{2,t} + \epsilon_{1,t}) \tag{16}$$

$$N_{2,t+1} = N_{2,t} \exp(2.1 + 0.5u_t - 0 \times N_{1,t} - 3.1N_{2,t} + \epsilon_{2,t}) \tag{17}$$

We run 500 simulations for each model. The noise is set so that $\epsilon_{i,t} \sim \mathcal{N}(0, \sigma^2)$ i.i.d. with $\sigma^2 = 0.01$.

Ten- and twenty-species interaction webs

We consider a 10 species model which generalises the two-species Ricker competition to more species and more interaction types, with added stochasticity ($\sigma^2 = 0.1$), and therefore represents a considerable challenge to interaction inference, due to the large quantity of potential false positives (many zero interactions) combined to both nonlinear dynamics and stronger stochasticity. The dynamical equation can be written as

$$\mathbf{N}_{t+1} = \mathbf{N}_t \circ \exp(\mathbf{r} + \mathbf{A}\mathbf{N}_t + \mathbf{e}_t), \mathbf{e}_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{18}$$

where $\mathbf{N}$ is the abundance vector, the error $\sigma^2 = 0.1$ and the interaction matrix $\mathbf{A}$ is defined to be

$$A = \begin{pmatrix} -4 & -2 & -0.4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.31 & -3.1 & -0.93 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.636 & 0.636 & -2.12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.111 & -0.111 & 0.131 & -3.8 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & -2 & -2 & -0.4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.31 & -3.1 & -0.93 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.636 & 0.636 & -2.12 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -4 & -2 & -0.4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.31 & -3.1 & -0.93 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.636 & 0.636 & -2.12 \end{pmatrix} \tag{19}$$

This Lotka-Volterra model has a stochastic Lyapunov exponent (SLE) of +0.33. This positive SLE clearly places this model in a noisy chaotic regime (Ellner & Turchin, 2005). In addition, we use the Jacobian matrix of model 18 as the interaction matrix of a MAR(1) model (VAR(1) in statistical parlance), which has therefore comparable interaction strengths but non-chaotic dynamics.

In this case, the dynamical equation is written as

$$\mathbf{x}_{t+1} = \mathbf{J}\mathbf{x}_t \quad (20)$$

$$\text{with } J_{ij} = \delta_{ij} + \alpha_{ij} N_j^* \quad (21)$$

where $\mathbf{x}_t = \mathbf{n}_t - \mathbf{n}^*$, with $n_t = \ln(N_t)$ and $\mathbf{n}^*$ being the equilibrium on log-scale, $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise (see derivation in the Appendices). By definition, such MAR(1) models have a single fixed point forced by stochasticity when stable (Ives *et al.*, 2003): they cannot exhibit chaos and therefore exhibit negative SLEs. We run 25 simulations over 500 time steps with different initial conditions, for both the chaotic LV model and its log-linear MAR(1) counterpart. We slightly modified this model to scale it up to 20 species, with a structure that is still very modular (eq. 29 in SI). For the 20-species model, we also compare Ricker and MAR(1) dynamics for 25 different simulations over 700 time steps. In the 20-species case, coefficients are drawn from a probability distribution (eq. 22 and eq. 23) and therefore differ from one simulation to the next, although we have taken care to avoid coefficients too close to zero by imposing a lower bound:

$$a_{i,j} = \chi_{i,j} \left[a_{min} + (a_{max} - a_{min})\text{Beta}(2,2)\right] \quad (22)$$

with the bounds of the interaction coefficient selected as

$$(a_{min}, a_{max}) = \begin{cases} (0.05, 0.1) & \forall i \neq j, \text{ with probability 0.2 (positive interaction)} \\ (-0.2, -0.1) & \forall i \neq j, \text{ with probability 0.8 (negative interaction)} \\ (-0.8, -0.3) & \forall i = j \end{cases} \quad (23)$$

This construction of the interaction coefficients allows to have some realistically strong dominance of the diagonal coefficients, a certain percentage of weak facilitation (20%), and a lot of competition between species whenever interactions are allowed by the network structure. The 20-species Ricker models thus constructed have SLEs slightly below zero (mean = -0.075, SD = 0.04), and are therefore less "nonlinear" than the 10-species models considered above.

For all datasets, real and simulated alike, the data are log-transformed and centered before analysis. We use a FDR of 20% in all pairwise high-dimensional analyses.

Results

In the following, we report both GC/MAR(p) modelling and CCM results for each dataset or model. The results are then summarized in Fig. 6.

Real data: Veilleux's predator-prey cycles

On those two datasets, both GC and CCM correctly identified the two-way predator-prey coupling. Surprisingly, CCM also identified reciprocal causal influences in the linear MAR(p) approximation.

Model selection of MAR(p) model by all information criteria selected a lag $p = 1$ for the CC05 dataset and a lag of 2 for the CC0.375 dataset (Fig. 2). The p-values for the GC test (null hypothesis: "no GC") and associated effect sizes demonstrate convincingly that the "no GC" hypothesis can be rejected, for both datasets (Table 1).

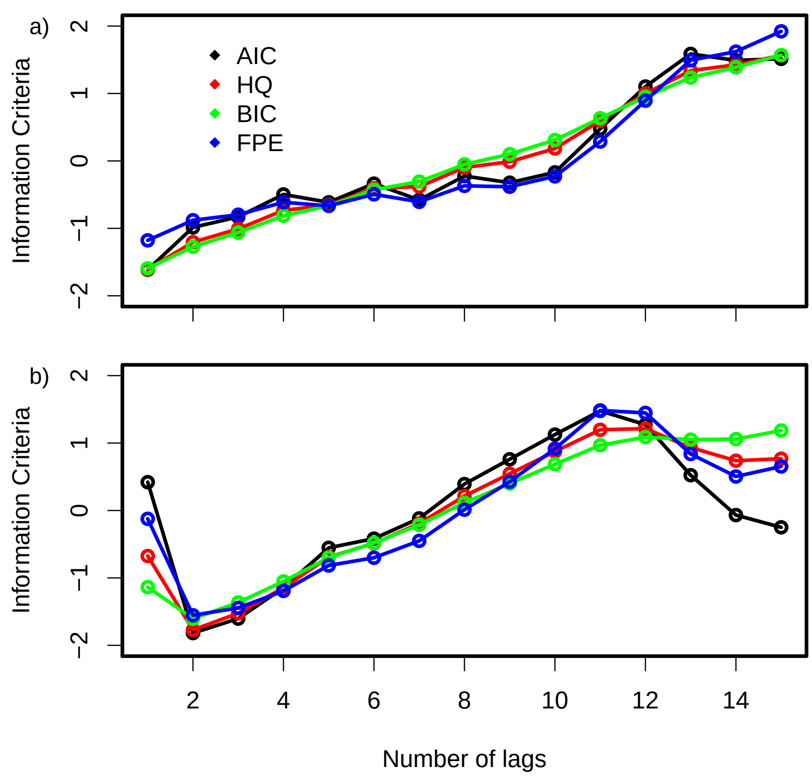

Figure 2: Results of model information criteria vs. lag order for the predator-prey data, for the two datasets. a) CC = 0.5 and b) CC = 0.375

Dataset	CC = 0.5 p-val	$G_{x \to y}$	CC = 0.375 p-val	$G_{x \to y}$
Lag p in VAR(p)	1		2	
$1 \to 2$	2.79×10^{-11}	0.76	0.0409	0.09
$2 \to 1$	1.76×10^{-14}	1.02	0.0464	0.10

Table 1: P-values for H_0: No Granger causality between x and y.

CCM also demonstrates bi-directional causality, as demonstrated by the substantial increase in $\rho(X, \hat{X}|M_Y)$ with library size L in both directions (Fig. 3a and c). This is true for the real data, but also many MAR(1)-simulated dataset using the fitted MAR(1) as the data-generating model (Fig. 3b and d).

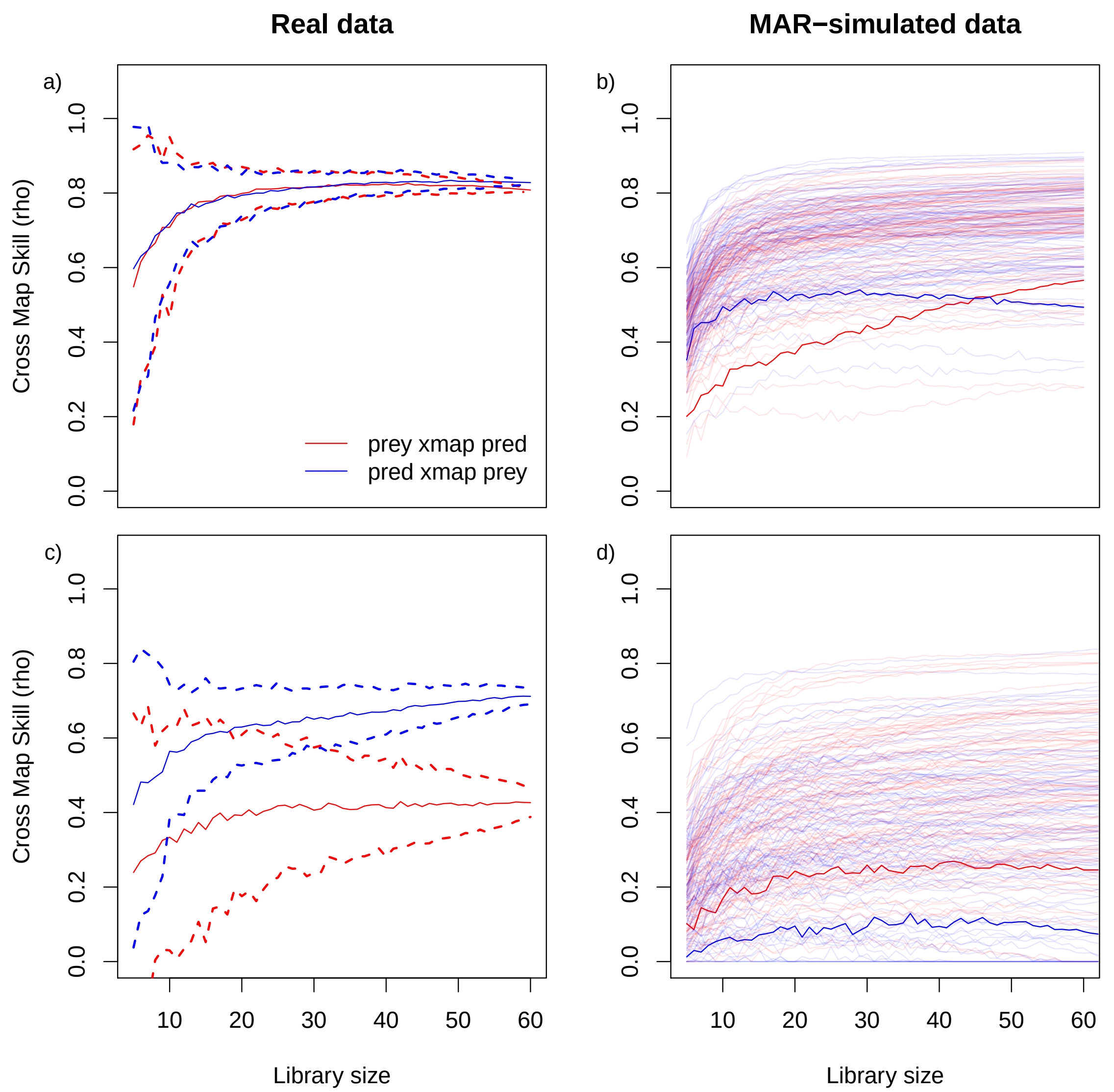

Figure 3: Convergent cross-mapping for Veilleux's CC05 dataset (a and b) and CC0.375 dataset (c and d). Dotted lines represent the confidence bands, obtained by bootstrapping. b) and d) present CCM analyses on data that were simulated using the best-fitting MAR(p) models to the Veilleux datasets.

Deterministic chaos in two-species competition models

In the 2-species chaotic competition model, high-order temporal lags tend to be selected (Supporting Information, Fig. S6) despite the single timelag considered in the simulation model (i.e., higher nonlinearity is expressed as high-order lags). The optimal lag is $p = 7$, for which we report the results in Table 2. Despite this potential overparameterization, the GC tests show that causality is detected for most lag orders (including $p = 7$) whenever causality is present (SI, Fig. S7). Further, the tests are able to reject the null hypothesis of no GC when GC is not present (Table 2, Fig. S7 in Appendix for $p < 7$). GC performs therefore surprisingly well in this chaotic context. CCM performs well when considering a simulation model with interactions, but barely more than GC concerning the weak causal effect $2 \to 1$ (Table 2). ~~Both methods produce some false positives, around 10% when considering p-values, which is to be expected since we test at the 10% level. These rates are somewhat higher in one causal direction for~~ CCM, up to 27% unless all ρ values below 0.2 are discarded. This is because a large number of simulations still show an increase of ρ with the library size L even though there is no causality (SI, Fig. S8). This was likely missed in Sugihara *et al.* (2012) because specific sets of initial conditions were selected, instead of drawing 500 at random as we do here.

Method	Granger	causality		CCM					I_{SM}
Thresholds	pval<0.1	$G_{x\to y}$ >0.04	both	pval<0.1	ρ>0.1	ρ>0.2	both0.1	both0.2	both
With									
$1 \to 2$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
$2 \to 1$	50.6	69.8	50.6	56.6	54.8	29.4	54.2	29.4	0.66
Without									
$1 \to 2$	11.2	100.0	11.2	17.4	10.0	0.6	10.0	0.6	0.83
$2 \to 1$	11.2	2.0	2.0	27.4	27.2	12.2	26.4	12.2	0.72

Table 2: Proportion of simulations with Granger-causality or CCM between x and y over 500 simulations, for the chaotic 2-species competition model. Similarity of causality estimates is indicated by the Sokal-Michener (I_{SM}) index with both conditions taken into account for GC and CCM (p-value and $G_{x\to y} > 0.04$ or $\rho > 0.1$, respectively).

CCM and GC are in general in agreement for specific simulations (i.e., specific initial conditions) corresponding to this model and parameter sets: the I_{SM} similarity index is close to 1, especially in the case when interactions are present.

Two-species stochastic and nonlinear dynamics

Without environmental driver

In our case with 2-species nonlinear competition and noise, we see that GC and CCM perform quite similarly (Table 3), with both methods able to select properly, in most cases, causality and non-causality. CCM has slightly better rates of interactions found (no false negatives), while GC is a little more conservative, especially when considering a threshold for $G_{x\to y}$, the logarithm of the sum of squares ratio (Table 3). Similarity indices for both methods are nonetheless very close to 1, so that they yield essentially similar conclusions when applied to the same time series.

Method	Granger	causality		CCM					I_{SM}
Thresholds	pval<0.1	$G_{x\to y}$>0.04	both	pval<0.1	ρ>0.1	ρ>0.2	both0.1	both0.2	both
With interactions									
$1\to 2$	98.4	94.0	94.0	100.0	100.0	100.0	100.0	100.0	0.94
$2\to 1$	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1.00
Without									
$1\to 2$	12.6	0.2	0.2	12.6	11.0	0.2	10.4	0.2	0.89
$2\to 1$	8.2	0.6	0.6	12.4	10.8	1.4	10.4	1.4	0.89

Table 3: Percentages of simulations with Granger-causality or CCM between x and y over 500 simulations, for the stochastic 2-species competition model without environmental driver, with interactions (top row) and without (bottom row). Similarity of causality estimates is indicated by the Sokal-Michener (I_{SM}) index, with both conditions taken into account for GC and CCM (p-value and $G_{x\to y} > 0.04$ or $\rho > 0.1$, respectively)

The MAR(p) model selected by BIC had a lag of $p = 3$ timesteps, confirming that small lags should be used in such models.

With an environmental driver

The two species model with a shared environmental driver (e.g., temperature) was considerably more complex and yielded less clear cut results than stochastic 2-species competition. Overall, both methods recognize the effect of temperature on the two species growth, with and without interactions (slightly lower score for CCM for species 2, but overall good performance, Fig. S10). Regarding interactions, CCM was better at uncovering interactions that were present in this case, as GC had a good performance for the strong interaction $2 \to 1$ but not the reverse $1 \to 2$. Both GC and CCM had difficulties indicating non-causality (when there were no interactions), and indicated false positives twice above the level of the test ($\approx 20\%$ instead of 10%, Table 4). Thresholding small effect sizes did not solve the issue. Conditional vs pairwise GC had overall similar performance, there was little gain in conditional Granger causality testing in this case.

We used seasonal surrogate time series to assess the significance of CCM, which clearly improved its power to detect interactions, but we still had spurious causalities in CCM when no interactions were present (Fig. S11). This is therefore a scenario where avoiding false causalities is difficult for both GC and CCM – though we should not forget that approximately 75-80% of absent interactions are still being discovered as such.

Method	GC	pairwise		GC	conditional		CCM	seasonal	surrogate	I_{SM}
Thresholds	pval<0.1	$G_{x\to y}$>0.04	both	pval<0.1	$G_{x\to y}$>0.04	both	pval<0.1	ρ>0.2	both	both
With										
$1 \to 2$	24.6	21.0	18.8	24.2	22.2	18.4	98.0	97.0	96.8	0.21
$2 \to 1$	89.8	86.0	85.6	83.8	81.0	79.6	98.6	96.8	96.8	0.76
Without										
$1 \to 2$	26.2	21.2	20.2	26.8	21.6	20.8	27.4	32.6	25.0	0.67
$2 \to 1$	27.0	20.6	19.4	24.8	18.6	17.4	24.0	17.2	17.2	0.73

Table 4: Percentages of simulations with Granger-causality or CCM between x and y over 500 simulations for a model with 2 species and a driver (temperature) Similarity of causality-estimates is indicated by the Sokal-Michener (I_{SM}) index, with both conditions taken into account for conditional GC and CCM (p-value and $G_{x\to y} > 0.04$ or $\rho > 0.2$, respectively). Causalities related to the temperature - not interactions - for CCM are shown in Appendix S2.7.

Larger interaction webs

Here we report the results of analyses for 10- and 20-species modular interaction webs. Lag order selection revealed that low-order MAR(p) models were selected (Fig. S12), with the BIC indicating $p = 1$ as the most parcimonious choice. Hence we have focused on MAR(1) models. The high-dimensional $S \times S$ MAR(1) models include clustering (see Methods and Appendix S1.1) because the basic LASSO-penalized VAR(1) models poorly identify modular interaction webs (Charbonnier *et al.*, 2010). The recall, or true positive rate, that records how many actual interactions are identified as such, is almost always above 60% for the structured LASSO (`SiMoNe`, Chiquet *et al.*, 2008; Charbonnier *et al.*, 2010) and goes up to 80%, which is a relatively good performance. Surprisingly, we found that pairwise (direct) Granger causality testing was even more efficient than the structured LASSO.

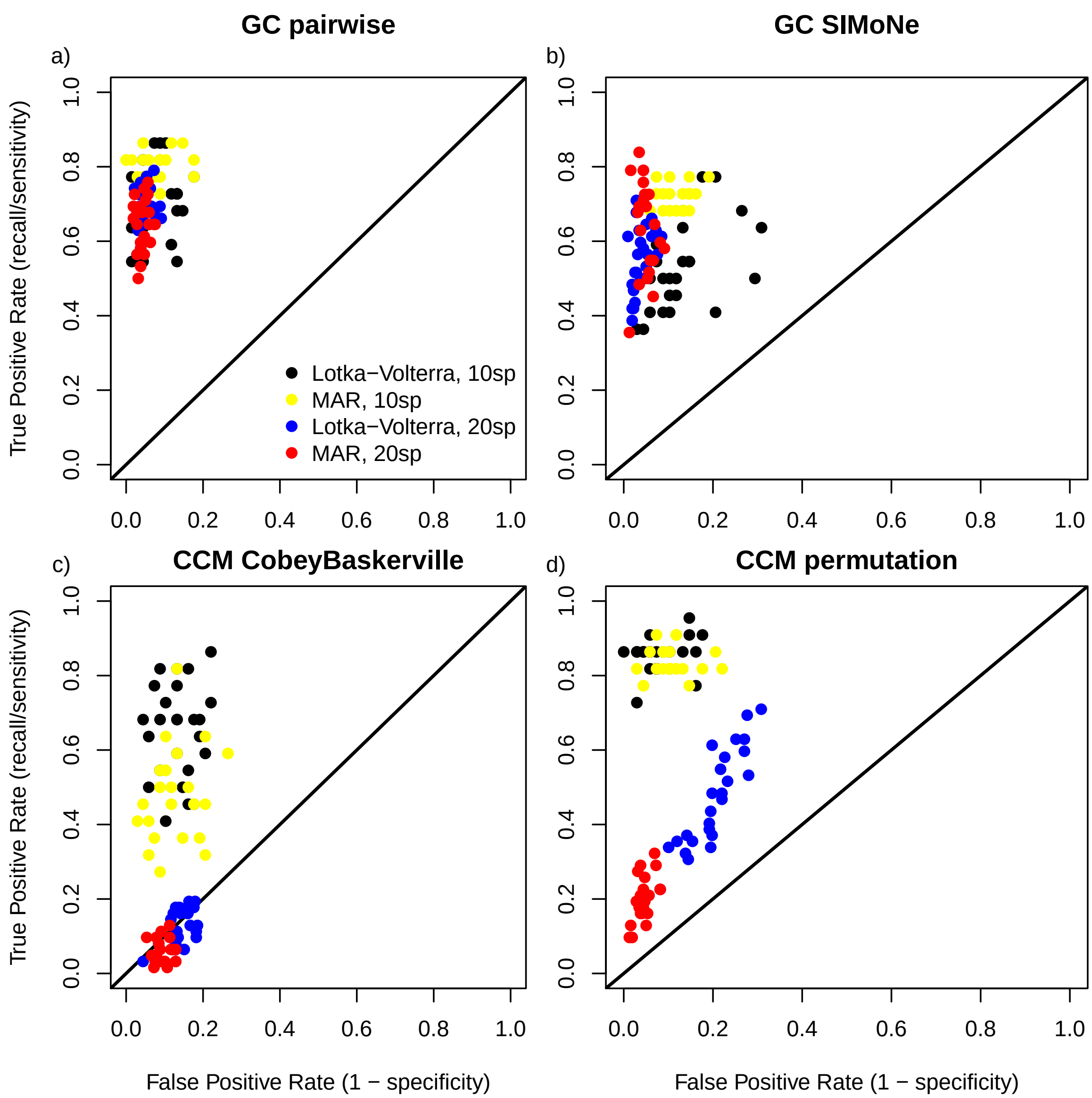

Figure 4: ROC curves for the 10 and 20-species model using Granger Causality (top) and CCM method (bottom), with different ways of computing the causality. For GC, (a) corresponds to results from pairwise GC; (b) is computed with structured-LASSO (SIMoNe). For CCM, we compared p-values computed as (c) $Pr(\rho(L_{min} < \rho(L_{max})$ as suggested by Cobey & Baskerville (2016) to (d) permutation-based p-values. In the 10-species system, one chaotic reference parameter set is considered with many initial conditions, while, in the 20-species model, parameters vary with each simulation. The 20-species model is a perturbed fixed point, with negative SLE. The MAR model is always the MAR(1) model obtained using the Jacobian of the Lotka-Volterra model as an interaction matrix, hence a linearization in log-scale of the Lotka-Volterra model.

Comparing GC and CCM in "ideal" conditions, with the best-performing algorithms for each method (pairwise GC with a Benjamini-Hochberg corrections and CCM with permutation-based p-values) reveals that they reconstruct similar networks for the fairly nonlinear (positive SLE) 10-species case (Fig. 5), both for Lotka-Volterra and equivalent VAR models (where although the dynamics are milder, interactions are still fairly strong since their Jacobian matrices match those of the Lotka-Volterra models).

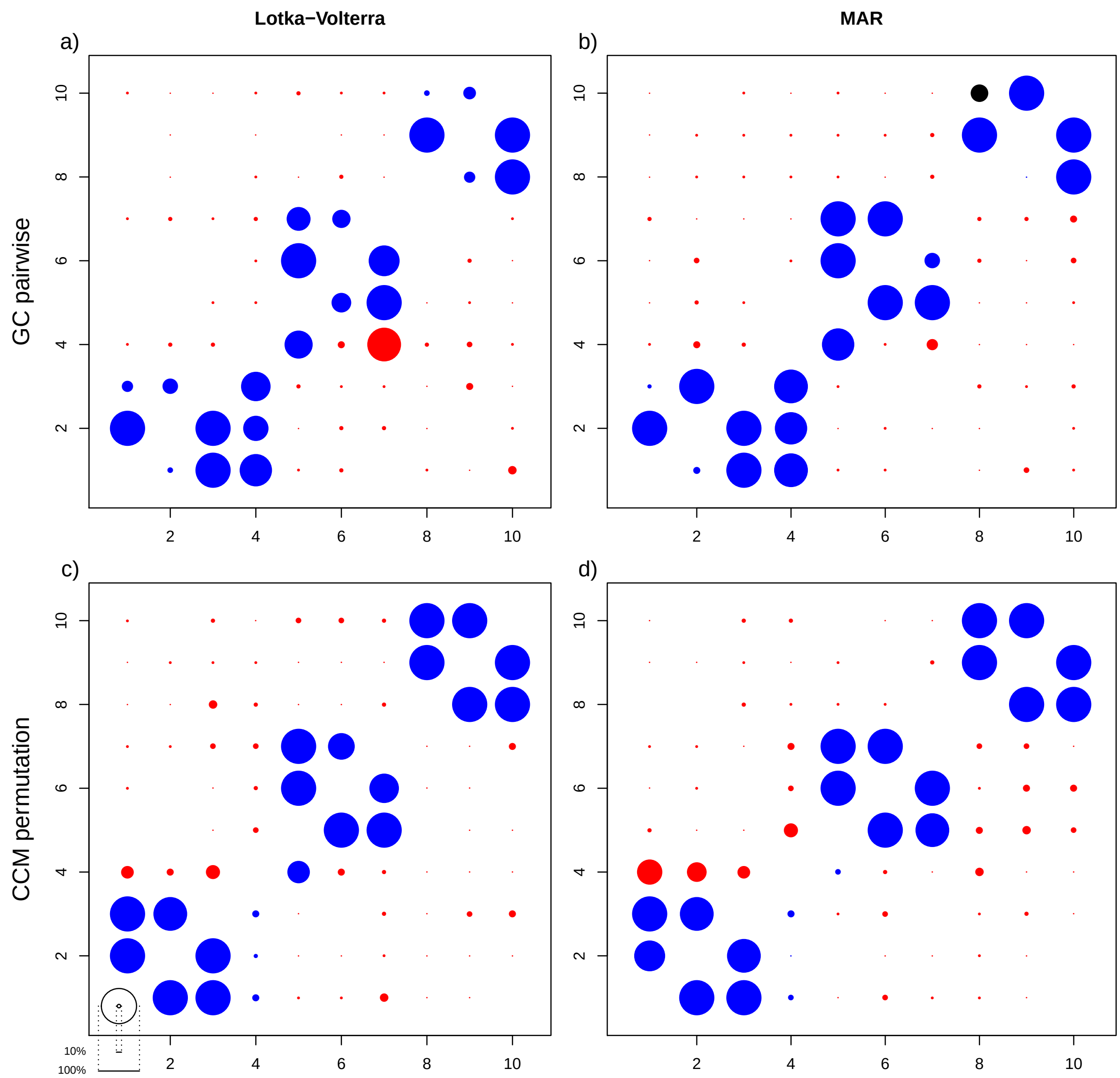

Figure 5: Interaction matrices obtained from pairwise GC (top) or pairwise CCM (bottom) for 10-species communities. Blue circles are the true positives, red circles are false positives and black circles are false negatives. For the true and false positives, the size of the circles is proportional to the proportion of detection over 25 simulations.

However, results with 20-species and a model with slightly weaker interactions tend to make GC the better option since CCM provides quite a number of false positives (Figs. 4 c,d and S15), even with the surrogate-based p-values that worked very well for smaller-dimensional examples. Pairwise GC testing had remarkable performance in this case, and was able to make out both all the modules and the connecting species between them (Fig. S15).

Summary of the results for all simulated case studies

In Fig. 6, we present the recall or sensitivity (fraction of true positives among all positives) and the specificity (fraction of true negatives among all negatives) for all simulated case studies present in this article (also presented as ROC curves for high-dimensional models). The case studies are ranked by descending order of nonlinearity, from the most nonlinear model (chaotic) to the most linear, for both the low-dimensional systems considered (2 species) and high-dimensional systems (10 and 20 species). Both metrics should be close to 1 for model performance to be high; a high recall is important when the objective of a study is to find all interactions present, and a high specificity is paramount when false positives are costly. While we would have expected a increase in performance of GC as the dynamics are less nonlinear and a decrease in performance of CCM, Fig. 6 shows a broad overlap in the performances of both methods, both in high- and low-dimensional systems. GC is a bit handicapped for weak interactions when there is a confounding abiotic driver, and CCM performs worse in weakly nonlinear 20 species systems; but clearly both methods display reasonable performances in most situations. A relatively high specificity, which is a key requirement of any interaction-finding method (otherwise, the method just outputs false positives) is found in all cases.

Discussion

The purpose of this paper was to evaluate the performance of linear and parametric methods for detecting Granger causality (GC) between time series, when ~~simulated according to nonlinear~~ community dynamics, and to compare such performances to a nonparametric and nonlinear popular alternative, convergent cross mapping (CCM). Our main results are that linear GC, implemented using MAR(p) models, is fairly robust to nonlinearities in ecological dynamics, when applied on the appropriate logarithmic-abundance scale and combined with model selection by information criteria. This was true for all considered nonlinear simulation models, including those demonstrating deterministic chaos (Fig. 6). This confirms and extends findings from an investigation of the robustness to nonlinearities of log-linear MAR(1) models (with p restricted to 1 lag, Certain *et al.*, 2018).

Comparison to the CCM framework by Sugihara *et al.* (2012) further revealed that CCM and MAR(p) / Granger causal modelling can in fact - surprisingly - yield relatively similar results in nonlinear and stochastic dynamical systems of interacting species. Evidence for this comes *both* from highly nonlinear systems for which CCM and GC both infer interactions (deterministic chaos, stochastic competition) and from cases where both methods seem to fail to some degree (i.e., two competing species forced similarly by a shared

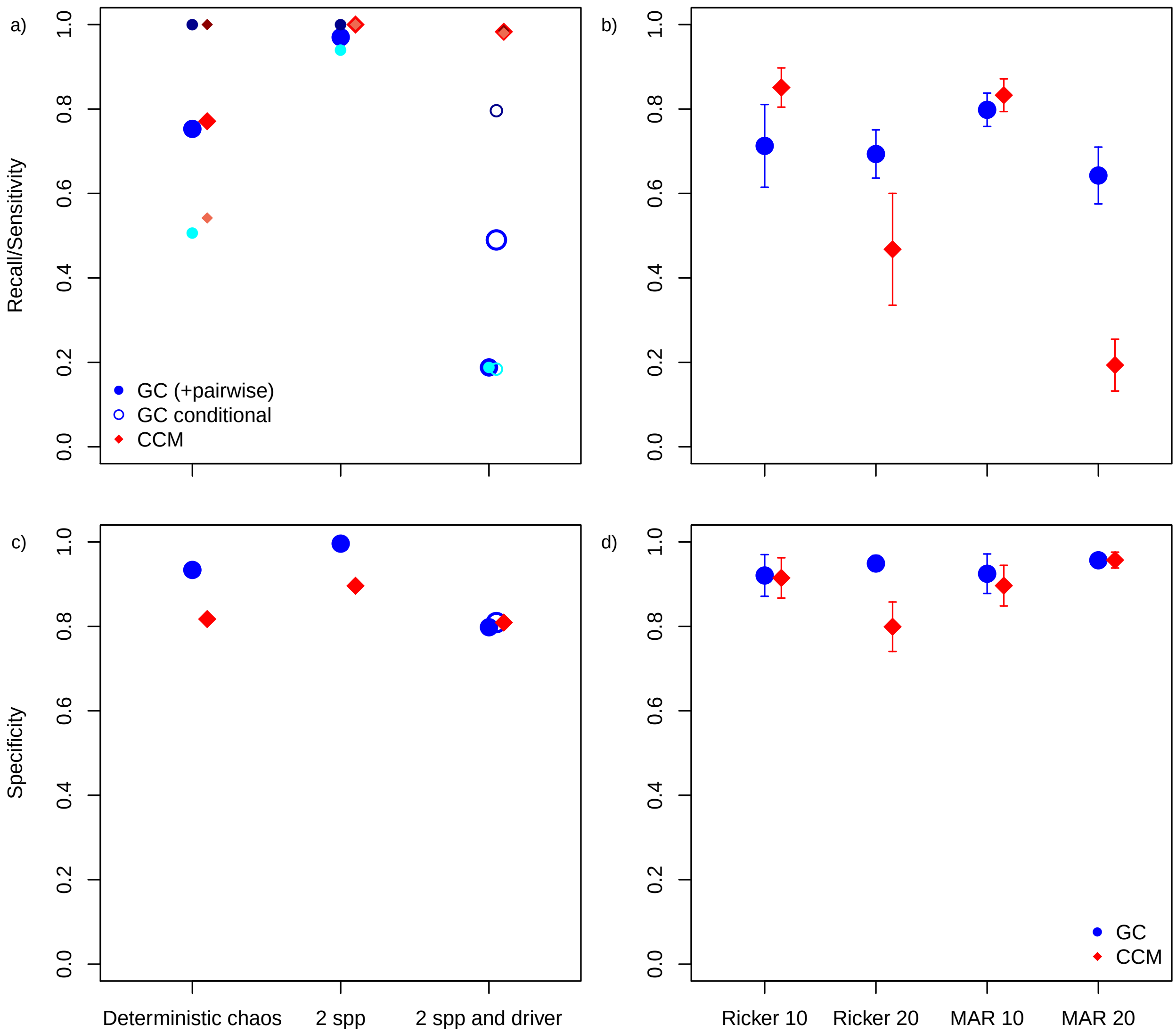

Figure 6: Recall and specificity for Granger-causality (GC, blue dots) and Convergent-Cross Mapping (CCM, red diamonds), ranked from most to least nonlinear dynamics. Causalities are considered only when they are significant at the 0.1 level (small dimension, a and c) or 0.2 level (large dimension, b and d). For small dimensionality (a and c), an additional threshold on effect sizes is considered. Large symbols represent the value of recall and specificity over 500 simulations for two interactions at a time for small communities. In a), the smaller symbols represent the stronger (respectively weaker) interaction in darker (respectively lighter) shade. For the 2spp. and driver simulation, unfilled circles in a) and c) are obtained through conditional Granger causality testing. The smaller circles correspond to the weak interaction (light blue) and the strong interaction (dark blue). For large dimensionality, error bars represent the mean (+/- standard deviation) value of recall and specificity over 25 simulations, since recall and specificity are already computed by comparing true and estimated interaction matrices.

environmental driver, where specificity is always a little lower). Therefore, an important conclusion from our study is that both Granger causality and CCM are able to yield similar inferences on similar datasets.

Moreover, we use here false discovery rate corrections and regularized models (i.e., LASSO-penalized MAR(1) models developed for modular interaction networks, Charbonnier *et al.*, 2010) to tackle relatively-high dimensional models (10 and 20 species). This allows to better infer Granger causality in these contexts that, we surmise, will be most exciting to ecologists working on interacting species using community-level data. The results demonstrate that simple pairwise Granger causality (i.e., using 2×2 MAR(p) models many times with a correction for multiple testing) are as good as the penalized MAR(1) models in finding the interaction network (and surprisingly, sometimes better). We elaborate on these results and possible explanations below.

Can Granger causality be applied to highly nonlinear coupled dynamical systems?

Our results showed that Granger causality, in its log-linear form, is robust to the presence of nonlinearities in the underlying dynamical systems. Nonlinear variants of Granger causality (Marinazzo *et al.*, 2008; Yang *et al.*, 2017; Hannisdal & Liow, 2018) can also be used to infer interactions in nonlinear and stochastic dynamical ecological systems. These conclusions are further supported by the neuroscience literature (Ding *et al.*, 2006; Chen *et al.*, 2006; Barnett & Seth, 2014; Papana *et al.*, 2013; Marinazzo *et al.*, 2008) which, unlike ecology, commonly use Granger causality on nonlinear (and stochastic) dynamical systems.

Ecologist views over Granger causality have likely been shaped by the influential paper of Sugihara *et al.* (2012), who suggested that Granger causality would work well for simulated (log)-linear systems (which they referred to as "stochastic") while CCM would work well for near-deterministic nonlinear dynamical systems. Given the history of both techniques, this makes intuitive sense. Our tests on simulated data revealed, however, that the domains of applicability of both techniques overlap to a great extent. Several differences between our analyses and those performed by Sugihara *et al.* (2012) (in their Supplementary Material) allow to explain this greater overlap, which we develop here.

First, Granger-causality analyses performed by Sugihara *et al.* (2012) on the Veilleux and other datasets rely on a slightly dated model selection procedure (pre-information criteria) which produced overparameterized autoregressive models with very long lags (e.g., $p > 10$). Here, re-analysing the data with a more classic, information-criteria motivated lag order selection, we have shown that in fact GC is perfectly able to find causality in the classic Veilleux *Paramecium-Didinium* predator-prey datasets.

Second, we sampled many chaotic datasets, corresponding to many initial conditions. Although some chaotic datasets may be difficult to identify for GC techniques, these are very few, as MAR(p) models and

GC inference found 95% of true interactions found where CCM finds 100%, in two-species chaotic models. This result was quite unexpected, as we thought that CCM would completely dominate the scores. Thus GC testing can be useful for highly nonlinear systems, and it tends to produce approximately correct rates of false positives when the null hypothesis of no interactions is true (an important aspect as well).

Third, we found that data simulated with log-linear autoregressive models can also be well-identified by CCM, even though CCM relies upon the possibility to reconstruct an attractor in state space. This is further proof of the overlap between the domains of applicability of linear GC and CCM.

How can Granger causality and convergent cross-mapping yield similar inferences, in spite of seemingly opposite assumptions?

Here, we would like to go back to the heart of the issue that Sugihara *et al.* (2012) highlighted, i.e., "causality reversion" in nonlinear dynamical systems. Note that while we offer some suggestions as to how GC and CCM performances can overlap to a large extent, we have no definite answer as to why (mathematically speaking); more research is needed on that point.

The standard Granger causality concept holds that whenever a model $Y_{t+h}|(Y_k, X_k)_{k\in[t-p+1,t]}$ better predicts the observed time series (y_t) than a model $Y_{t+h}|(Y_k)_{k\in[t-p+1,t]}$, then x is causal for y. Most often the time horizon for prediction considered is $h = 1$, which is the perspective adopted here. CCM instead holds that causality flows from x to y whenever $\rho(x, \hat{x}|M_Y)$ increases strongly with the library size L used to reconstruct x from the shadow manifold M_Y. It seems that in the latter method, x causes y whenever knowledge about y can be used to reconstruct x. However, verbal reasoning is treacherous there. To determine whether x causes y:

- GC compares knowledge about Y_t vs. knowledge about X_t, Y_t in prediction of Y_{t+1}
- CCM compares knowledge about M_Y vs. no knowledge about M_Y in prediction of X_t.

There is no direct conditionality upon past X_t values in the prediction step of the algorithm for CCM. Thus there is no causality reversion that is intrinsic to nonlinear dynamic testing: GC and CCM are simply two different types of causal inference that are based upon *different assumptions on the conditioning set* and *ways to select models.*

Finally, GC and CCM both try, as do other approaches based on continuous-time stochastic processes and martingale theory (Aalen *et al.*, 2012; Commenges & Gégout-Petit, 2009), to reconstruct a stochastic dynamical system where interactions are defined as influences of state variables upon the rates of change of the system. No matter how different are their historical origins, GC and CCM are bound to exhibit some similarities because they define interactions in a similar manner.

Issues in calculating p-values and confidence intervals

So far, we mainly discussed the performance of CCM and GC in terms of sensitivity and specificity to detect interactions. Such a continuous, unit-less interpretation of the model outputs dominates benchmarks and tests in the ecology and physics literature (e.g., see recently Krakovská *et al.*, 2018). This implicitly requires setting cut-offs for p-values and effect sizes to decide when an interaction is present. This is possible in a simulation context, but in practice, p-values, Bayes factors, and confidence intervals are the quantities are typically reported. Therefore, a question of interest is: do GC and CCM consistently produce precise p-values and confidence intervals? Our results show that while overall GC and CCM produce sound results, statistical indicators for both methods are not always very well calibrated. This is exemplified by all the Tables in our manuscript in that, in the case where the null hypothesis of no interactions is true, the percentage of p-values below 10%, for both GC and CCM, do not always match exactly the 10% level of the test employed.

In the Granger case, this is easily explained by the fact that the model that generated the data (nonlinear) and the model used to analyze it (log-linear) are not the same, and thus there is no reason to expect perfectly calibrated p-values. Our results support other findings that confidence intervals for MAR(1) models, when fitted to the data generated by more nonlinear models, tend to be 'too narrow' (Certain *et al.*, 2018), in the sense that there is poor coverage of the point estimate. Nonlinear Granger causality methods (Schreiber, 2000, see Paluš, 2008; Amblard & Michel, 2013; Papana *et al.*, 2013, for reviews), could be of use to improve causality detection by obtaining more exact p-values. Transfer entropy (Appendix S1), in particular, admits linear GC as a special case (Barnett & Bossomaier, 2012), and therefore provides an interesting bridge to classical MAR(p) modelling.

Why p-values for CCM were also imperfectly calibrated is unclear. While the original CCM article (Sugihara *et al.*, 2012) method did not directly calculate p-values, further work has recommended to use surrogate time series to do so (e.g., Deyle *et al.*, 2016). Cobey & Baskerville (2016) proposed another method to calculate p-values for CCM based on the increase of ρ with library size L, but although their choice seemed sensible in theory, the resulting p-value was, unfortunately, not working very well in practice (see our SI, section S2.1). We therefore considered different surrogate-based p-values and chose the best-performing ones (Figs. S2, S3 in Appendices), but in some cases - with a confounding abiotic driver or many species - this was not completely satisfactory. More work on formal statistical inference for CCM remains to be done. Another idea could be to combine both worlds and perform surrogate-based nonlinear Granger-causality inference (Schreiber, 2000; Schreiber & Schmitz, 2000; Paluš, 2008).

The specific challenges of high-dimensional, many species interacting systems

We found ~~here~~ that both GC and CCM were scalable to larger interaction networks (10 or 20 species) for relatively long time series by ecological standards (i.e., 500 to 700 time steps). We used both false discovery rate corrections and regularized models (i.e., LASSO-penalized MAR(1) models developed for modular interaction networks, Charbonnier *et al.*, 2010).

One surprising find, for Granger analyses, was that the structured LASSO did not outperform the FDR-corrected pairwise analyses. One way to interpret these results is in terms of correcting for confounders vs collider bias. Fitting a high-dimensional model, even with regularization through the LASSO, has the benefits of including in the estimation of an interaction $j \to i$ the other (potentially many) interacting species. This can be construed as correcting for potential confounders when estimating an interaction. However, any incorrectly included species the network can generate what is known in causal inference as collider bias (Pearl, 2009): if species 6 does not truly affect species 1, but is included in the dynamical model for species 1, then the effect $2 \to 1$ might be poorly estimated. Therefore, there is a trade-off between accounting for confounding factors and avoiding collider bias. Pairwise FDR-corrected analyses seemed to realize the best-trade off, for the fairly modular 10 and 20 autonomous networks that we have considered here.

However, the performances of the structured LASSO MAR(1) models may be diminished by other choices: (1) we used MAR(1) not MAR(p), which limits the ability of the autoregressive model to mimic the nonlinear system and (2) we did not use iterative model fitting, e.g., using the first inferred network as prior for the latent structure or initial condition for further estimates. Both of these ideas may improve the network inference. One reason why we used MAR(1) modelling for the high-dimensional systems, outside of just simplificity, was that $p = 1$ was selected in 20 species case based on BIC, SI section S2.9. But this selection of the lag itself did not use regularization. Selecting both the interaction matrix sparsity and the lag order through regularization in high-dimensional MAR(p) modelling is extremely challenging because there are many ways to connect the number of time lags p to the LASSO penalties (Michailidis & d'Alché Buc, 2013; Nicholson *et al.*, 2017). Mainali *et al.* (2019) recently used the R package `BigVAR` (Nicholson *et al.*, 2017), a promising package for penalized MAR(p) fitting, but few interactions were found and model performance was not evaluated with simulations; we have found here in contrast that without a latent network structure, unstructured LASSO-based methods perform poorly on large networks. Hence our choice of `SIMoNe` (Charbonnier *et al.*, 2010; Chiquet *et al.*, 2008) which sticks to $p = 1$ but allows the latent structure to be specified as a stochastic block model (Daudin *et al.*, 2008) (see Appendix S1.1 for details). Combining structured LASSO modelling with models more sophisticated than MAR(1) remains an area where development is needed.

Going further with causal inference for nonlinear and stochastic ecological dynamical systems

Overall, both linear Granger causality and convergent cross-mapping can show good recall (sensitivity) and specificity for highly nonlinear and stochastic dynamical systems. Their domains of applicability overlap to a great degree. Rather than choosing one of these frameworks based on the supposed degree of nonlinearity or stochasticity of the ecological system studied (e.g., Runge *et al.*, 2019), we suggest that which one to use may be chosen based on the goal and constraints of the analysis.

For instance, (log)-linear GC, being a fully parametric framework can easily be extended to situations where we have small counts that preclude data transformation. This requires using a log-link function rather than an actual log transformation, as in Poisson Log-Normal models (e.g., Chiquet *et al.*, 2018) and other flavours of latent variable modelling (e.g., Warton *et al.*, 2015; Ovaskainen *et al.*, 2017). It may likewise be very useful when one wants to introduce compositionality constraints in microbiome studies. Conversely, convergent cross-mapping or nonlinear Granger techniques allow for a much finer reconstruction of the attractor shape, which can be very useful to compare to the attractor shape of candidate mechanistic models (e.g., coupled differential equations models), if those exist. In some cases, using both frameworks, to increase the robustness of the interaction inference, is another idea (Hannisdal *et al.*, 2017).

Looking at the various implementations of GC and CCM, it seems that the most critical methodological choices are rarely located along a "linear vs nonlinear model" gradient, but instead boil down to two characteristics. First, details matter: faulty selection of the lag order p of autoregressive models results in nonsensical GC inference, and yet proper p selection yields causal inferences fairly robust to nonlinearities. Likewise, versions of CCM including significance testing are quite sensitive to the p-value definition, and surrogate-based p-values should be preferred. In other words, the devil is always in the details of the test or model selection, for Granger-based or CCM-based methods alike. Second, a key choice to make is what constitutes the "conditioning set", i.e., the variables that are known to be important confounders and are de facto included in the time series model (Eichler, 2013). For instance, an unknown confounder such as seasonal temperature or an invading species can massively thwart any attempt at interaction inference if not corrected for. And even when corrected for (i.e., adding the confounder to the autoregression or considering surrogate time series), this is the scenario where we observed the largest proportion of false positives for both GC and CCM. Strategies to better understand how to choose and handle this conditioning set when performing causal inference will be, we believe, a very important feature of ecological interaction inference for the years to come. Several algorithms have been already put forward (Eichler, 2013; Runge, 2018; Runge *et al.*, 2019), and much remains to be done to better incrementally select variables in order to assemble networks.

Acknowledgements

FB thanks Julien Chiquet for advice on the structured LASSO and Grégoire Certain for discussions on MAR modelling. FB and CP were supported by the French ANR through LabEx COTE (ANR-10-LABX-45).

Author contributions

All authors contributed to the project design. FB and CP constructed the case studies, wrote the computer code, and analysed the real and simulated data. All authors contributed to the interpretation of the results. FB wrote a first draft of the manuscript, which was then edited by all authors.

Data accessibility

Codes for the analyses presented in this paper are available at https://github.com/fbarraquand/GCausality.

References

Aalen, O.O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987, 177–190.

Aalen, O.O., Røysland, K., Gran, J.M. & Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175, 831–861.

Adler, P., Ellner, S. & Levine, J. (2010). Coexistence of perennial plants: an embarrassment of niches. *Ecology letters*, 13, 1019–1029.

Adler, P., Smull, D., Beard, K., Choi, R., Furniss, T., Kulmatiski, A., Meiners, J., Tredennick, A. & Veblen, K. (2018). Competition and coexistence in plant communities: intraspecific competition is stronger than interspecific competition. *Ecology Letters*, 21, 1319–1329.

Amblard, P.O. & Michel, O. (2013). The relation between granger causality and directed information theory: A review. *Entropy*, 15, 113–143.

Ambroise, C., Chiquet, J., Matias, C. *et al.* (2009). Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3, 205–238.

Barnett, L., Barrett, A.B. & Seth, A.K. (2009). Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103, 238701.

Barnett, L. & Bossomaier, T. (2012). Transfer entropy as a log-likelihood ratio. *Physical review letters*, 109, 138105.

Barnett, L. & Seth, A.K. (2014). The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223, 50–68.

Barraquand, F., Picoche, C., Maurer, D., Carassou, L. & Auby, I. (2018). Coastal phytoplankton community dynamics and coexistence driven by intragroup density-dependence, light and hydrodynamics. *Oikos*, 127, 1834–1852.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Berlow, E.L., Neutel, A.M., Cohen, J.E., De Ruiter, P.C., Ebenman, B., Emmerson, M., Fox, J.W., Jansen, V.A., Iwan Jones, J., Kokkoris, G.D. *et al.* (2004). Interaction strengths in food webs: issues and opportunities. *Journal of animal ecology*, 73, 585–598.

Carr, A., Diener, C., Baliga, N.S. & Gibbons, S.M. (2019). Use and abuse of correlation analyses in microbial ecology. *The ISME Journal*.

Cazelles, K., Araújo, M.B., Mouquet, N. & Gravel, D. (2016). A theory for species co-occurrence in interaction networks. *Theoretical Ecology*, 9, 39–48.

Certain, G., Barraquand, F. & Gårdmark, A. (2018). How do MAR(1) models cope with hidden nonlinearities in ecological dynamics? *Methods in Ecology and Evolution*, 9, 1975–1995.

Charbonnier, C., Chiquet, J. & Ambroise, C. (2010). Weighted-lasso for structured network inference from time course data. *Statistical applications in genetics and molecular biology*, 9.

Chen, Y., Bressler, S.L. & Ding, M. (2006). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *Journal of neuroscience methods*, 150, 228–37.

Chiquet, J., Mariadassou, M., Robin, S. *et al.* (2018). Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12, 2674–2698.

Chiquet, J., Smith, A., Grasseau, G., Matias, C. & Ambroise, C. (2008). Simone: Statistical inference for modular networks. *Bioinformatics*, 25, 417–418.

Cobey, S. & Baskerville, E.B. (2016). Limits to causal inference with state-space reconstruction for infectious disease. *PloS one*, 11, e0169050.

Coenen, A.R. & Weitz, J.S. (2018). Limitations of correlation-based inference in complex virus-microbe communities. *mSystems*, 3.

Commenges, D. & Gégout-Petit, A. (2009). A general dynamical statistical model with causal interpretation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 719–736.

Coyte, K.Z., Schluter, J. & Foster, K.R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science*, 350, 663–666.

Daudin, J.J., Picard, F. & Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, 18, 173–183.

Dennis, B., Desharnais, R.A., Cushing, J.M., Henson, S.M. & Costantino, R.F. (2001). Estimating chaos and complex dynamics in an insect population. *Ecological Monographs*, 71, 277–303.

Detto, M., Molini, A., Katul, G., Stoy, P., Palmroth, S. & Baldocchi, D. (2012). Causality and persistence in ecological systems: a nonparametric spectral granger causality approach. *The American Naturalist*, 179, 524–535.

Deyle, E.R., Fogarty, M., Hsieh, C.h., Kaufman, L., MacCall, A.D., Munch, S.B., Perretti, C.T., Ye, H. & Sugihara, G. (2013). Predicting climate effects on pacific sardine. *Proceedings of the National Academy of Sciences*, 110, 6430–6435.

Deyle, E.R., Maher, M.C., Hernandez, R.D., Basu, S. & Sugihara, G. (2016). Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences*, 113, 13081–13086.

Ding, M., Chen, Y. & Bressler, S. (2006). Granger causality: Basic theory and application to neuroscience. *Handbook of time series analysis*, pp. 437–460.

Ebisuzaki, W. (1997). A method to estimate the statistical signifiance of a correlation when the data are serially correlated. *Journal of Climate*, 10, 7.

Eichler, M. (2013). Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371, 20110613.

Ellner, S. & Turchin, P. (2005). When can noise induce chaos and why does it matter: a critique. *Oikos*, 111, 620–631.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American statistical association*, 77, 304–313.

Geweke, J.F. (1984). Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79, 907–915.

Gibbons, S.M., Kearney, S.M., Smillie, C.S. & Alm, E.J. (2017). Two dynamic regimes in the human gut microbiome. *PLoS computational biology*, 13, e1005364.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424–438.

Grziwotz, F., Strauß, J.F., Hsieh, C.h. & Telschow, A. (2018). Empirical dynamic modelling identifies different responses of aedes polynesiensis subpopulations to natural environmental variables. *Scientific reports*, 8, 16768.

Hampton, S.E., Holmes, E.E., Scheef, L.P., Scheuerell, M.D., Katz, S.L., Pendleton, D.E. & Ward, E.J. (2013). Quantifying effects of abiotic and biotic drivers on community dynamics with multivariate autoregressive (MAR) models. *Ecology*, 94, 2663–2669.

Hannisdal, B., Haaga, K.A., Reitan, T., Diego, D. & Liow, L.H. (2017). Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20170722.

Hannisdal, B. & Liow, L.H. (2018). Causality from palaeontological time series. *Palaeontology*, 61, 495–509.

Harford, W.J., Karnauskas, M., Walter, J.F. & Liu, H. (2017). Non-parametric modeling reveals environmental effects on bluefin tuna recruitment in atlantic, pacific, and southern oceans. *Fisheries oceanography*, 26, 396–412.

Ives, A., Dennis, B., Cottingham, K. & Carpenter, S. (2003). Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, 73, 301–330.

Ives, A.R. (1995). Predicting the response of populations to environmental change. *Ecology*, 76, 926–941.

Jost, C. & Ellner, S.P. (2000). Testing for predator dependence in predator-prey dynamics: a non-parametric approach. *Proceedings of the Royal Society of London B: Biological Sciences*, 267, 1611–1620.

Krakovská, A., Jakubík, J., Chvosteková, M., Coufal, D., Jajcay, N. & Paluš, M. (2018). Comparison of six methods for the detection of causality in a bivariate time series. *Phys. Rev. E*, 97, 042207.

Link, J.S. (2002). What does ecosystem-based fisheries management mean. *Fisheries*, 27, 18–21.

Loreau, M. & de Mazancourt, C. (2008). Species synchrony and its drivers: neutral and nonneutral community dynamics in fluctuating environments. *The American Naturalist*, 172, E48–E66.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Mainali, K., Bewick, S., Vecchio-Pagan, B., Karig, D. & Fagan, W.F. (2019). Detecting interaction networks in the human microbiome with conditional granger causality. *PLoS computational biology*, 15, e1007037.

Marinazzo, D., Pellicoro, M. & Stramaglia, S. (2008). Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77, 1–9.

May, R. (1973). *Stability and complexity in model ecosystems*. Princeton University Press, Princeton, USA.

Mayr, E. (1961). Cause and effect in biology. *Science*, 134, 1501–1506.

Michailidis, G. & d'Alché Buc, F. (2013). Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues. *Mathematical biosciences*, 246, 326–334.

Mønster, D., Fusaroli, R., Tylén, K., Roepstorff, A. & Sherson, J.F. (2017). Causal inference from noisy time-series data—testing the convergent cross-mapping algorithm in the presence of noise and external influence. *Future Generation Computer Systems*, 73, 52–62.

Mukhopadhyay, N.D. & Chatterjee, S. (2006). Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23, 442–449.

Mutshinda, C.M., O' Hara, R.B. & Woiwod, I.P. (2011). A multispecies perspective on ecological impacts of climatic forcing. *Journal of Animal Ecology*, 80, 101–107.

Mutshinda, C.M., O'Hara, R.B. & Woiwod, I.P. (2009). What drives community dynamics? *Proceedings of the Royal Society B: Biological Sciences*, 276, 2923–2929.

Nicholson, W., Matteson, D. & Bien, J. (2017). Bigvar: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:1702.07094*.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. & Abrego, N. (2017). How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology Letters*, 20, 561–576.

Paluš, M. (2008). From Nonlinearity to Causality: Statistical testing and inference of physical mechanisms underlying complex dynamics. *Contemporary Physics*, 48, 307–348.

Papana, A., Kyrtsou, C., Kugiumtzis, D. & Diks, C. (2013). Simulation study of direct causality measures in multivariate time series. *Entropy*, 15, 2635–2661.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96–146.

Pikitch, E., Santora, E., Babcock, A., Bakun, A., Bonfil, R., Conover, D., Dayton, P., Doukakis, P., Fluharty, D., Heheman, B. *et al.* (2004). Ecosystem-based fishery management. *Science*, 305, 346–347.

Runge, J. (2014). *Detecting and quantifying causality from time series of complex systems.* Ph.D. thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät.

Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28, 075310.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M.D., Muñoz-Marí, J. *et al.* (2019). Inferring causation from time series in earth system sciences. *Nature communications*, 10, 2553.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, 85, 461.

Schreiber, T. & Schmitz, A. (2000). Surrogate time series. *Physica D: Nonlinear Phenomena*, 142, 346–382.

Schweder, T. (1970). Composable markov processes. *Journal of applied probability*, 7, 400–410.

Sims, C. (1980). Macroeconomics and reality. *Econometrica*, 48, 1–48.

Stone, L. & Roberts, A. (1991). Conditions for a species to gain advantage from the presence of competitors. *Ecology*, 72, 1964–1972.

Sugihara, G., May, R., Ye, H., Hsieh, C.h., Deyle, E., Fogarty, M. & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, 338, 496–500.

Tibshirani, R., Wainwright, M. & Hastie, T. (2015). *Statistical learning with sparsity: the Lasso and generalizations.* Chapman and Hall/CRC.

Veilleux, B.G. (1979). An analysis of the predatory interaction between paramecium and didinium. *J Anim Ecol*, 48, 787–803.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K. (2015). So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, 30, 766–779.

Wootton, J. & Emmerson, M. (2005). Measurement of interaction strength in nature. *Annual Review of Ecology, Evolution and Systematics*, 36, 419–444.

Yang, G., Wang, L. & Wang, X. (2017). Reconstruction of complex directional networks with group lasso nonlinear conditional granger causality. *Scientific reports*, 7, 2991.

Ye, H., Clark, A., Deyle, E., Munch, S., Cai, J., Cowles, J., Daon, Y., Edwards, A., Keyes, O., Stagge, J., Ushio, M., White, E. & Sugihara, G. (2018). rEDM: Applications of Empirical Dynamic Modeling from Time Series.

Ye, H., Deyle, E.R., Gilarranz, L.J. & Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5.

Ye, H. & Sugihara, G. (2016). Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science*, 353, 922–925.

Yodzis, P. (1998). Local trophodynamics and the interaction of marine mammals and fisheries in the benguela ecosystem. *Journal of Animal Ecology*, 67, 635–658.

Supporting Information – Appendices

S1 Extensions of Granger causality

S1.1 LASSO-based MAR(1) models

We follow here the presentation of Charbonnier *et al.* (2010); Chiquet *et al.* (2008), with some notational adaptations from Ives *et al.* (2003) and keep our notations in line to those of the main text. We start with the MAR(1) model without external input for the log-abundance vector $\mathbf{x}_t$, which we assume to be scaled and centered. The model is given by

$$\mathbf{x}_t = \mathbf{a} + \mathbf{B}\mathbf{x}_{t-1} + \mathbf{e}_t, \mathbf{e}_t \sim \mathcal{N}_d(\mathbf{0}, \sigma^2\mathbf{I}) \tag{24}$$

where matrix $\mathbf{B}$ has dimension $d \times d$, same as in the main text, $\mathbf{a}$ is a d-dimensional vector, and the noise elements are independent of $\mathbf{x}_t$ and each other. The model is observed for times $t = 1, ..., T+1$ which then defines a $T \times p$ matrix of observed densities $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_1, ..., \mathbf{x}_T]'$ (the prime denotes matrix transposition) and a $T \times p$ matrix of densities observed just one time step after $\mathbf{Y} = [\mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_{T+1}]'$.

As remarked by Charbonnier *et al.* (2010), this model can be fitted to data by using the following relations:

- $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ is the empirical variance-covariance matrix
- $\mathbf{V} = \frac{1}{n}\mathbf{X}'\mathbf{Y}$ is the temporal autocovariance matrix

The log-likelihood of the MAR(1) process is then equivalent to

$$\max_{\mathbf{B}}\{\mathrm{Tr}(\mathbf{V}'\mathbf{B}) - \frac{1}{2}\mathrm{Tr}(\mathbf{B}'\mathbf{S}\mathbf{B})\} \tag{25}$$

The solution of this maximization problem is then given by $\mathbf{B}^{\mathrm{mle}} = \mathbf{S}^{-1}\mathbf{V}$. This is proved by (1) reducing eq. 25 to an OLS problem and (2) compute $\mathbf{B}^{\mathrm{ols}}$ as $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{S}^{-1}\mathbf{V}$. This solution requires that $\mathbf{S}$ is invertible, which typically won't be the case when $T > d$, for which we need some degree of regularization.

LASSO-based B estimate Sparsity can be enforced with a classical L_1 penalty, so that

$$\max_{\mathbf{B}}\{\mathrm{Tr}(\mathbf{V}'\mathbf{B}) - \frac{1}{2}\mathrm{Tr}(\mathbf{B}'\mathbf{S}\mathbf{B}) - \rho||\mathbf{B}||_1\} \tag{26}$$

Unfortunately, this tends to (a priori) penalize all coefficients alike, and therefore to consider by default that the network has no structure. We have always found this method to lead to worse results that those assuming some structure, for modular ground truth networks like those considered in our 10- and 20-species simulations.

When the network is structured, one can introduce a latent structure by assuming that a network $\mathcal{P}$ is structured into $\mathcal{Q}$ classes. We note Z_{iq} the indicator function (a random variable) whose value is 1 if species i belongs in class q (this can be a module, for instance).

The choice of the latent structure follows Ambroise *et al.* (2009), who use the mixture framework of Daudin *et al.* (2008). A Laplace distribution on the network weights is chosen, a Laplace prior on coefficients being equivalent to LASSO optimization (Tibshirani *et al.*, 2015). It is therefore assumed that the a priori link strength between species i and species j is distributed as

$$f_{ijql}(x) = \frac{1}{2\lambda_{ql}} \exp(-\frac{|x|}{\lambda_{ql}}) \tag{27}$$

where λ_{ql} describe the intensity of the link between classes q and l.

Implementing this prior on the interaction strength then equates to the following optimization problem for the likelihood $\mathcal{L}$ with latent network structure $\mathbf{Z}$.

$$\hat{\mathbf{B}} = \operatorname{argmax} \log \mathcal{L}(\mathbf{Y}, \mathbf{B}; \mathbf{Z}) = \operatorname{argmax} \left\{ \operatorname{Tr}(\mathbf{V}'\mathbf{B}) - \frac{1}{2}\operatorname{Tr}(\mathbf{B}'\mathbf{S}\mathbf{B}) - ||\mathbf{P}^{\mathbf{Z}} \star \mathbf{B}||_1 \right\} \tag{28}$$

where $\mathbf{P}^{\mathbf{Z}} = \left(P_{ij}{}^{\mathbf{Z}}_{i,j\in\mathcal{P}} \right) = \sum_{q,l\in\mathcal{Q}} \frac{Z_{iq}Z_{jl}}{\lambda_{ql}}$ are the penalties encapsulating the network structure (see Ambroise *et al.* 2009 for details on such penalties).

We refer to Charbonnier *et al.* (2010) for the details of the algorithm used here. In essence the particular structure of the model allows to reduce this global LASSO optimization to d LASSO-style problems, which makes it much faster. The tuning of the penalty parameter is then done using BIC (Charbonnier *et al.*, 2010).

S1.2 Transfer entropy and nonlinear Granger causality

Transfer entropy can be defined as

$$\mathcal{T}_{x\to y|z} = H(\mathbf{y}^{T+1}|\mathbf{y}^T, \mathbf{z}^T) - H(\mathbf{y}^{T+1}|\mathbf{y}^T, \mathbf{x}^T, \mathbf{z}^T)$$

where $\mathbf{y}^{T+1} = (y_2, ..., y_{T+1})$ and $\mathbf{y}^T = (y_1, ..., y_T)$ and $\mathbf{x}^T, \mathbf{z}^T$ are similarly defined.The quantity $H(x|y) = H(x,y) - H(y)$ is a conditional entropy, defined with $H(x)$ the Shannon entropy. It has then been shown that the Granger causal measure $\mathcal{G}_{x\to y|z} = \ln(\frac{\sigma_\eta^2}{\sigma_\epsilon^2})$ where the residuals errors are taken from eqs. 2 can be generalized to $\mathcal{T}_{x\to y|z}$. In the linear case, Barnett *et al.* (2009) proved that $\mathcal{G}_{x\to y|z} = 2\mathcal{T}_{x\to y|z}$, so that Granger causality through MAR(1) modelling is a special case of causality defined through transfer entropy.

In general, any method which evaluates whether adding a new time series $\mathbf{x}$ to a dynamical system for variables $y_1, ..., y_n$ improves prediction of y_i can be defined as a generalised conditional GC method evaluating $x \to y_i | (y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_n)$. Quite a number of nonlinear Granger causality inference techniques then fall within this category (e.g., Marinazzo *et al.*, 2008; Paluš, 2008).

S1.3 20-species model interaction matrix

The 20-species model has an interaction structure that is still fairly modular (eq. 29) yet some species act as links between the different modules (e.g., species 4 and 5).

$$\chi = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1
\end{pmatrix} \tag{29}$$

S2 Additional results

S2.1 Choice of p-values and thresholds on effect sizes

During preliminary simulations, we discovered that false causalities in absence of interactions could arise in a number of cases, which could indicate that the usual p-values and associated thresholds ($\alpha = 10\%$ for 2-species simulations, 20% for 10- and 20-species simulations) were not sufficient to deduce interactions. We thus searched for additional conditions on the estimates, such as effect sizes, to conclude to causality. We based our analyses on the stochastic model described in eq. 12.

For Granger-causality, we computed the log-ratio of the residuals sum of squares (using notations from eq. 2 and 3, $\log\left(\frac{\sum \eta_i{}^2}{\sum \epsilon_i{}^2}\right)$) and the average effect of the causal species over all causal lags up to p ($\frac{\sum_j |a_{ji}|}{p}$). We see on Fig. S1 that the log-ratio tends to be a more efficient indicator of causality and that fixing a threshold of 0.04 on this ratio seems to achieve a good balance between false negatives and positives.

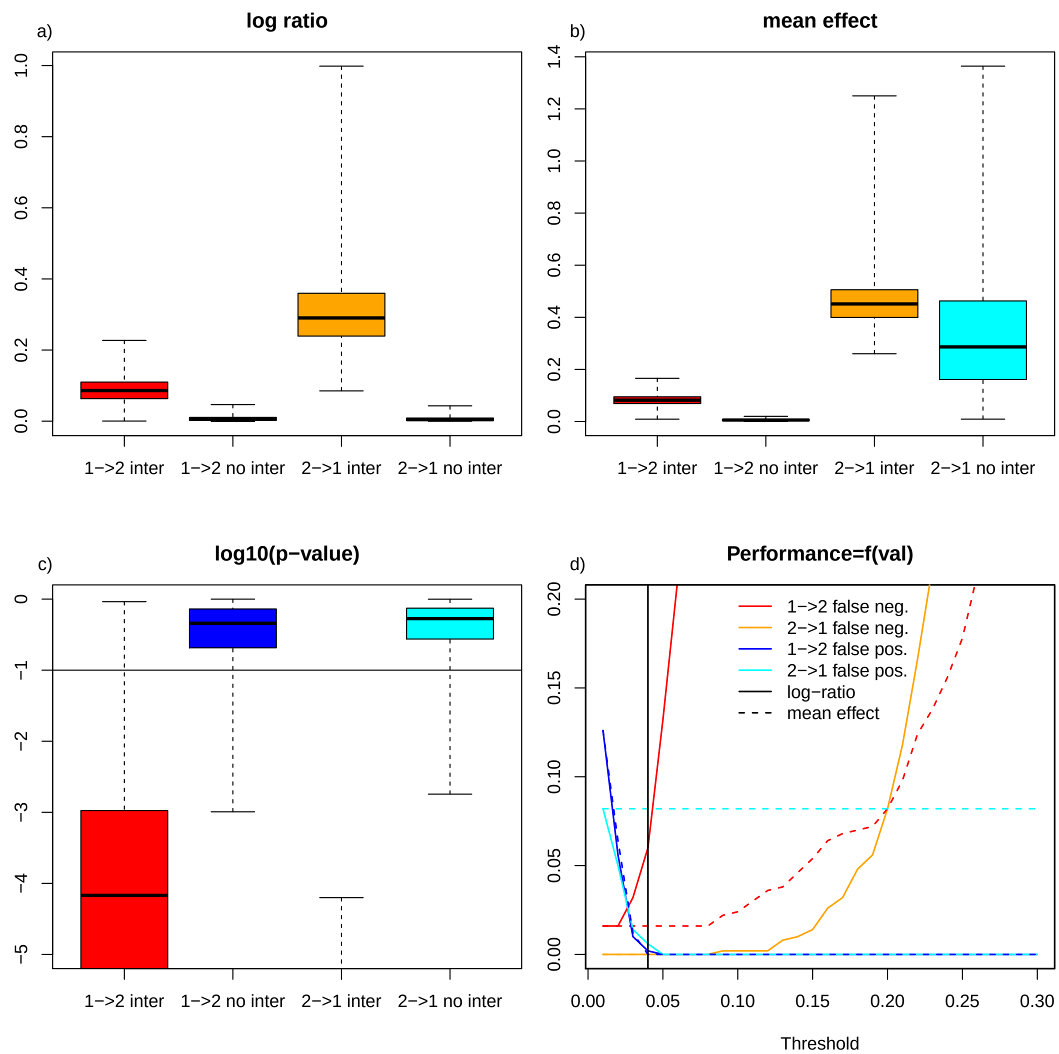

Figure S1: Comparison of methods to determine Granger-causality between two variables in a stochastic model. Log ratio of residuals sum of squares (a) and average effect of the causal species (b) are compared, and the proportions of false negatives (red and orange) and false positives (blue and cyan), depending on the p-value and threshold imposed on these effects, are shown in d)

For the convergent cross-mapping, the computation of the p-value itself was an issue (as discussed in Methods). We compared the p-value described by Cobey & Baskerville (2016), and three different types of surrogates: permutation, distance-based ('twin', the sampling replaces one point by another which remains close in value) or frequency-based ('Ebisuzaki', the time series spectrum is kept during resampling, Ebisuzaki 1997). We also examined the effect of putting a threshold on the value of ρ. We see on Fig. S2 that surrogate-based p-values are more efficient to detect causalities. As they have very similar behaviors, we chose to keep the simplest (and least computationally intensive) method, based on permutation. We also considered a threshold at 0.1 or 0.2 on ρ values to avoid the majority of false positives and false negatives.

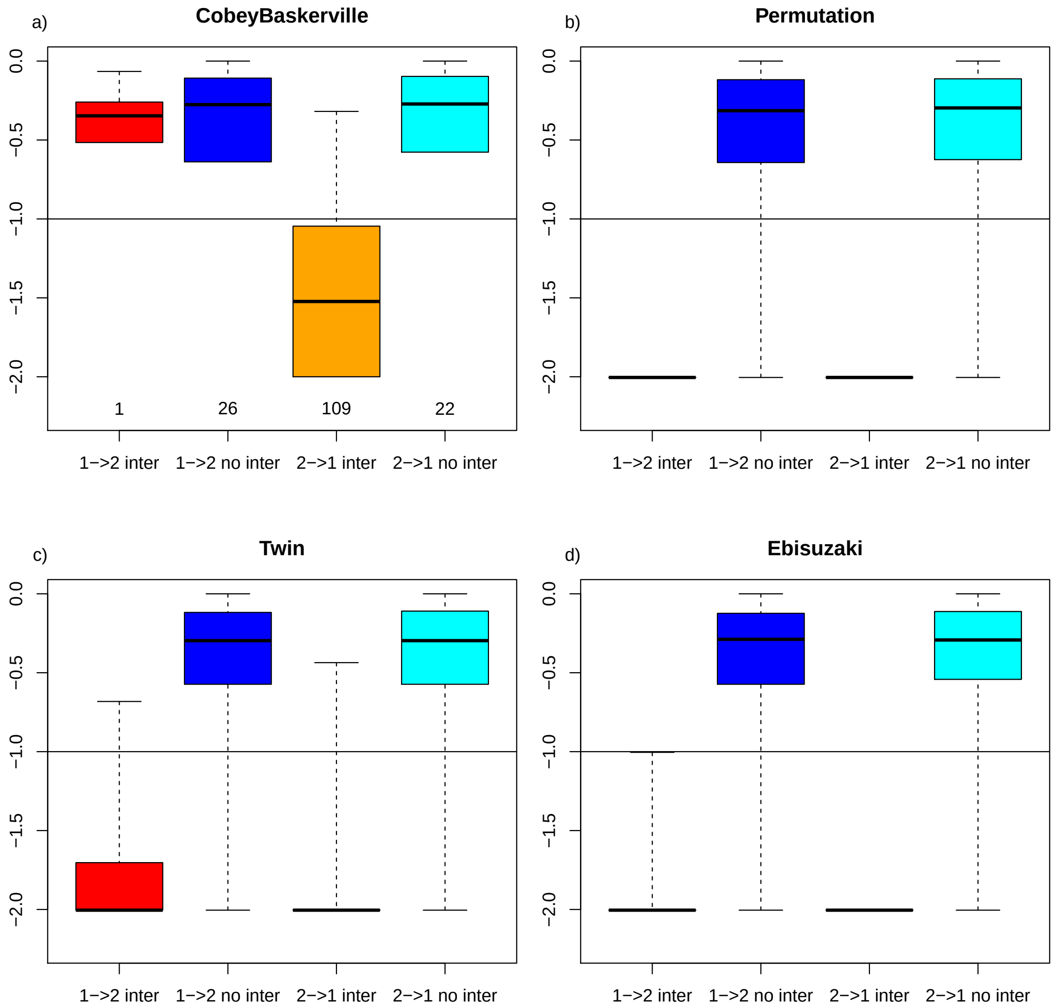

Figure S2: Log10(p-values) for the stochastic 2-species model, using different methods to compute p-values. Cobey & Baskerville (2016) method and permutation-, twin- and Ebisuzaki-based surrogates are compared. The number of p-values which are found to be 0, among the 500 simulations estimated, is written at the bottom of the p-value boxplot.

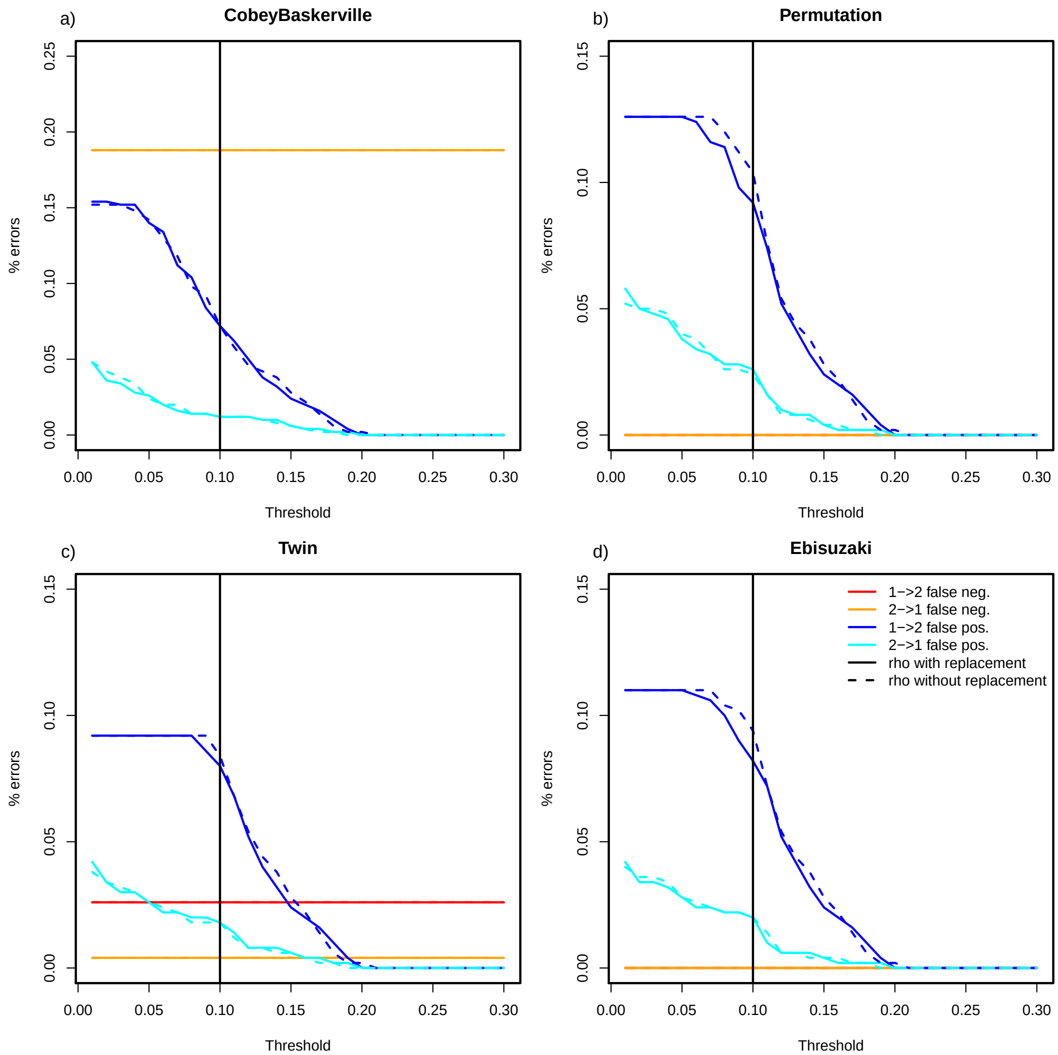

Figure S3: Comparison of the proportion of false negatives and false positives when combining p-values and thresholds on final ρ value for CCM.

S2.2 Example simulation of a 2-species stochastic Ricker model

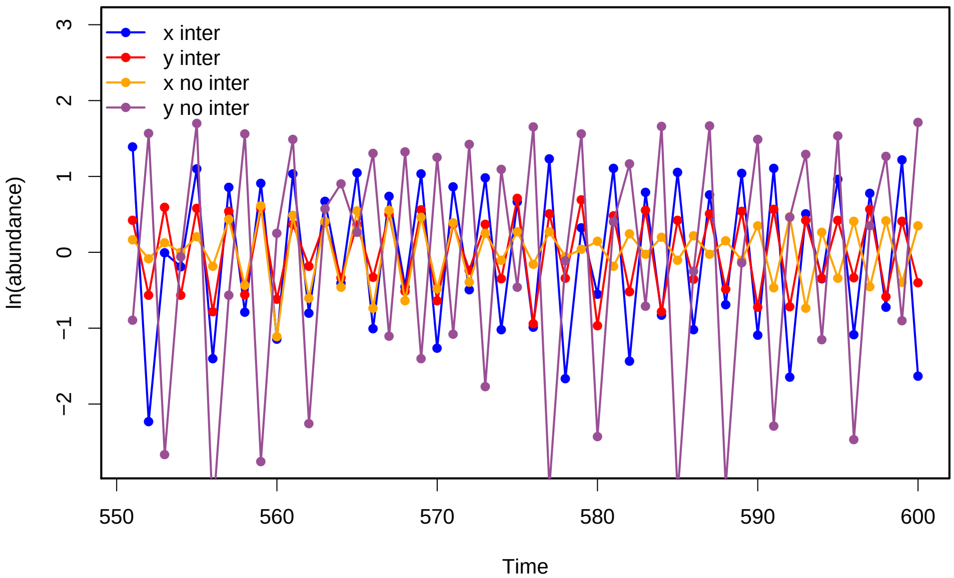

Figure S4: 2-species Ricker model (eq. 12,13) with (blue = species 1, red = species 2) and without (purple = species 1, orange = species 2) competition.

817 ## S2.3 Effect of log-transformation on CCM

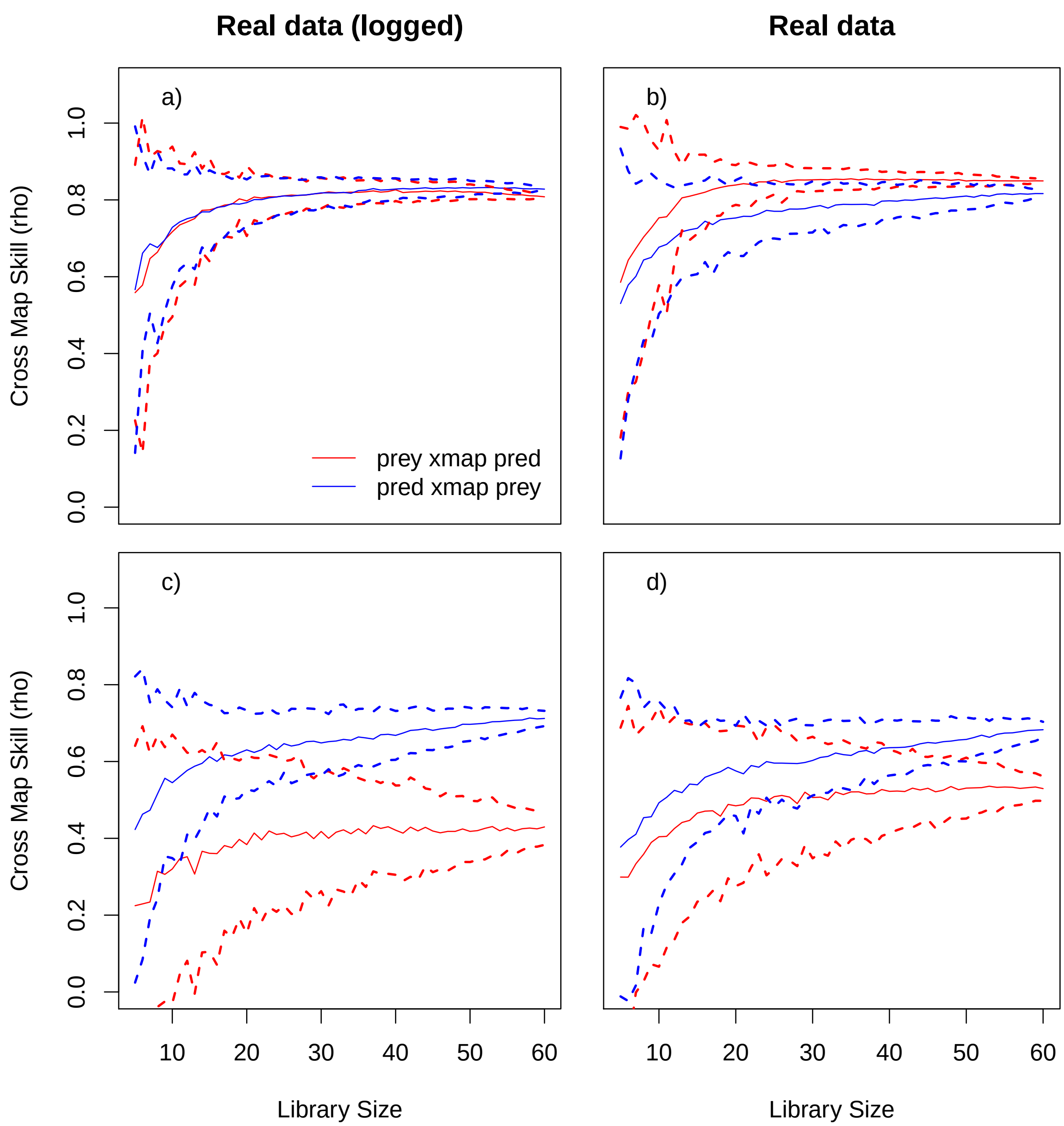

Figure S5: Convergent cross-mapping with (left) and without (right) log-transform of the data for the Veilleux dataset

S2.4 Deterministic two-species competition model

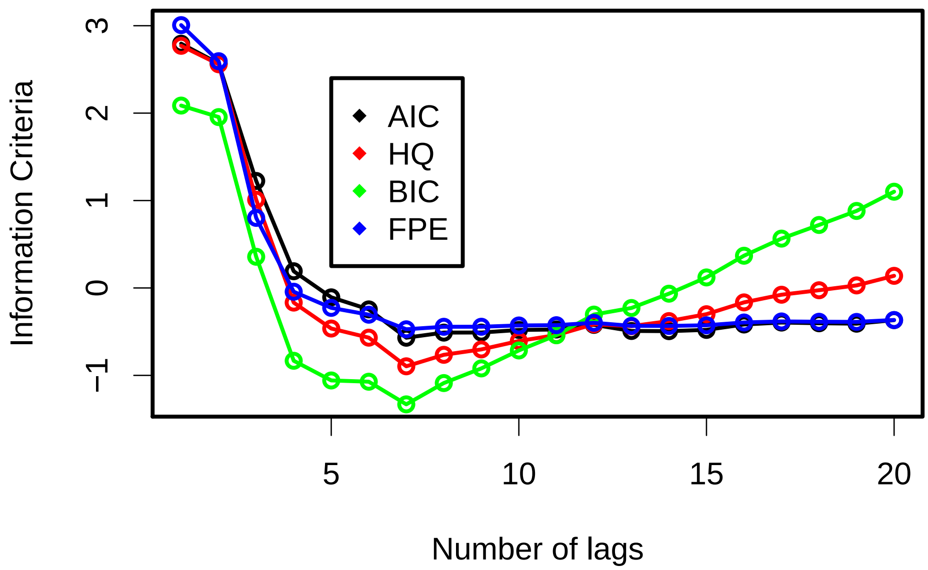

Figure S6: Results of model information criteria vs. lag order for the simulated deterministic competition model of eq. 8.

Lag order p selection in the MAR(p) framework

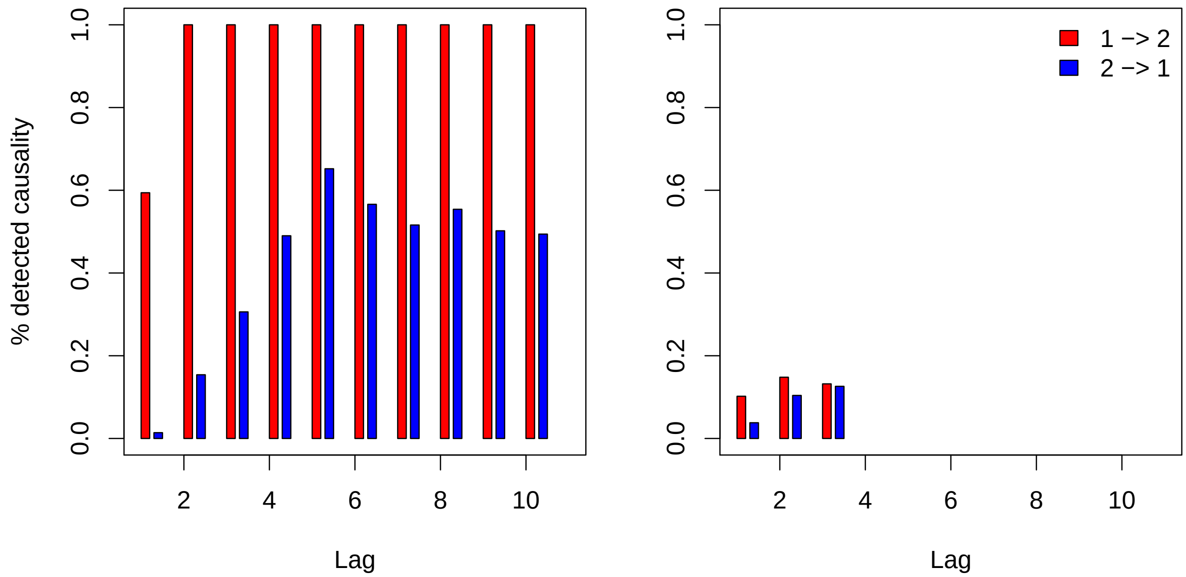

Figure S7: Proportion of detected Granger-causality, at the 10% significance threshold, over 500 chaotic simulations with (left) and without (right) actual interactions between species, depending on the number of time lags taken into account (x-axis). Without interactions, the optimal lag is 3 and the Wald test cannot be performed for p>3.

Granger causality at various lag orders p

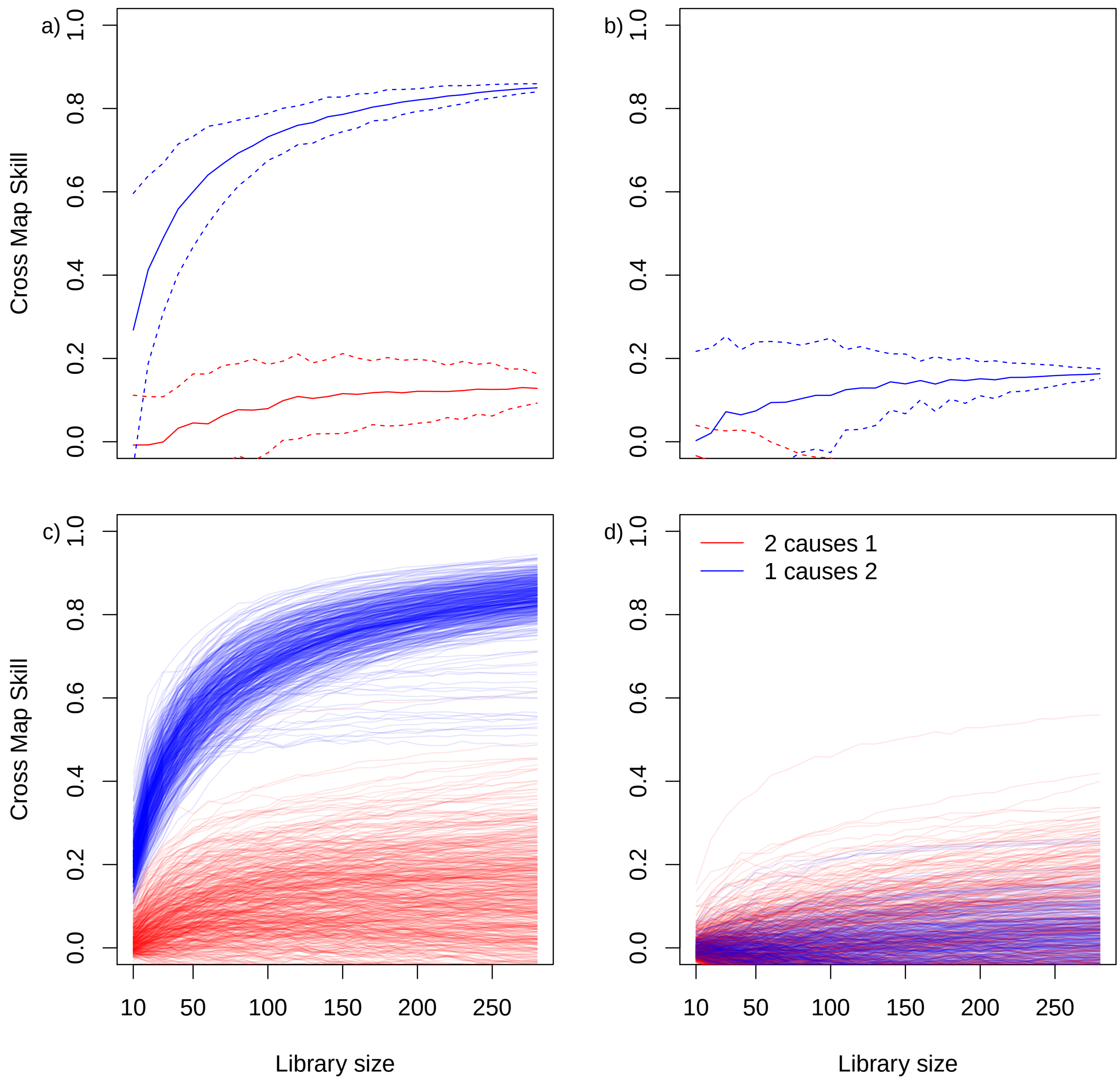

Figure S8: Convergent-cross mapping on simulated a deterministic 2-species competition model, with (left) and without (right) competition between the two species. On the top row, one simulation with (a) and without (b) interactions with associated confidence bands; bottom row, cross-map skill (ρ) for 500 simulations.

821 **Convergent cross-mapping for 500 randomly drawn initial conditions**

822 ## S2.5 CCM on simulated stochastic competition data

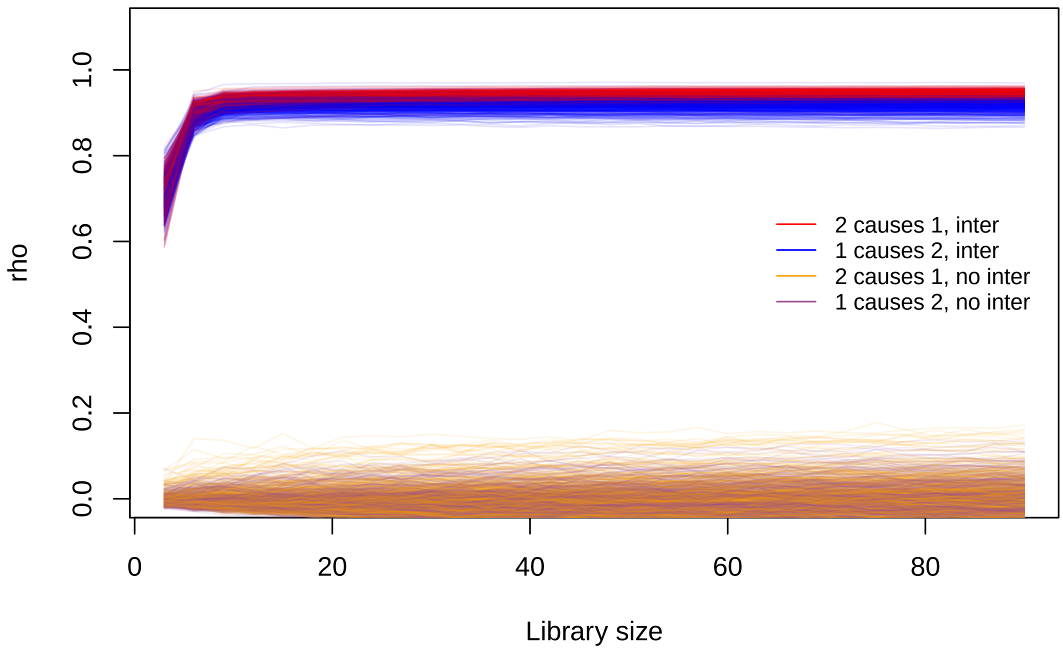

Figure S9: Convergent-cross mapping on simulated stochastic data, with (red, blue) and without (orange, purple) competition between the two species, over 500 simulations.

S2.6 CCM on simulated stochastic competition model forced by a shared abiotic driver

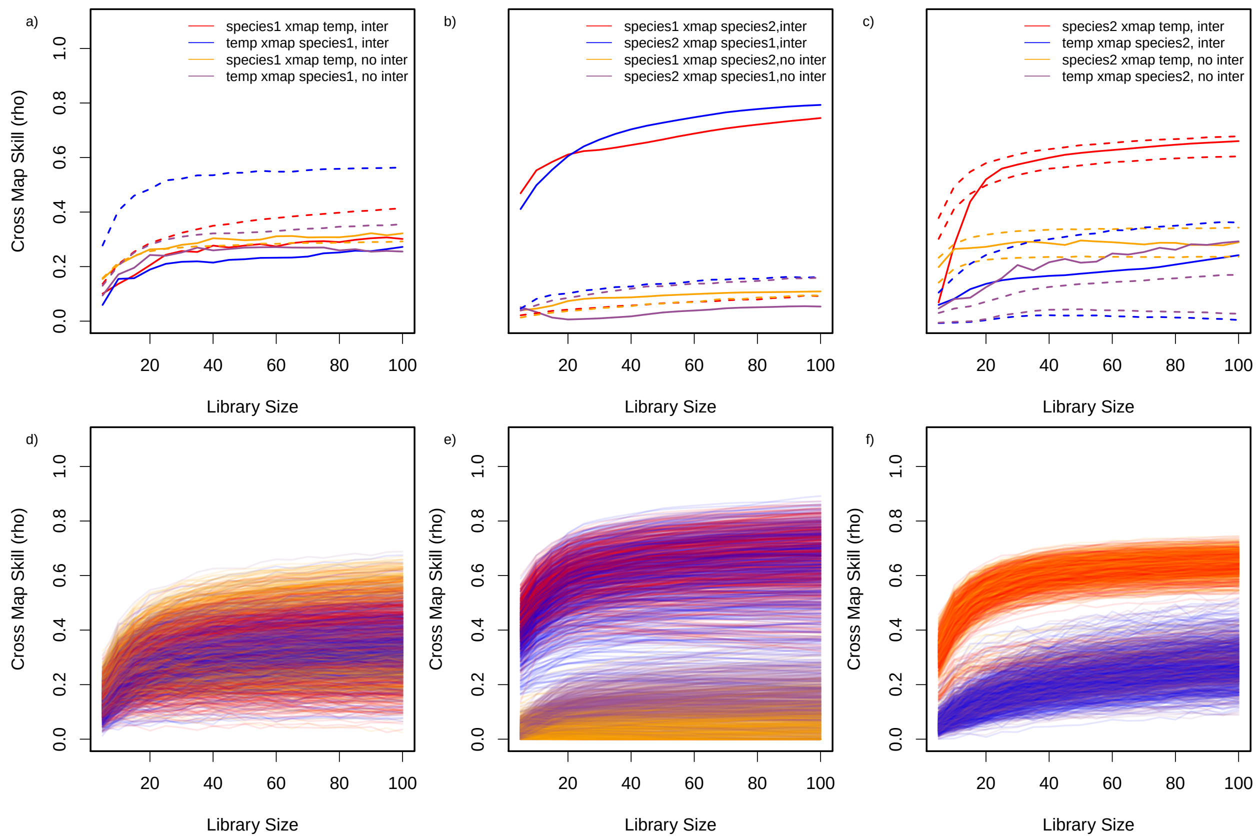

Figure S10: Convergent cross-mapping for the two species forced by an environmental driver (denoted as temp), when interactions are present (blue, red) and when interactions are absent (purple, orange), for 500 simulations. Dashed lines indicate the 10% interval for rho-values obtained from surrogate time series, i.e., time series that have the same seasonal forcing but whose cross-correlations are altered.

S2.7 Causality with respect to the abiotic driver

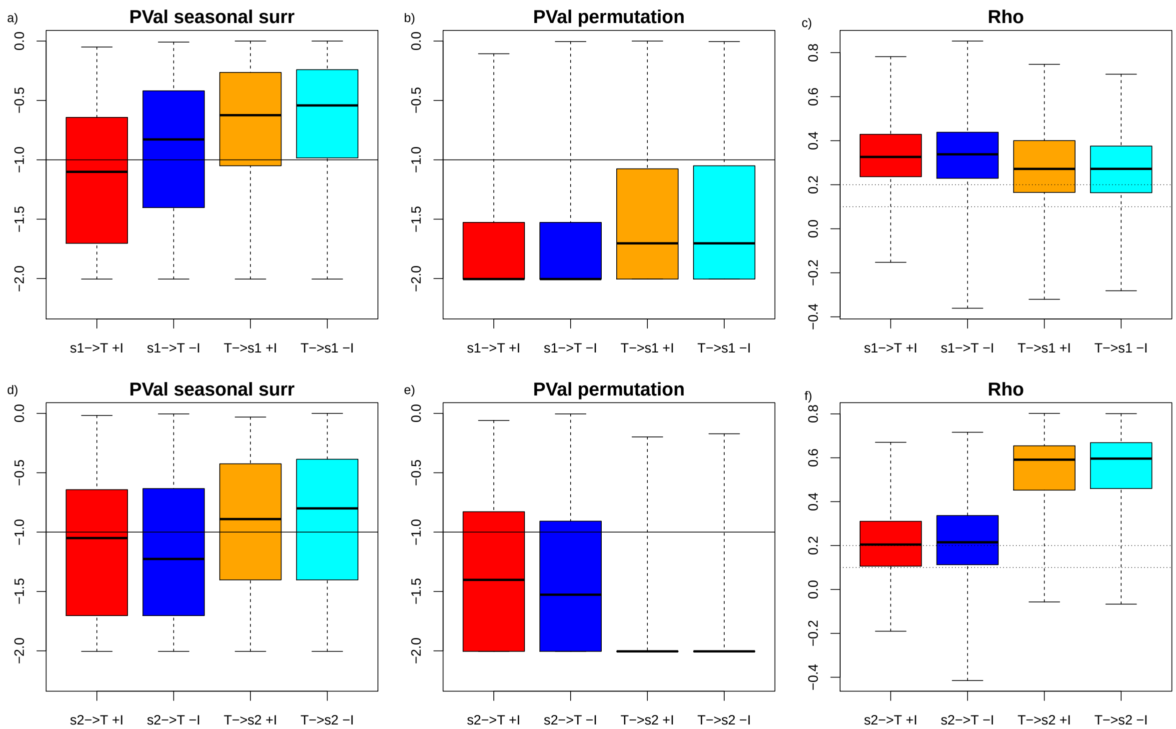

Figure S11: Comparison of log10(p-values) and CCM skill (rho) values to examine effects of temperature (T) on species 1 and 2 (s1 and s2), and the spurious reverse causality (species 1 or 2 causing temperature). Simulations were ran with (+I) and without (-I) interactions between species 1 and 2. The 10% false positive threshold is indicated by a line on the pval plots (p-value must be below this line for the causality to be inferred) while the 0.1 and 0.2 thresholds that could be imposed on rho values are dotted lines in the right panel (ρ must be above the line for the causality to be inferred)

S2.8 From Lotka-Volterra to multivariate autoregressive model

Our objective here is to keep a commensurate interaction matrix between the Lotka-Volterra model and its corresponding log-linearised autoregressive version. We therefore need to compute the corresponding Jacobian matrix $\mathbf{J}$ (see also Ives *et al.* 2003; Certain *et al.* 2018). The Lotka-Volterra model can be written, after centering and log-transformation, as:

$$
\begin{aligned}
\mathbf{n}_{t+1} &= \mathbf{n}_t + \mathbf{A}\mathbf{N}_t + w_t, w_t \sim \mathcal{N}(0, \Sigma) && (30) \\
\Rightarrow n_{i,t+1} &= f_i(n_{k,t})_{k \in [1,S]} + \epsilon_{i,t}, \epsilon_{i,t} \sim \mathcal{N}(0, \sigma') && (31)
\end{aligned}
$$

where $\mathbf{N}_t = (e^{n_{1,t}}, ..., e^{n}_{S,t})^T$ and $f_i(\mathbf{n}) = n_i + a_{i\bullet}\mathbf{N}_t$, with $a_{i\bullet}\mathbf{N}_t = \sum_{k=1}^{S} a_{ik} N_{k,t}$.

We can write the Jacobian matrix elements as $J_{ij} = \frac{\partial f_i}{\partial n_j}$. Then,

$$J_{ij} = \frac{\partial n_i}{\partial n_j} + \sum_{k=1}^{S} a_{ik} \frac{\partial e^{n_k}}{\partial n_j} \tag{32}$$

$$J_{ij} = \delta_{ij} + a_{ij} e^{n_j} = \delta_{ij} + a_{ij} N_j \tag{33}$$

835 S2.9 Lag order selection for the 10- and 20-species model

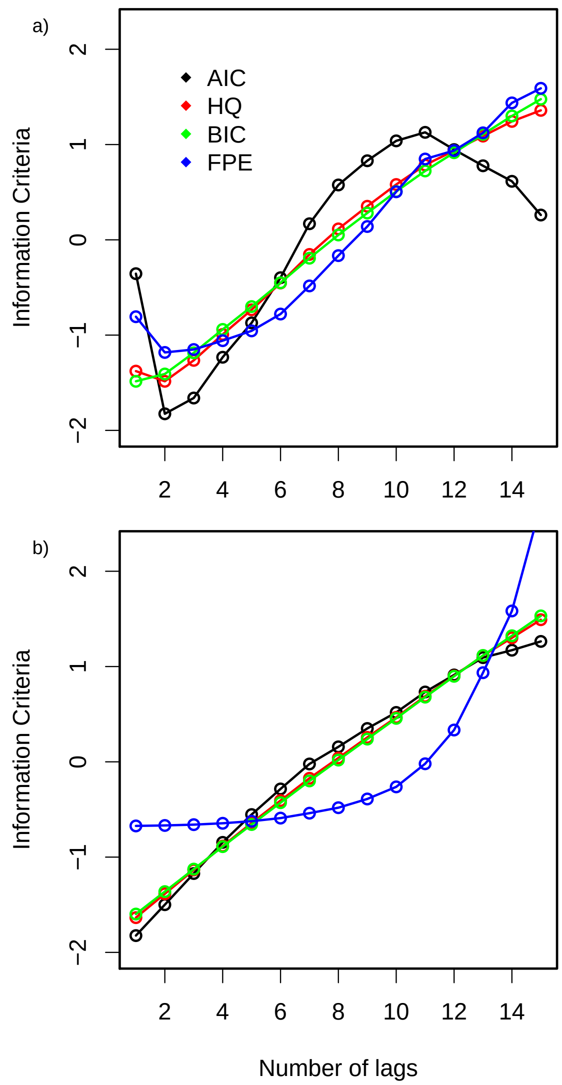

Figure S12: Lag order selection for (a) the 10-species and (b) one of the 20-species stochastic community model.

S2.10 Interaction matrix for the 20-species model

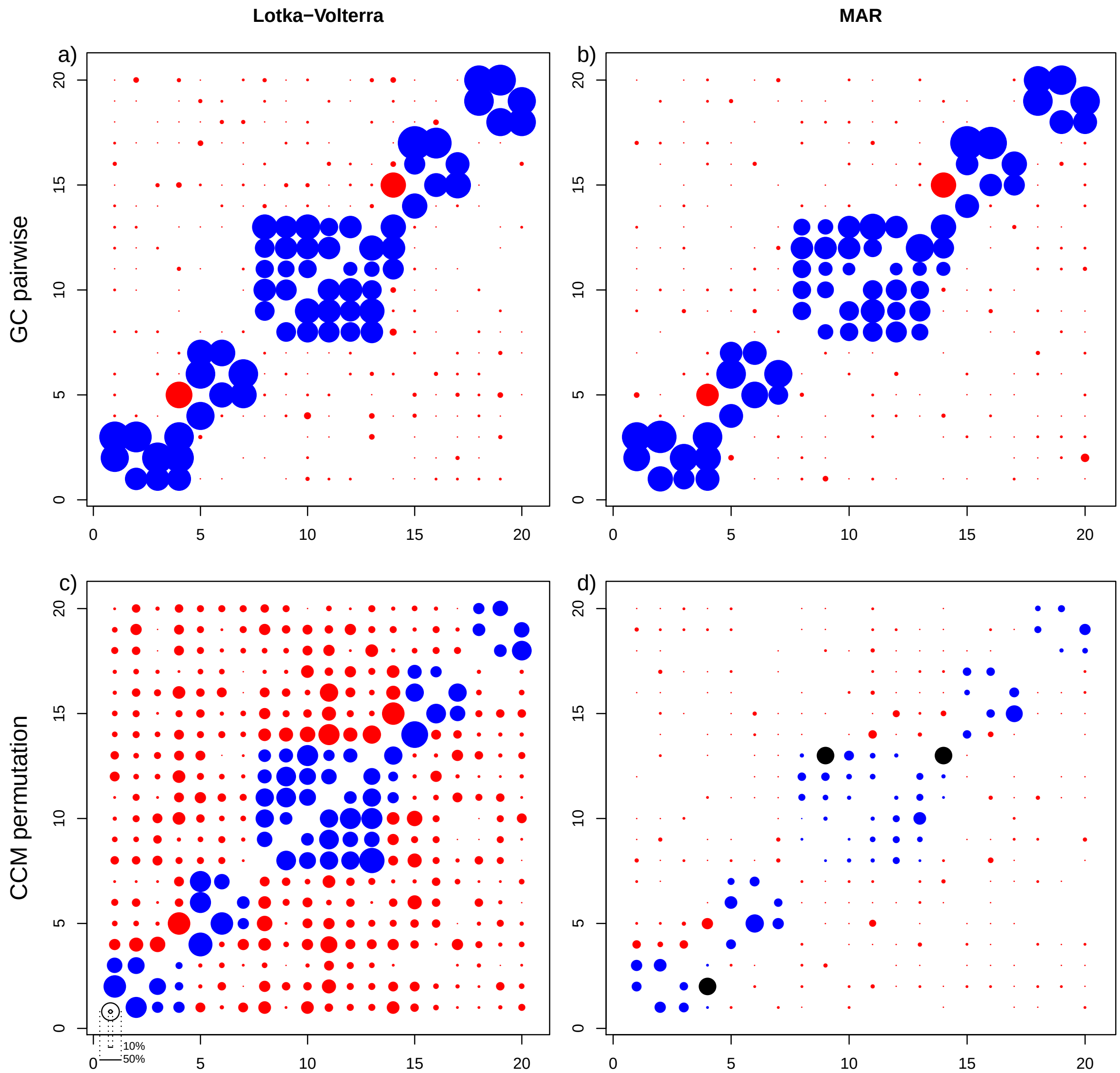

Figure S13: Interaction matrices obtained from pairwise GC (top) or permutation-based surrogates for CCM (bottom) for 20-species communities. Blue circles are the true positives, red circles are false positives and black circles are false negatives. For the true and false positives, the size of the circles is proportional to the proportion of detection over 25 simulations.

837
S2.11 Alternative network reconstruction methods

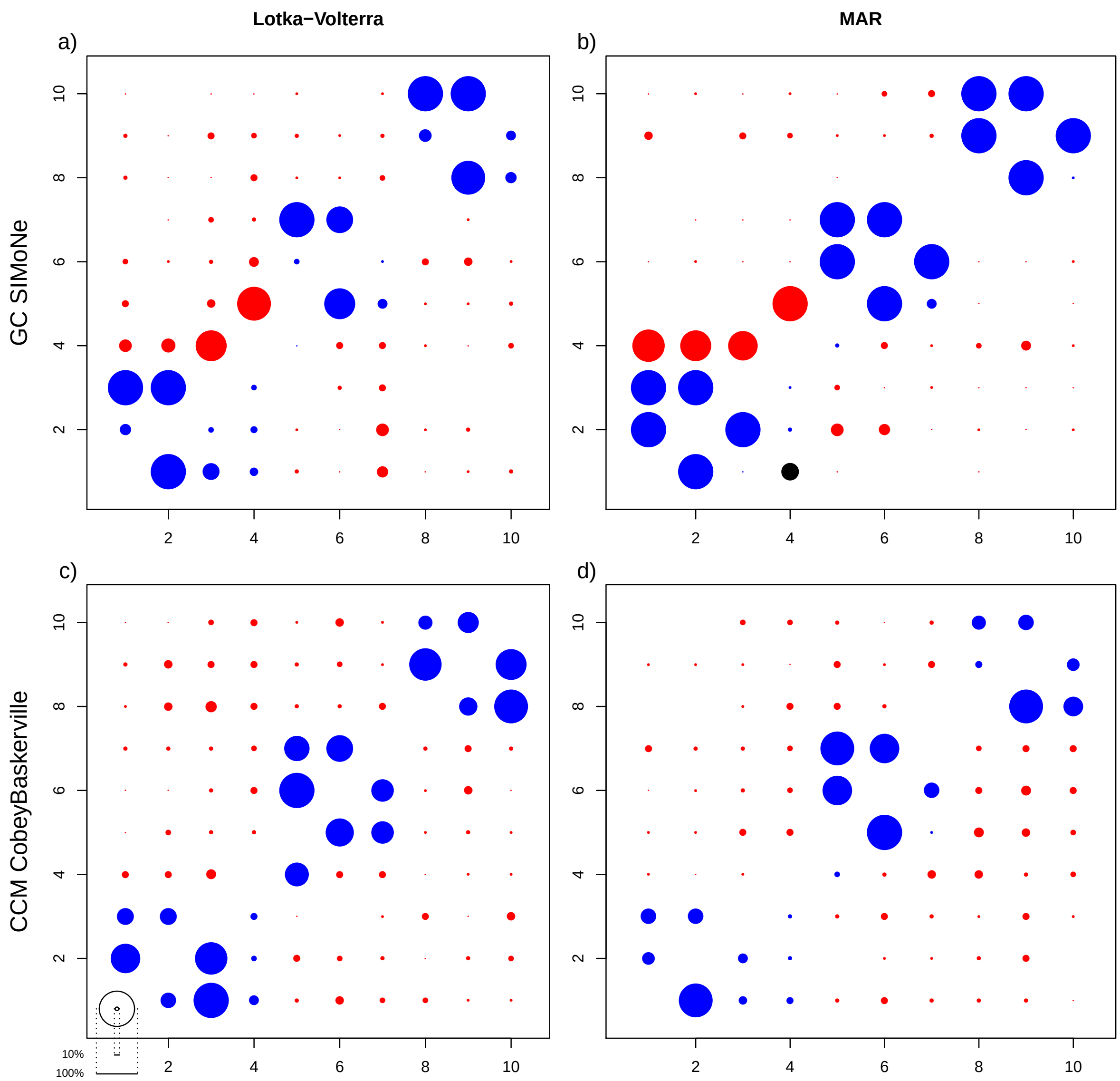

Figure S14: Interaction matrices obtained from GC (top) or CCM (bottom) for 10-species communities, based on alternative ways of computing p-values (see main text). Blue circles are the true positives, red circles are false positives and black circles are false negatives. For the true and false positives, the size of the circles is proportional to the proportion of detection over 25 simulations.

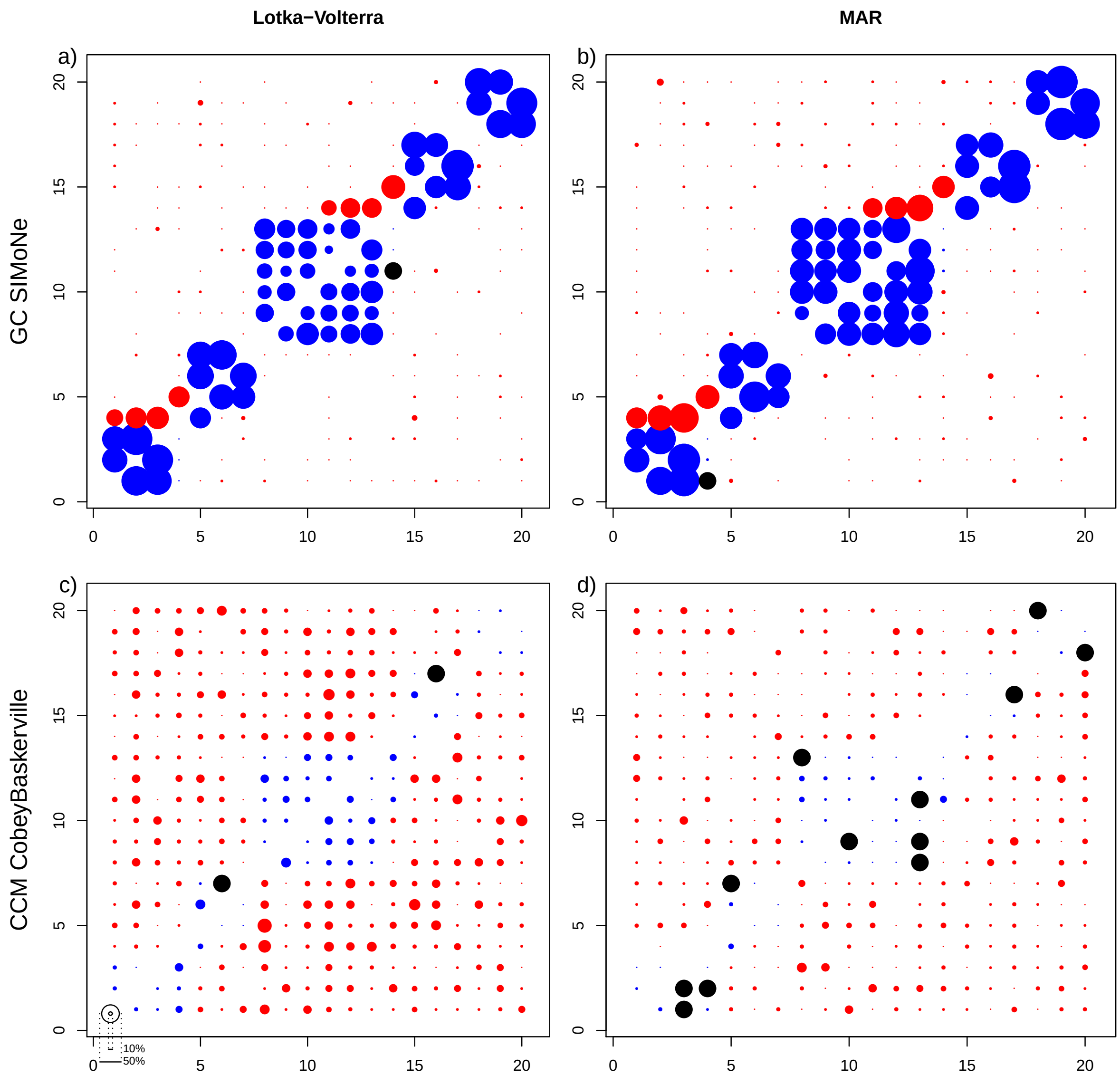

Figure S15: Interaction matrices obtained from GC (top) or CCM (bottom) for 20-species communities, based on alternative ways of computing p-values (see main text). Blue circles are the true positives, red circles are false positives and black circles are false negatives. For the true and false positives, the size of the circles is proportional to the proportion of detection over 25 simulations.