Identifying dynamic coupling between interacting populations using Granger causality and convergent cross-mapping: a reconciliation

F. Barraquand, C. Picoche, M. Detto & F. Hartig

April 30, 2019

Abstract

Finding who interacts with whom based on time series data is a key endeavour of statistical ecology. Here, we show that multivariate autoregressive (MAR) modelling of finite but arbitrary order can be used, with appropriate log transformation and model selection based on information criteria, in order to infer causal links between interacting populations. Causality is here understood in the sense of Granger, i.e. time series x Granger causes y if information on present and past values of x helps predicting future values of y. Recent nonlinear approaches such as convergent cross-mapping (CCM) have been put forward to alleviate potential issues of Granger causality (GC), notably due to the assumption of linearity in most implementations of GC. This has led to a relative skepticism about the ability of (log)-linear parametric methods to infer causal links in nonlinear dynamical systems. The merits of nonlinear and nonparametric models notwithstanding, the present results show that the demise of parametric and linear modelling has been somewhat exaggerated. Not only parametric MAR(p) models are able to infer causal links for a number of nonlinear (stochastic) systems, but they can even do so for datasets with highly nonlinear dynamical behaviour (e.g., limit cycles in real data, simulated chaotic dyamics). GC demonstrates, surprisingly, similar performances to CCM in a number of highly nonlinear cases. We further show that the (log)-linear framework can be extended to large interaction networks, provided some assumptions about the interaction structure. Finally, we discuss nonparametric, spectral and nonlinear extensions of Granger causality approaches that could be of use to ecologists.

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

Introduction

There is a diversity of viewpoints regarding the importance of interactions between species for predicting their joint population dynamics; a division that can be traced back at least to the Clementsian (integrated) vs Gleasionan (stochastically dominated) views of plant communities (Chase, 2003). Should one predict a species dynamics together with other species in the same community or separately? Measuring linkages between different species' population dynamics is a key statistical endeavour that has ramifications for fields as varied as coexistence theory (where the ratio intra/inter measures niche differentiation, Adler *et al.* 2010, 2018) or ecosystem-based fisheries management (Link, 2002; Pikitch *et al.*, 2004). There is, however, a continued tension within ecology and related fields regarding the importance of species interactions for community dynamics. This could arises not only from disagreements between experimental and field studies (Tuck *et al.*, 2018) but also from different definitions of what the term interaction even means (Berlow *et al.*, 2004; Wootton & Emmerson, 2005).

In many branches of statistical and theoretical ecology that have access to temporal records of population sizes, an interaction between species can be defined at the population level: two species i and j are deemed to interact if species i's population growth rate is affected by the temporal variation in the population density of species j or vice-versa (Berlow *et al.*, 2004). This definition, which we will use in the remainder of this paper, maps well both 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)

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

and statistical time series models (Ives *et al.*, 2003; Mutshinda *et al.*, 2009, 2011; Hampton *et al.*, 2013), that make similar assumptions for stochastic models in discrete time. In other words, an interaction between species can be defined as a causal influence from species j to species i in a stochastic dynamical system describing the community dynamics, which 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 statistically-orientated literature.

To quantify dynamic causation between species using time series data, ecologists have used so far a number of statistical models ranging from mechanistic to purely phenomenological. They include most notably multivariate autoregressive models or order one, or MAR(1) models (also called VAR(1) - vector autoregressive models). These are statistical multispecies generalizations of the Gompertz discrete-time single-species models (Ives *et al.*, 2003; Mutshinda *et al.*, 2009). MAR(1) can to a large extent infer linkages between species population dynamics (Hampton *et al.*, 2013; Certain *et al.*, 2018). MAR(p) models, with a maximum time

lag of order $p \geq 1$, generalize the MAR(1) framework familiar to ecologists and are interestingly linked to one of the most celebrated causality concept, the so-called Granger-Wiener causality (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 a causality concept that is strongly tied to the idea that the cause must precede in time the effect. This corresponds to the intuition of many biologists (Mayr, 1961) and especially ecologists, familiar with predators lagging 1/4 of cycle behind their prey population dynamics (May, 1973). There are other ways to define causality, through interventions notably (e.g. Pearl, 2009), which are very useful in experimental setting, but usually less practical in a dynamical setting (Aalen *et al.*, 2012). If a dynamical model for time series y can see its in-sample predictive ability improve by inclusion of time series x, we say that x Granger-causes y. This definition is rather general and does not specify, in general, any sort of model framework. It can in principle be applied to phenomenological and mechanistic frameworks alike, as well as in nonparametric settings (Detto *et al.*, 2012). However, GC testing is often used in contexts were the a priori information on community dynamics is limited, so that one needs to reduce the universe of possible model formulations to a handful of contrasted scenarios. In this context, GC is usually tested within the context of statistically friendly MAR(p) models for whom confidence intervals for coefficients, model selection, and other inferential tools are well understood. In an ecological setting, we use mainly log-linear models because the MAR framework applies to log-abundances (Ives *et al.*, 2003). Ecologists have therefore been using Granger causality many times (reviewed in Hampton *et al.*, 2013) without perhaps knowing it.

In the last decade, other methods such as Convergent Cross-Mapping (CCM; Sugihara *et al.*, 2012) have been put forward as a new way to infer interactions between species, and alleviate problems due to the limitation of linear autoregressive models (although, it should be noted that MAR(p) models are log-linear, which makes them essentially power-law models, hence able to model some nonlinearities).Sugihara *et al.* (2012) made a somewhat strong criticism of the Granger causality concept, which they deemed best suited for linear systems dominated by stochasticity, and therefore unfit to model strongly nonlinear dynamical systems with a strong deterministic skeleton. However, subsequent work has shown that even the simplest MAR(1) models can be surprisingly robust to nonlinearities (Certain *et al.*, 2018), uncovering interactions in the case of nonlinear interactions with a fixed point and even limit cycles. Moreover, CCM performed relatively similarly to linear Granger causality for a large array of 2-species simulation models (Krakovská *et al.*, 2018). Given the great conceptual and practical simplicity of MAR(p) models, there is therefore a need to better understand in which ecological scenarios linear GC can be a good approximation for interaction inference, and in which case more elaborate techniques, such as CCM, are warranted.

In this article, we first evaluate the performance of 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 in fact infer interactions suprisingly well in nonlinear cases. We then highlight intriguing parameter configurations and empirical case studies where Granger causality and Sugihara et al.'s convergent cross-mapping (CCM) either both fail or both work, which suggests that seemingly foreign causality concepts might in fact share some underlying mathematical similarities [To edit as a function of what we find]. Throughout, we take particular care to output both effect sizes and statistical significance of causal influences, and we discuss the meaning of uncertainties around causal effects, when p-values are not calibrated [To edit again as a function of what we eventually report]. We finally demonstrate that Granger causality / MAR(p) modelling can be scaled up to large interaction networks using appropriate model regularization techniques (LASSO-based).

Methods and models

In the following, we recall the basics of Granger causality concepts and MAR(p) - Multivariate AutoRegressive - modelling, which is the most common way (not the only one) to assess Granger causality. For completeness, we also present some common nonlinear alternatives to MAR(p) modelling which stays within the Granger causality purview in Appendix . We describe shortly thereafter convergent-cross mapping (Sugihara *et al.*, 2012), which is implemented as well but 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

$\mathbf{x} = (x_t)_{t\in[|1:T|]}$Granger-causes $\mathbf{y} = (y_t)_{t\in[|1:T|]} \iff$ 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 time series autoregressions to explain time series y, one with only y and one with y and x:

$$y_t = \sum_{i=1}^{L} a_i y_{t-i} + \eta_t \tag{2}$$

$$y_t = \sum_{i=1}^{L} a_{1i} x_{t-i} + \sum_{i=1}^{L} a_{2i} y_{t-i} + \epsilon_t \tag{3}$$

Granger causality then occurs if $\sigma_\epsilon^2 < \sigma_\eta^2$. When more than two variables are considered, simple GC has to be differentiated from conditional GC. Conditional GC occurs whenever a third variable z is considered and corrected for.

When fitting a MAR(p) model, we typically estimate conditional GC. 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):

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

so that its $\mathbf{B}$ (interaction) matrix is defined by

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

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

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

We drop $\mathbf{u_t}$ which corresponds to a special case where a subset of the variables has a one-way causal impact. The condition for an interaction from species j to species i given all other species densities and variables then becomes, in a general MAR(p) setting (according to eq. 6,):

$$\exists b_{ij}^{(q)} \neq 0 \Leftrightarrow x_j \rightarrow x_i|(x_1, ..., x_{j-1}, x_{j+1}, ..., x_d) \tag{7}$$

where each time lag is indexed by q. This highlights immediately the high model dimensionality that can be attained with high temporal lag order p and high number of species. Conversely, direct or pairwise GCausality testing between x_i and x_j is assessed through a bivariate autoregressive model for each (i, j) pair, and therefore uses a considerably lower-dimensional model. Direct GC testing requires, however, some false discovery correction to attain meaningful statistical significance (Mukhopadhyay & Chatterjee, 2006). In the following, we use a Benjamini-Hochberg correction (Benjamini & Hochberg, 1995).

MAR(p) model fitting has been performed using the package `vars` in `R`, which uses ordinary least squares for estimation. We 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. 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 (Sugihara *et al.*, 2012) 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 $w_i = \ u_i / \sum_{i=1}^{E+1} u_j$, 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.

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 library size L if Y causes X. The surprising thing here being that prediction of 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 $\rho(\mathbf{x}, \hat{\mathbf{x}}|M_Y)$.

Even though there are no standard ways of computing associated p-values for the convergent cross-mapping skill ρ, several 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 different versions of this p-value can be computed depending on whether one uses sampling with replacement (the bootstrap) for the librairies or sampling without replacement (in which case $M_{Y,\mathrm{Lmin}}$ varies but not $M_{Y,\mathrm{Lmax}}$).

- for periodic signals, or, more generally, a shared forcing driver, spurious causality can be found between time series. We compute surrogate time series which keep the periodicity of the signal but modify its residuals so that the 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, $p(X \nrightarrow Y) = \frac{1}{n}\sum_{i=n}^{n} \mathbf{1}_i\left((\rho(\mathbf{x}_{\mathrm{real}}, \hat{\mathbf{x}}_{\mathrm{real}}|M_Y) < \rho(\mathbf{x}_{\mathrm{surr}}, \hat{\mathbf{x}}_{\mathrm{surr}}|M_Y)\right)$
- for large simulations (10 and 20 species) without a confounding abiotic driver, we also computed a surrogate-based p-value, but surrogates were only computed by a random permutation of the values within the time series.

The analyses have been performed in R [3.4.4] 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 library size depending on the length of the time series . The libraries were obtained with random draws without replacement from the original time series.

Granger causality in high-dimensional models

Full MAR(p) model fitting is highly impractical for high-dimensional models, unless very long time series are considered or special constraints are implemented to reduce the dimensionality (Michailidis & d'Alché Buc, 2013). If there are d species and p timelags, a $d \times d \times p$ dimensional model needs to be fitted to the data. For instance, let us imagine that a system involving 10 species with at least one cycling species is considered (and its cyclicity is probably not induced by the species in the network). To model it properly, we need $p = 2$ (long, regular cyclic behaviour in a one species AR(p) model requires $p > 1$). We then have 200 parameters in the interaction $\mathbf{B}^{(q)}$ matrices only. Even a MAR(1) model will have 100 elements, and therefore would be impossible to fit properly without a set of time series of length above 100. 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 for time series of long yet reasonable length (100 to 300 timesteps), we have considered two solutions:

- Pairwise Granger causality testing with False Discovery Rate (FDR) correction (Benjamini-Hochberg), with a philosophy similar to Mukhopadhyay & Chatterjee (2006). This done by fitting bivariate MAR(p) models, test for Granger causality in both directions, and then re-adjust 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 approach to MAR(1) estimation would simply use the 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 (for instance, modular) in the network. We have therefore used a technique that explicitly accounts for network structure in addition to selecting coefficients with the LASSO (Charbonnier *et al.*, 2010).

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 which demonstrated two-way coupling (Jost & Ellner, 2000).

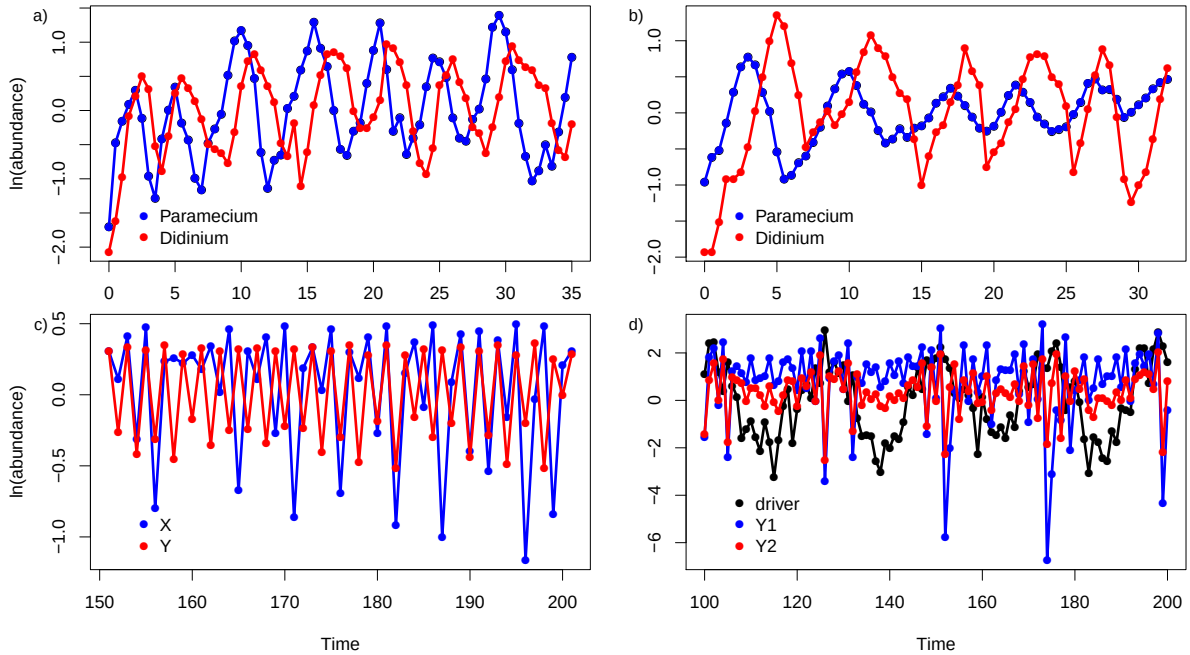

Figure 1: Time series of small-community models. Veilleux's predator-prey data are shown in (a) (dataset CC05) and (b) (dataset CC0375); an example simulation for the 2-species chaotic model is shown in panel (c) and a simulation of the competition model including an environmental driver is illustrated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

Deterministic chaos in two-species competition models

Here we use the same model as in Sugihara *et al.* (2012), a two-species discrete-time logistic competition model

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

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

The models are, as in Sugihara *et al.* (2012), in the chaotic regime, which constitutes a strong test of the log-linear MAR(p) framework. The only thing that is varied is the initial condition, which is taken at random 500 times. Although we acknowledge that "mirage correlations" can occur in some datasets, we aimed at reproducting a realistic sample of what this model can provide, as there are no justifications to favor one specific set of initial conditions (outside of illustration purposes). The sample size is taken to be $t_m = 300$ as in Sugihara *et al.* (2012), with 500 time steps discarded as burn-in.

Because a method that finds no interactions whenever absent is as important as one that finds interactions whenever they are present, we also consider the following deterministic "null competition" model:

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

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

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

Two-species stochastic and nonlinear dynamics, including environmental drivers

First we consider a stochastic two-competition model, with Lokta-Volterra interactions in discrete time and a Ricker type of multispecies density-dependence.

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

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

This case was already investigated in Certain *et al.* (2018), but including an environmental driver on species 1. Then we consider a variant of this model 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)

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

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

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

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

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

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, the error $\sigma^2 = 0.1$ and the interaction matrix is defined to be

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

We use the same interaction matrix with Gompertz (MAR(1)) dynamics, and run 25 simulations over 500 time steps with different initial conditions. This model has a (stochastic) Lyapunov exponent of +0.33, thus being clearly in a noisy chaos regime (Ellner & Turchin, 2005).

We slightly modified this model to scale it up to 20 species. Its structure is still fairly modular (eq. 20)

$$
B = \begin{pmatrix}
b_{1,1} & b_{1,2} & b_{1,3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b_{2,1} & b_{2,2} & b_{2,3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b_{3,1} & b_{3,2} & b_{3,3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b_{4,1} & b_{4,2} & b_{4,3} & b_{4,4} & b_{4,5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & b_{5,4} & b_{5,5} & b_{5,6} & b_{5,7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & b_{6,5} & b_{6,6} & b_{6,7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & b_{7,5} & b_{7,6} & b_{7,7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{8,8} & b_{8,9} & b_{8,10} & b_{8,11} & b_{8,12} & b_{8,13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{9,8} & b_{9,9} & b_{9,10} & b_{9,11} & b_{9,12} & b_{9,13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{10,8} & b_{10,9} & b_{10,10} & b_{10,11} & b_{10,12} & b_{10,13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{11,8} & b_{11,9} & b_{11,10} & b_{11,11} & b_{11,12} & b_{11,13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{12,8} & b_{12,9} & b_{12,10} & b_{12,11} & b_{12,12} & b_{12,13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{13,8} & b_{13,9} & b_{13,10} & b_{13,11} & b_{13,12} & b_{13,13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{14,11} & b_{14,12} & b_{14,13} & b_{14,14} & b_{14,15} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{15,14} & b_{15,15} & b_{15,16} & b_{15,17} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{16,15} & b_{16,16} & b_{16,17} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{17,15} & b_{17,16} & b_{17,17} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{18,18} & b_{18,19} & b_{18,20} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{19,18} & b_{19,19} & b_{19,20} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & b_{20,18} & b_{20,19} & b_{20,20}
\end{pmatrix} \tag{20}
$$

For the 20-species model, we also compare Ricker and Gompertz dynamics for 25 different simulations over 1000 time steps, but in this case, coefficients are drawn from a random distribution (eq. 21 and eq. 22).

$$
b_{i,j} = b_{min} + (b_{max} - b_{min})\text{Beta}(2,2) \tag{21}
$$

with the bounds of the interaction coefficient selected as

$$
(b_{min}, b_{max}) = \begin{cases}
(0.05, 0.1) & \forall i \neq j, \text{ with probability 0.2 (positive interaction)} \\
(-0.2, -0.1) & \forall i \neq j, \text{ with probability 0.8 (negative interaction)} \\
(-0.8, -0.3) & \forall i = j
\end{cases} \tag{22}
$$

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 before analysis.

Results

In each section, we apply both GC/MAR(p) modelling and CCM.

Real data: Veilleux's predator-prey cycles

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") demonstrate convincingly that the "no GC" hypothesis can be rejected, for both datasets (1).

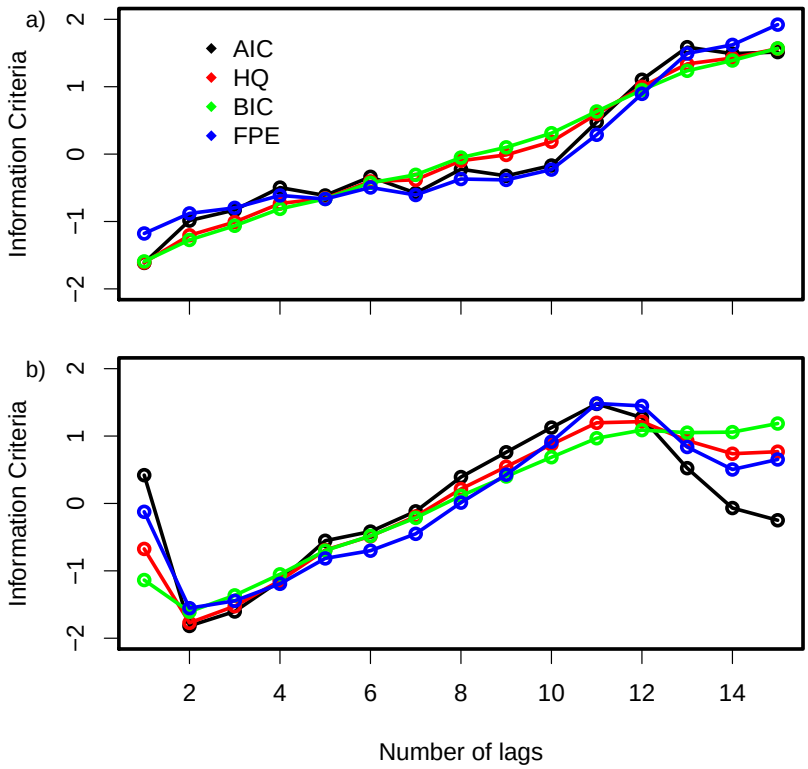

Figure 2: Results of model information criteria vs. lag order for the predator-prey data, for the two datasets.

Dataset	CC = 05	CC = 0.375
Lag p in VAR(p)	1	2
$1 \to 2$	2.79×10^{-11}	0.0409
$2 \to 1$	1.76×10^{-14}	0.0464

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

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

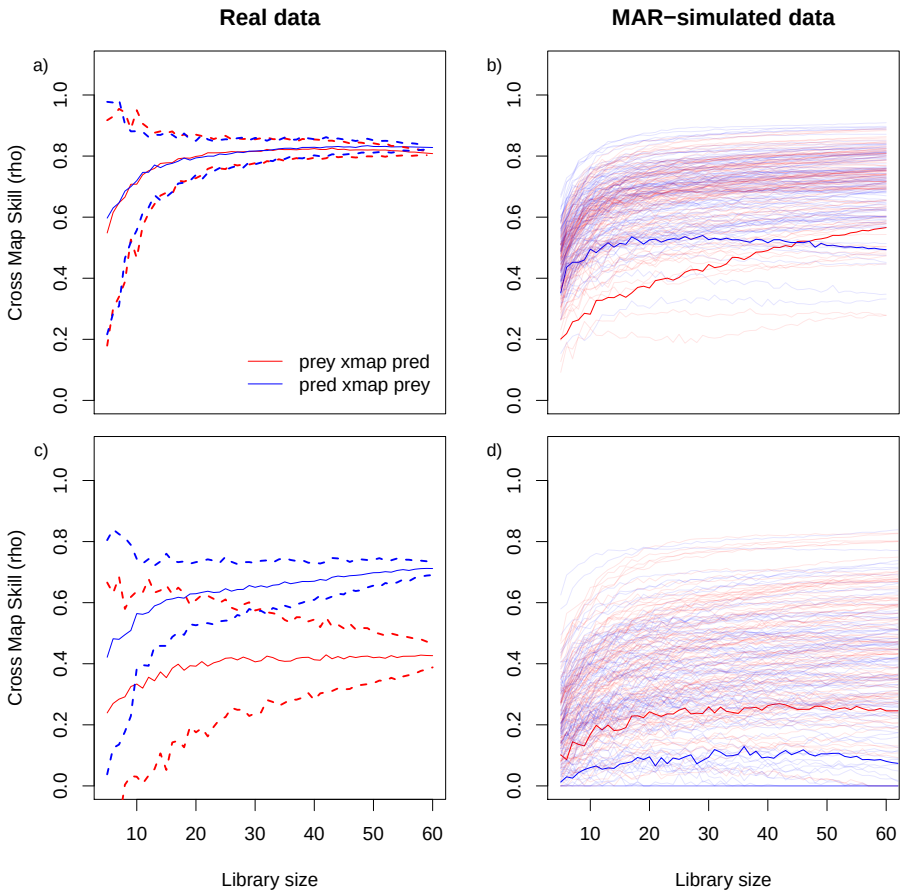

Figure 3: Convergent cross-mapping for Veilleux's CC05 dataset (a and b) and CC0.375 dataset (c and d). Dotted lines represent the confidence bands, obtained by bootstrapping.

Deterministic chaos in two-species competition models

In the 2-species chaotic competition model high-order temporal lags tend to be selected (Fig. 15) despite the single timelag considered in the simulation model (i.e., higher nonlinearity is expressed as high-order lags). Despite this potential overparameterization, the GC tests show that causality is detected for most timelags (including the optimal one, $p = 7$) whenever causality is present. Further, the tests are able to reject the null hypothesis of no GC when GC is not present (Fig. 4). 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 $2 \rightarrow 1$ (Table 2). Moreover, both methods produce some false positives (around 10% when considering p-values, which is to be expected since we test at the 10% level), though these rates are somewhat higher in one causal direction for CCM (up to 30% 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 (Fig. 5). 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.

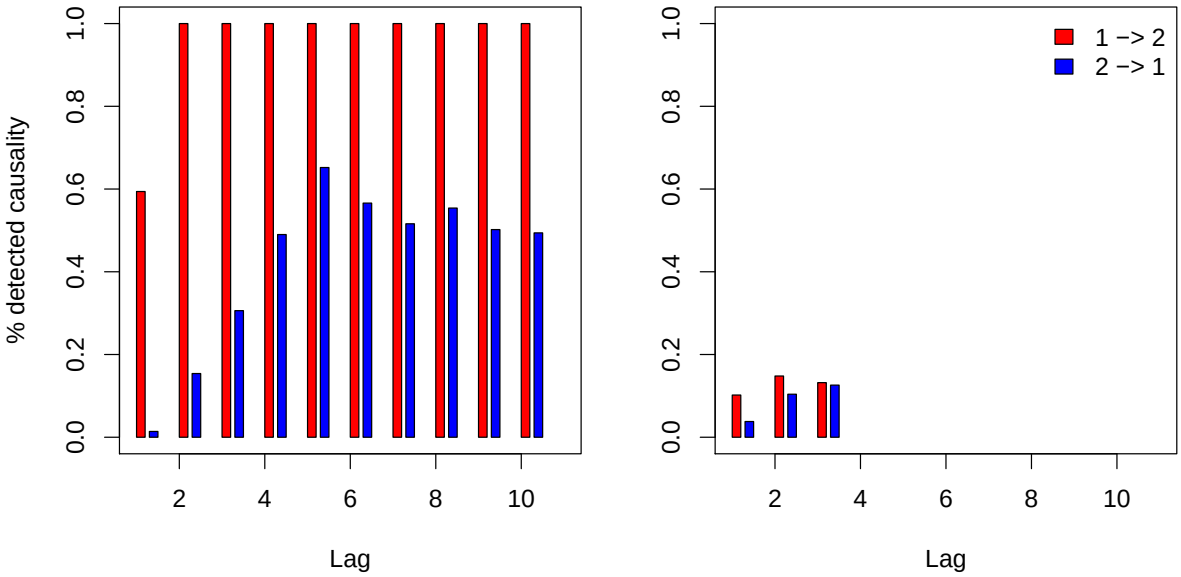

Figure 4: Proportion of detected Granger-causality, at the 10% significance threshold, over 500 chaotic simulations with (left) and without (right) actual interactions between species, depending on the number of time lags taken into account (x-axis). Without interactions, the optimal lag is 3 and the Wald test cannot be performed for p>3.

Interactions	Granger causality			CCM				
	pval<0.1	log-ratio>0.04	both	pval<0.1	rho>0.1	rho>0.2	both0.1	both0.2
With								
$1 \to 2$	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
$2 \to 1$	50.6%	69.8%	50.6%	56.8%	54.8%	29.4%	54.2%	29.4%
Without								
$1 \to 2$	14.2%	100%	14.2%	14.8%	8.2%	0.4%	8.2%	0.4%
$2 \to 1$	12%	2.6%	2.6%	30.8%	29%	12.8%	28.8%	12.8%

Table 2: Proportion of simulations with Granger-causality or CCM between x and y over 500 simulations, for the chaotic 2-species competition model.

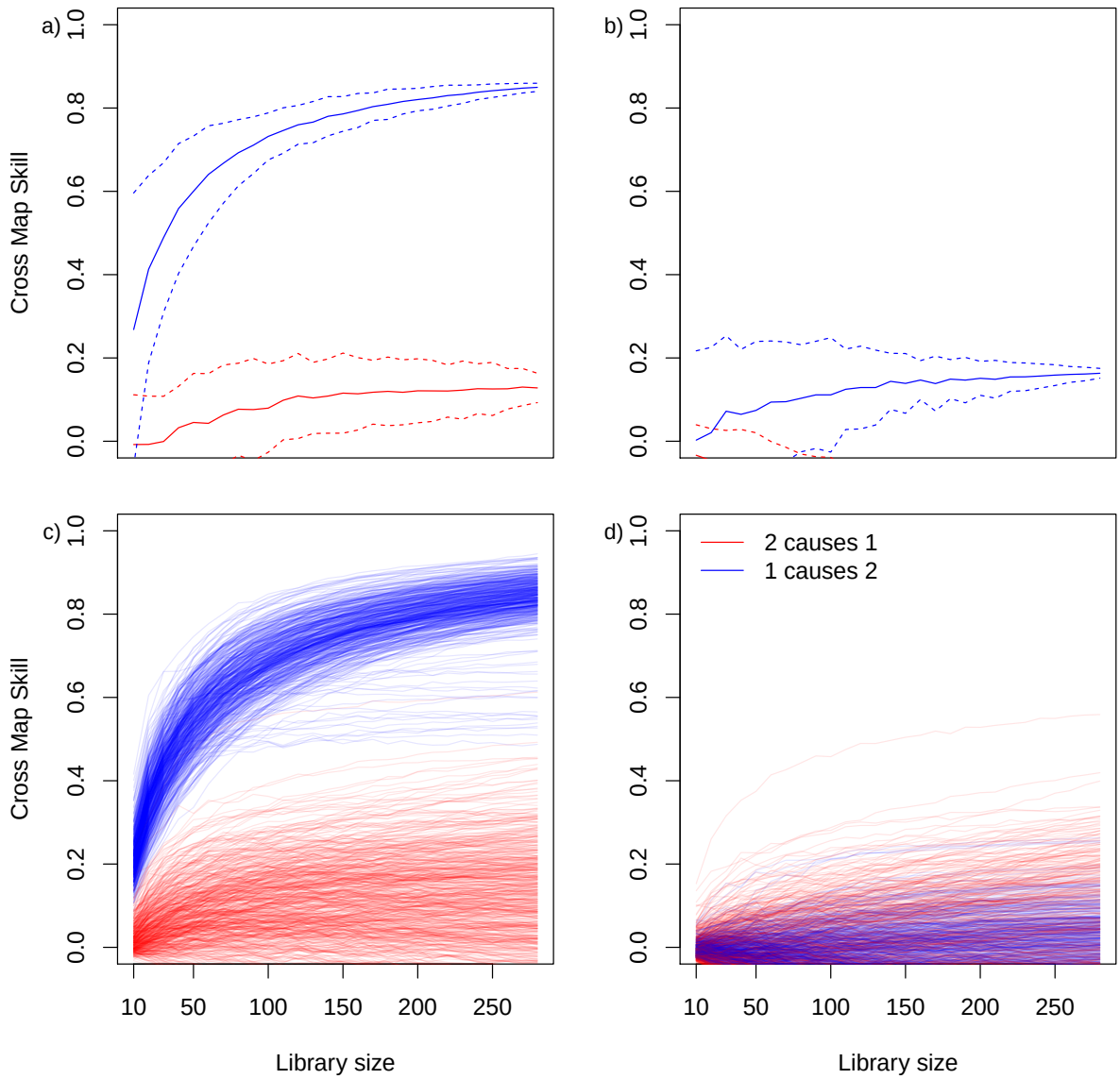

Figure 5: Convergent-cross mapping on simulated deterministic data, with (left) and without (right) competition between the two species. On the top row, one simulation with (a) and without (b) interactions with associated confidence bands; bottom row, cross-map skill (ρ) for 500 simulations. [**This figure may go to the Appendix**]

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 in most cases causality and non-causality. CCM has slightly better rates of interactions found (no falses negatives) but also a little more false positives, while GC is a

little more conservative, especially when considering a threshold for the logarithm of the sum of squares ratio (Table 3).

Method	Granger causality			CCM				
Thresholds	pval<0.1	log-ratio>0.04	both	pval<0.1	rho>0.1	rho>0.2	both0.1	both0.2
With interactions								
$1 \to 2$	98.4%	94 %	94 %	100 %	100 %	100 %	100 %	100 %
$2 \to 1$	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Without								
$1 \to 2$	12.8%	0.2%	0.2%	11.6%	10.6%	1%	10%	1%
$2 \to 1$	9.6%	0.4%	0.4%	12.8%	12%	1.2%	11.4%	1.2%

Table 3: Proportion 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).

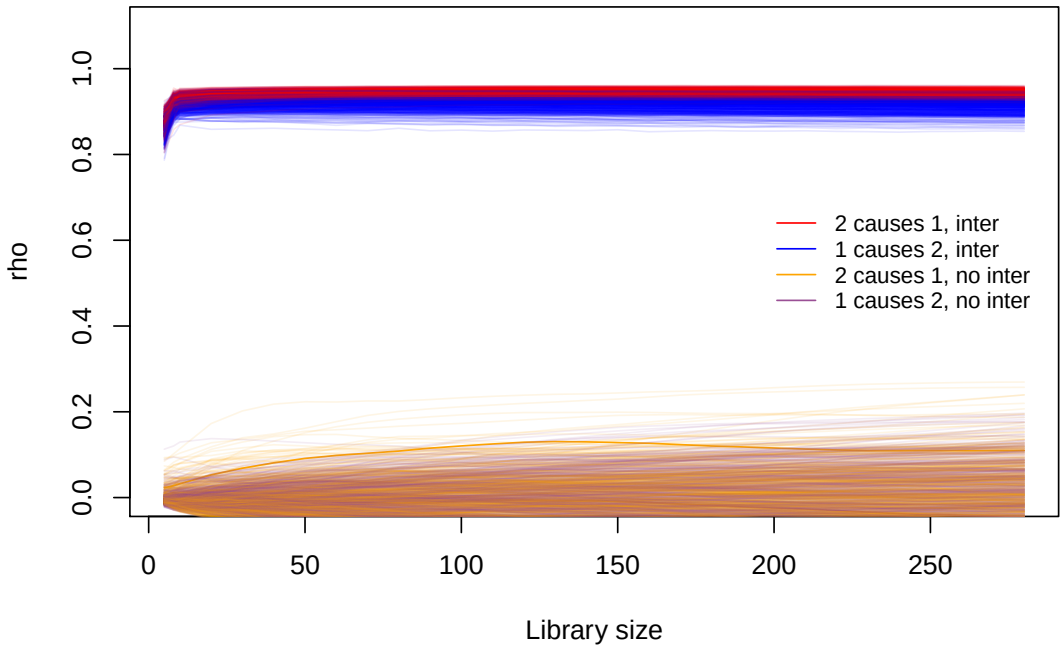

Figure 6: Convergent-cross mapping on simulated stochastic data, with (red, blue) and without (orange, purple) competition between the two species, over 500 simulations. **[This figure may go to the Appendix]**

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

This considerably more complex case where the two species are influenced by a shared environmental driver (e.g., temperature) 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. 7). Regarding interactions, CCM was better at uncovering interactions that were present in this case, GC had a good performance for the strong interaction $2 \to 1$ but not the reserve $1 \to 2$. Both GC and CCM had difficulties indicating non-causality in this case (when there are no interactions), and indicated false positives twice above the level allowed (≈ 20 intead of 10%, Table 4). Thresholding using 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.

CCM was considered with signifiance assessed through seasonal surrogates time series, which clearly improved its power to detect interactions, but we still had spurious causalities in CCM when no interactions were present (Fig. 16). This is therefore a scenario where avoiding false causalities is difficult for both GC and CCM.

Method	GC	pairwise		GC	conditional		CCM	seasonal	surrogate
Thresholds	pval<0.1	log-ratio>0.04	both	pval<0.1	log-ratio>0.04	both	pval<0.1	rho>0.2	both
With									
$1 \to 2$	25.8%	21.4%	19%	24.6%	20.8%	19.4%	94.8%	94.8%	94.2%
$2 \to 1$	92.4%	88.6%	88.2%	87.6%	84.2%	83.6%	97.2%	96.6%	96.6%
Without									
$1 \to 2$	23.8%	18.6%	17.4%	24.8%	19.2%	18.2%	23.6%	27.6%	22%
$2 \to 1$	23.4%	17.4%	16.4%	21.6%	16.8%	15.6%	20.6%	16.2%	16%

Table 4: Proportion of simulations with Granger-causality or CCM between x and y over 500 simulations for a model with 2 species and a driver (temperature) [Causalities related to temperature - not interactions - for CCM are in the Appendices for now.]

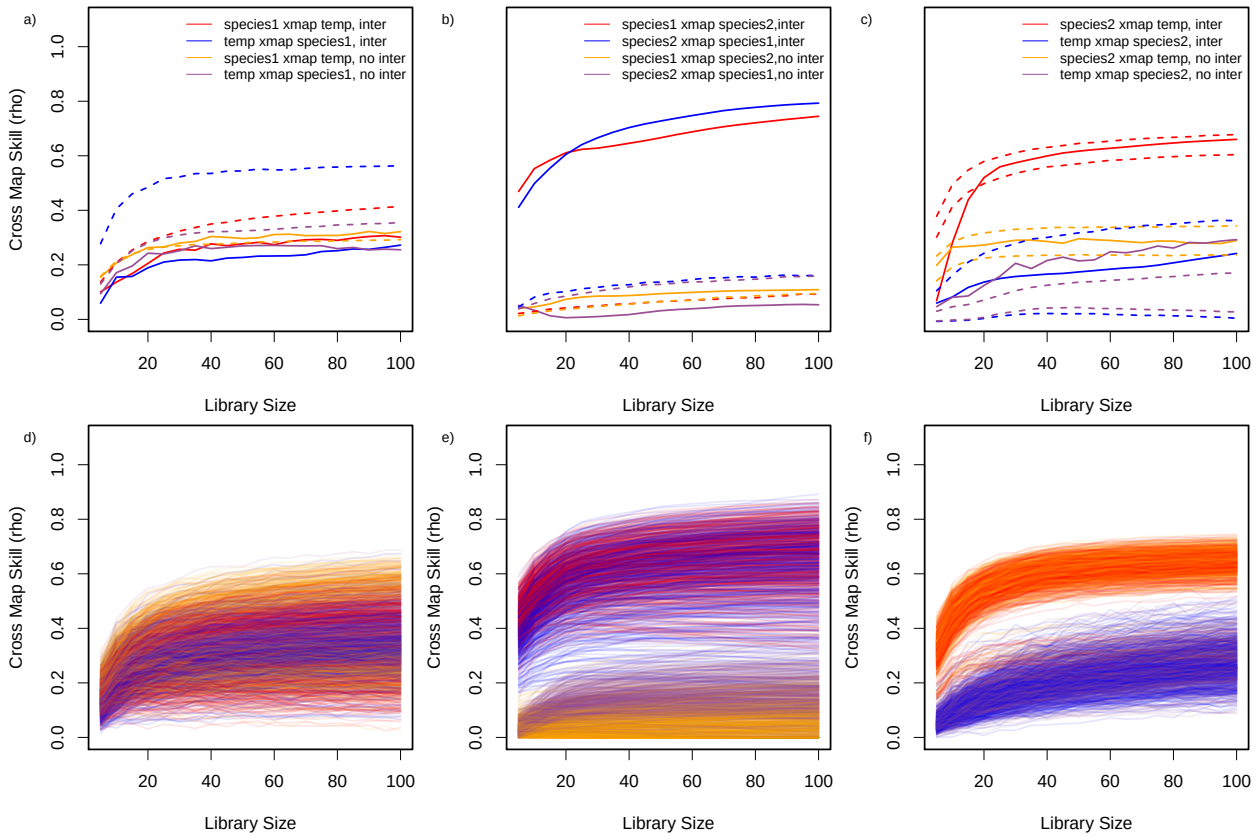

Figure 7: Convergent cross-mapping for the two species forced by an environmental driver (denoted as temp), when interactions are present (blue, red) and when interactions are absent (purple, orange), for 500 simulations. Dashed lines indicate the 10% interval for rho-values obtained from surrogate time series, i.e., time series that have the same seasonal forcing but whose cross-correlations are altered. [**This figure may go to the Appendix**]

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. **??**), 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 SXX) because the basic LASSO-penalized VAR(1) models poorly identify modular interactions webs (Charbonnier *et al.*, 2010). The recall, or true positive rate, that records how many true positives are identified as such, is almost always above 60% for the LASSO-based method (siMoNe, Chiquet *et al.*, 2008; Charbonnier *et al.*, 2010) and up to 80%, which is a relatively good performance. Surprinsingly, we found that pairwise (direct) Granger causality testing was quite efficient, in this case more than the LASSO-based method. One issue though is that the level of significance has to be adjusted (we use a global α= 20%), which is always somewhat subjective (but there are also always several tuning parameters in LASSO-based inferences, so some degree of subjectivity is everywhere).

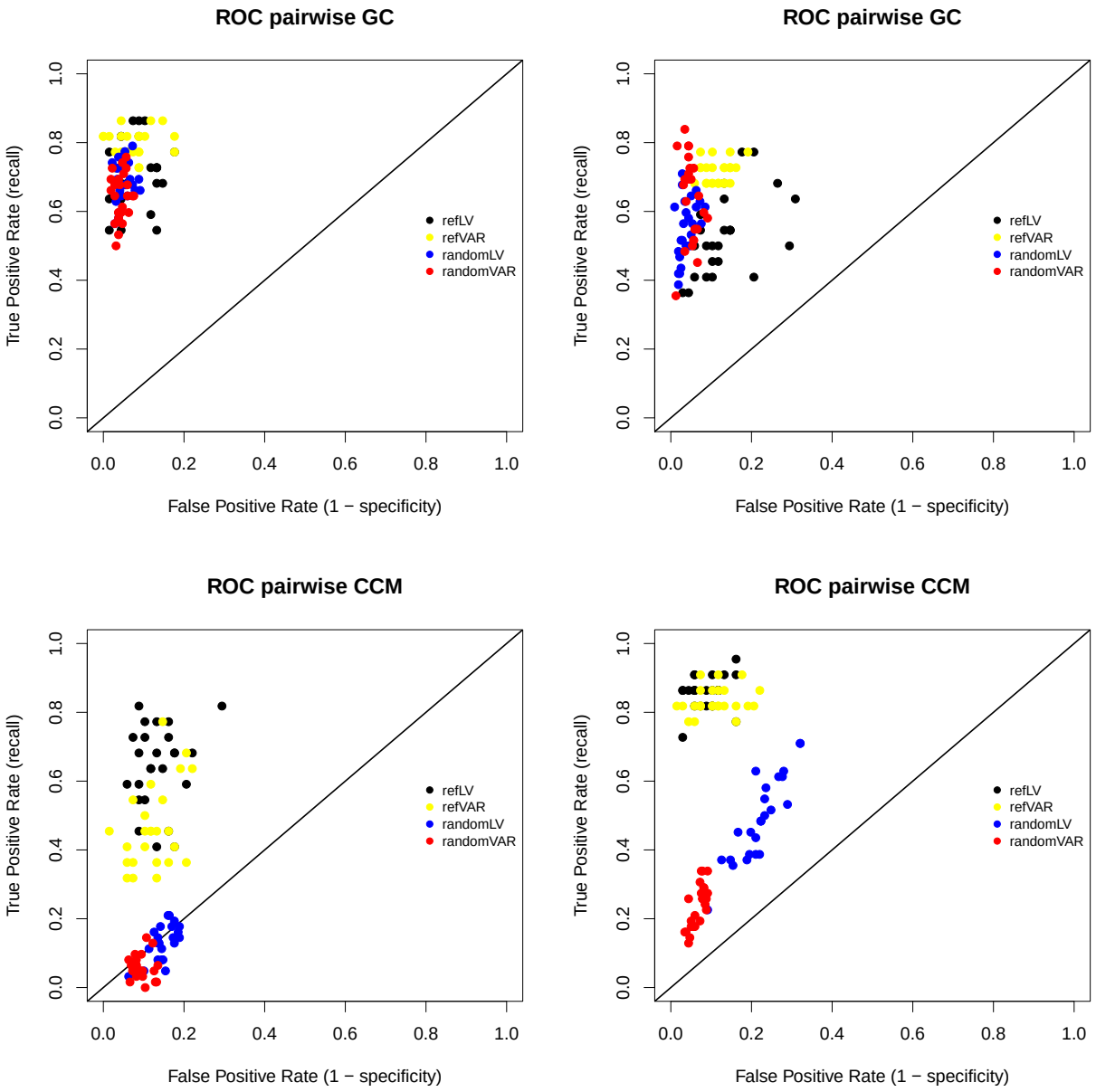

Figure 8: ROC curves for the 10 and 20-species model using Granger Causality (top) and CCM method (bottom), with different ways of computing the causality. For GC, left is pairwise GC; right is LASSO (SIMoNe). For CCM, we used Cobey-Baskerville p-values on the left, our "shuffled"-p-values on the right panel. refLV and refVAR refer to the 10-species system, for which one chaotic parameter set with many initial conditions is considered, while randomLV and randomVAR refer to the 20-species model (which is a perturbed fixed point, with negative SLE). The VAR case is always the VAR(1) model associated to the Jacobian of the community matrix, hence a linearization in log-scale of the Lotka-Volterra model.

Comparing GC and CCM in "ideal" conditions, with the best-performing algorithms that we tested (pairwise GC with a Benjamini-Hochberg corrections and CCM with our 'shuffled' p-values) reveals that they reconstruct similar networks for the fairly nonlinear (positive SLE) 10-species case (Fig. 9), 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).

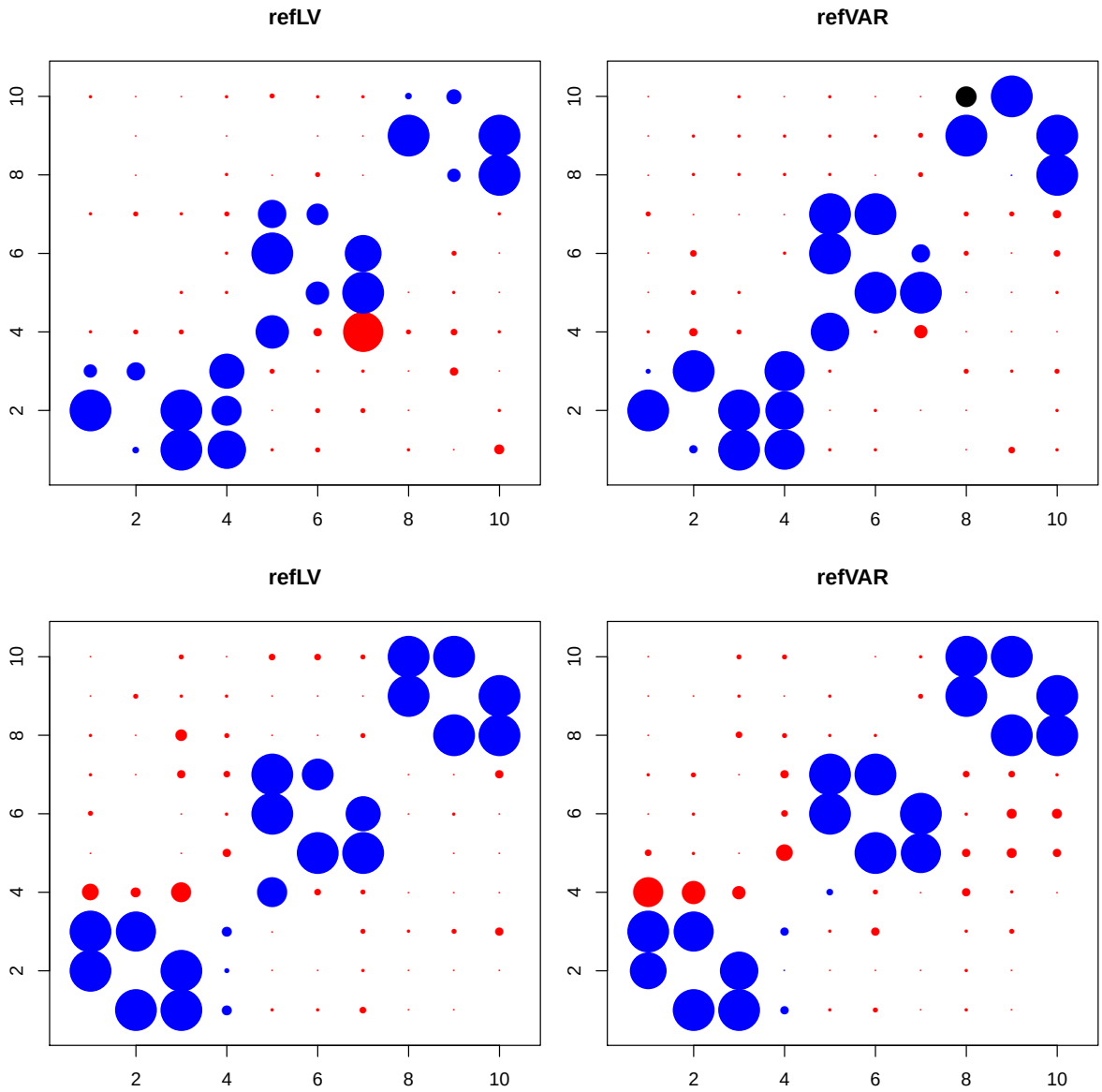

Figure 9: Interaction matrices obtained from pairwise GC (top) or 'shuffling' CCM (bottom) for 10-species communities. Blue circles are the true positives, red circles are false positives and black circles are false negatives. For the true and false positives, the size of the circles is proportional 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 provide quite a number of false positives (Figs. 8c,d and 18), even with the 'shuffled' p-values that worked very well for smaller-dimensional examples).

Discussion

We have shown above that MAR(p) modelling and Granger causality testing are fairly robust to nonlinearities in ecological dynamics, when applied on the appropriate logarithmic-abundance scale, and combined with model selection by information criteria. This is true for many nonlinear simulation models, including deterministic chaos. 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 (stochastic competition) and from cases where both methods seem to fail to some degree (i.e., two competing species forced similarly by a shared environmental driver). Therefore, an important conclusion from our study is that both Granger causality and CCM are able to yield reasonable inferences on similar datasets, unless there is a very high degree of confounding (see also Cobey & Baskerville, 2016, on the latter point).

Moreover, we use here false discovery rates 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 almost as good as the penalized MAR(1) models in finding the interaction network. We elaborate on these results and possible explanations below.

Is Granger causality a useful concept for highly nonlinear dynamical systems, and why?

Sugihara *et al.* (2012) suggested that Granger causality would work well for simulated (log)-linear systems while CCM would work well for nonlinear dynamical systems. Reality, however, is less clear-cut. First, the analyses performed by Sugihara *et al.* (2012) et al. on the Veilleux and other datasets seem to rely on a slightly dated model selection procedure (pre-information criteria) which produced overparameterized autoregressive models with very long lags. Reanalysing 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 predator-prey datasets.

We also reported above on results where even deterministic chaos can, surprisingly, be very well-identified

by MAR(p) models / GC inference (we expected CCM to completely dominate the scores and GC to fail entirely, but GC provided very similar performances, e.g. 95% of true interactions found where CCM finds 100%). Thus GC testing can be useful for highly nonlinear systems, and it usually tends to approximately correct rates of false positives when the null hypothesis of no interactions is true (an important aspect as well). Moreover, 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. Of course, as CCM has been developed specifically from nonlinear dynamical systems theory (and relies upon state-space reconstruction of chaotic models), it seems more appropriate for highly nonlinear and weakly stochastic cases, while MAR(p) models are best performers in weakly nonlinear and more stochastic situations. Nonetheless, it was very surprising to see how well MAR(p) models can perform in deterministic chaos in a number of cases, unlike what could have been expected from the previous study of Sugihara *et al.* (2012). There is most definitely a very broad degree of overlap in the models and data for which Granger causality (even in its classic log-linear form) and convergent cross-mapping can be applied.

Several recents attempts to perform nonlinear inference of interactions support our findings. Suzuki *et al.* (2017) use stochastic GLV models and a model selection algorithm derived from Fisher & Mehta (2014) to infer interactions. Even though they use the S-map (a derivative of the simplex reconstruction used to perform CCM) to reconstruct a dynamical system, because their evaluation of causality is based on whether in-sample prediction improves, they actually perform a nonlinear Granger inference. The neuroscience, gene regulation networks and related dynamical systmems literatures are awash with nonlinear or nonparameteric GC inferences, most often confirmed by simulations of nonlinear stochastic systems (Marinazzo *et al.*, 2008; Dhamala *et al.*, 2008; Papana *et al.*, 2013; Yang *et al.*, 2017). Hence there is excellent evidence that either linear GC after log-transformation (i.e., the associated autoregressive model is in power-law form), as done here, or nonlinear GC inference is possible to infer causality in nonlinear stochastic systems of interacting species. As shown by our analyses for large interaction webs, GC can even be extended to quite large webs, both in a pairwise GC context and in a conditional GC context (full model fitting, using the LASSO and clustering for model reduction).

When testing for causality in high-dimensional models, we had to rely on p-values rather than effect sizes (Fig. 8), except when the LASSO-based method was considered. This is because, due to the additive effects of many species on the growth rates, the effect sizes due to the causal influence of 1 species on the focal species become very small, and therefore very difficult to tune (tuning on effect sizes was, by contrast, possible in low-dimensional systems). Both ρ in cross-mapping and the log-ratio of sum of squares in a Granger context are vulnerable to this issue. Despite this, both pairwise GC and pairwise CCM performed quite well. Although we found CCM to perform a little less well in the 20 species example, it might be

possible for somebody better used to this method to make it work more reliably and decrease the number of false positives.

The problem of uncalibrated p-values

Overall, we found that both CCM and GC performed similarly, in the sense that they managed to find most of the time interactions when there where interactions and no interactions when they were no interactions. These sort of 'qualitative agreements' are usually the way such results are reported in the ecology and physics literature (e.g., see recently Krakovská *et al.*, 2018). This lacks some statistical flavour, however.

A fact exemplified by all the Tables in our manuscript is 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 the 10% level of the test employed. A test that produces accurate p-values should have, in this case, p-values below the threshold that are exactly 10 percent. Of course, in the Granger case, the model that generated the data (nonlinear) and the model used to analyze it (log-linear) is not the same, but still we could have expected some robustness. These results are not completely surprising though, as confidence intervals for MAR(1) models 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, perhaps using transfer entropy (Schreiber, 2000, see Paluš, 2008; Amblard & Michel, 2013; Papana *et al.*, 2013, for reviews as well as Appendix XX), could be of use to improve causality detection by obtaining more exact p-values under the null.

Why exactly these p-values issues are true for CCM as well is unclear, as we crafted surrogate-based tests, and we reckon that much work on more formal statistical inference for CCM remains to be done. Cobey & Baskerville (2016) proposed a p-value that made logical sense, but we found unfortunately that although very sensible in theorty, this p-value was not working very well in practice. We considered different surrogate-based p-values and chose the best-performing ones (Figs. 12,13 in Appendices), but in some cases - with a confounding abiotic driver or many species - this was not completely satisfactory.

One idea could be to combine both worlds and use surrogate-based nonlinear Granger-causality (Schreiber, 2000; Schreiber & Schmitz, 2000; Paluš, 2008).

How can Granger causality and convergent cross-mapping yield similar inferences?

Here, we would like to go back to the heart of the issue that Sugihara *et al.* (2012) highlighted, causality reversion in nonlinear dynamical systems. The standard Granger causality holds that whenever a model

$Y_{t+h}|(Y_t, X_{tt\in A_t})$ better predicts the observed time series (y_t) than a model $Y_{t+h}|(Y_{tt\in A_t})$, then x is causal for y. CCM instead holds that whenever $\rho(x, \hat{x}|M_Y)$ increases strongly with the library size L use 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 recontruct 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+h}
- CCM compares knowledge about Y_t vs. no knowledge about Y_t in prediction of X_t.

Using a standard autogressive model, the equivalent of a CCM test $x \to y$ would be predicting X_t by a model $X_t|Y_t$ vs X_t. There is no conditionality upon past X_t values in the prediction step of the algorithm. Thus there is no causality reversion that is intrinsic to nonlinear dynamic testing: GC and CCM are simply two different types of causality testing that are based upon *different assumptions on the conditioning set* and *ways to select models*. We therefore conclude for these methods and concepts to work relatively similarly, they must share some underlying similarities that not yet evident to theoretical and statistical ecologists. More mathematical research on the possible connections and differences between these methods is obviously needed to better see in which scenarios each should be favored (if one of those should be favoured – they could as well be combined).

Conclusion: Going further with causal inference of nonlinear dynamical systems

Based on our simulations, complemented with those of the neuroscience literature (Ding *et al.*, 2006; Chen *et al.*, 2006; Barnett & Seth, 2014; Papana *et al.*, 2013; Marinazzo *et al.*, 2008), Granger causality, in its log-linear (in this paper) or nonlinear varieties (Marinazzo *et al.*, 2008; Yang *et al.*, 2017) varieties, is found to be appropriate to infer ecological interactions. Convergent cross mapping (Sugihara *et al.*, 2012) is another interesting method to infer interactions, which has been found here to perform similarly to log-linear Granger causality in most cases. Although it has been shown that some scenarios like seasonal forcing might render interaction inference difficult (Cobey & Baskerville, 2016) with CCM, we think that as always details of implementation are important, thus it is important not to disqualify a method in the early stages of its development. Previous simulation work by Krakovská *et al.* (2018) also concluded that GC and CCM had similar performances, although in their setting of extremely long simulations, and in absence of p-value calibration, other methods were performing better. Importantly, 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., a few hundreds time steps).

From an almost sociological viewpoint, we conclude that recommending to abandon established statistical methods like GC (Sugihara *et al.*, 2012), or linear modelling after transformation more generally (REF?), in favor of promising yet not fully developped statistical methods may not help ecological science to get the best statistical inference tools. More fruitful methodological development might highlight the pros and cons of new developments in statistical inference under a broad range of simulated scenarios.

From a statistical perspective, looking at the various implementations of GC and CCM, it seems that the hardest methodological choice to make is never linear vs nonlinear but instead boils down to:

- The details of the test implementation: we have shown that a bad selection of the lag order of autoregressive models results in wrong GC inference, while CCM is sensitive to the p-value definition. In other words, the devil is always in the details of the test.
- The conditioning set, i.e. the information that is considered to be known for the prediction (Eichler, 2013). Strategies to better understand how to choose the conditioning set when doing 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 (Runge, 2018).

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.

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. & 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.

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.

Chase, J.M. (2003). Community assembly: when should history matter? *Oecologia*, 136, 489–498.

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., 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.

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

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., 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.

Dhamala, M., Rangarajan, G. & Ding, M. (2008). Analyzing information flow in brain networks with nonparametric Granger causality. *NeuroImage*, 41, 354–62.

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

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.

Fisher, C.K. & Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One*, 9, e102451.

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

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

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.

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

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.

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

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.

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.

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

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.

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.

Suzuki, K., Yoshida, K., Nakanishi, Y. & Fukuda, S. (2017). An equation-free method reveals the ecological interaction networks within complex microbial ecosystems. *Methods in Ecology and Evolution*, 8, 1774–1785.

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

Tuck, S., Porter, J., Rees, M. & Turnbull, L. (2018). Strong responses from weakly interacting species. *Ecology Letters*, 21, 1845–1852.

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

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.

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

Appendices and Supplements

Extensions of Granger causality

LASSO-based MAR(1) models

[Description to include here]

Transfer entropy and nonlinear Granger causality

Transfer entropy can be defined as

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

where $y^{t_m+1} = (y_2, ..., y_{t_m+1})$ and $y^{t_m} = (y_1, ..., y_{t_m})$. 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 take 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 GC method $x \to y_i|(y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_n)$. Quite a number of nonlinear Granger causality inference techniques then fall within this category (e.g., Marinazzo *et al.*, 2008; Paluš, 2008).

Additional results

Example simulation of 2-species stochastic Ricker model

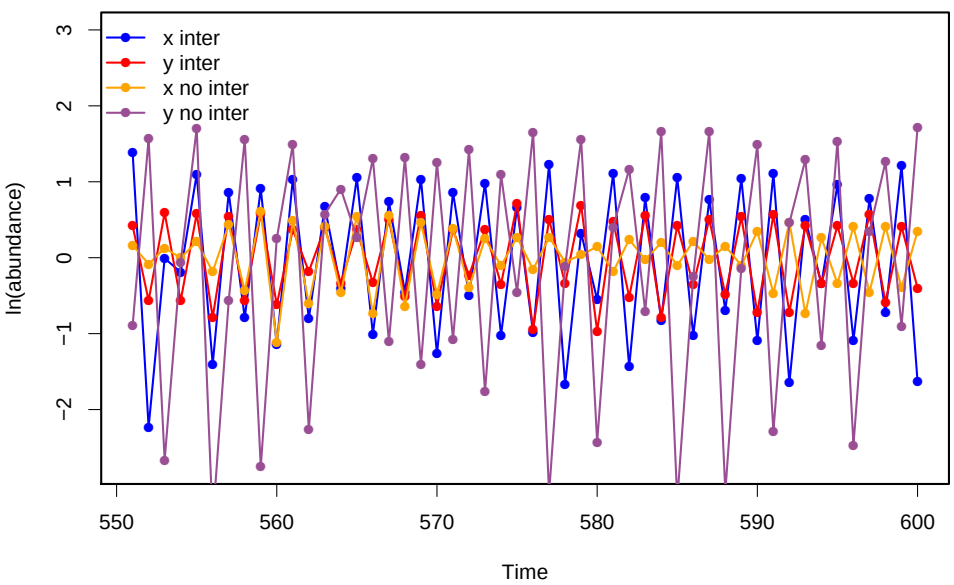

Figure 10: 2-species Ricker model [ref to eqs. of main text] with (blue = species 1, red = species 2) and without (purple = species 1, orange = species 2) competition.

Choice of p-values and additional thresholds

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 ($\frac{\sum_i |a_{1i}|}{L}$). We see on Fig. 11) 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.

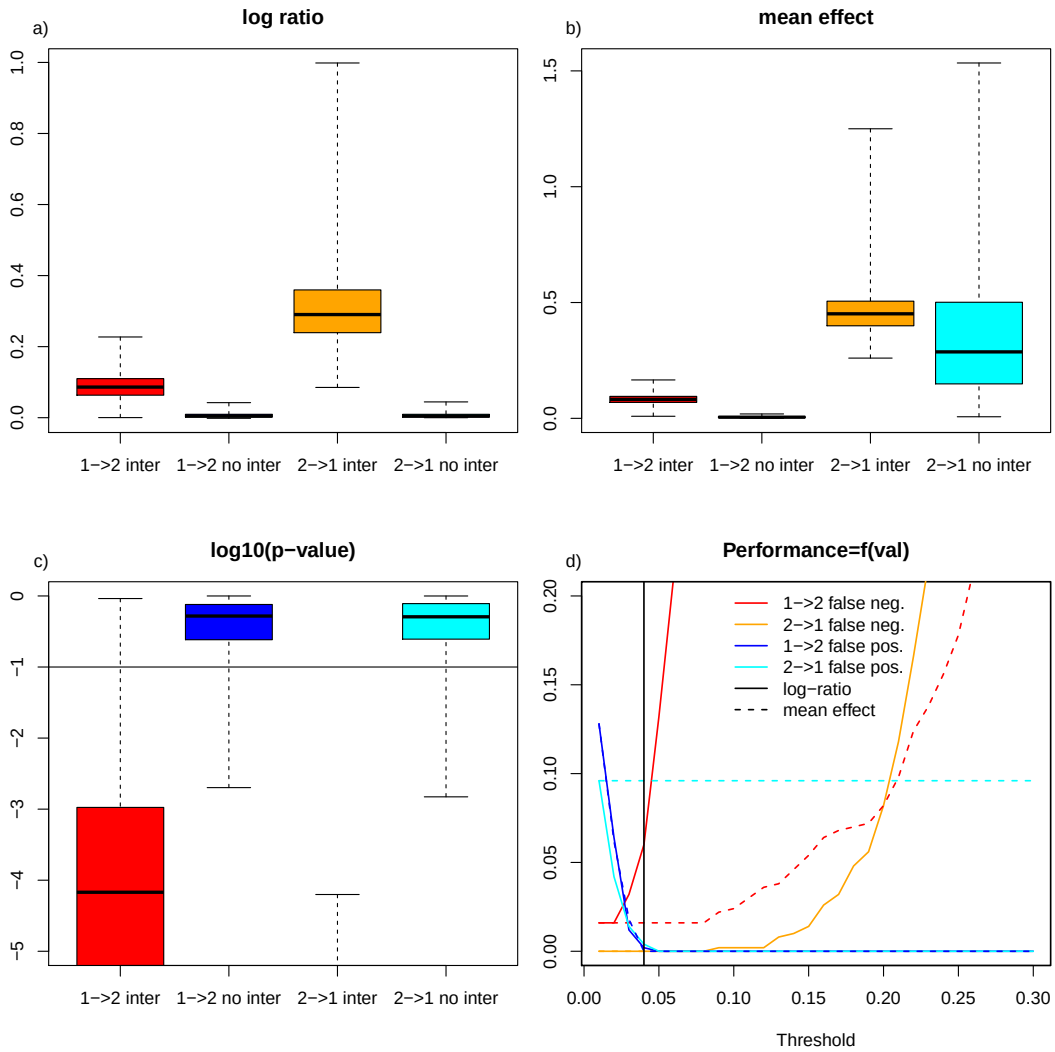

Figure 11: Comparison of methods to determine Granger-causality between two variables in a stochastic model. Log ratio of residuals sum of squares (a) and average effect of the causal species (b) are compared, and the proportions of false negatives (red and orange) and false positives (blue and cyan), 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 above). 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). We also examined the effect of putting a threshold on the value of ρ. We see on Fig. 12 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.

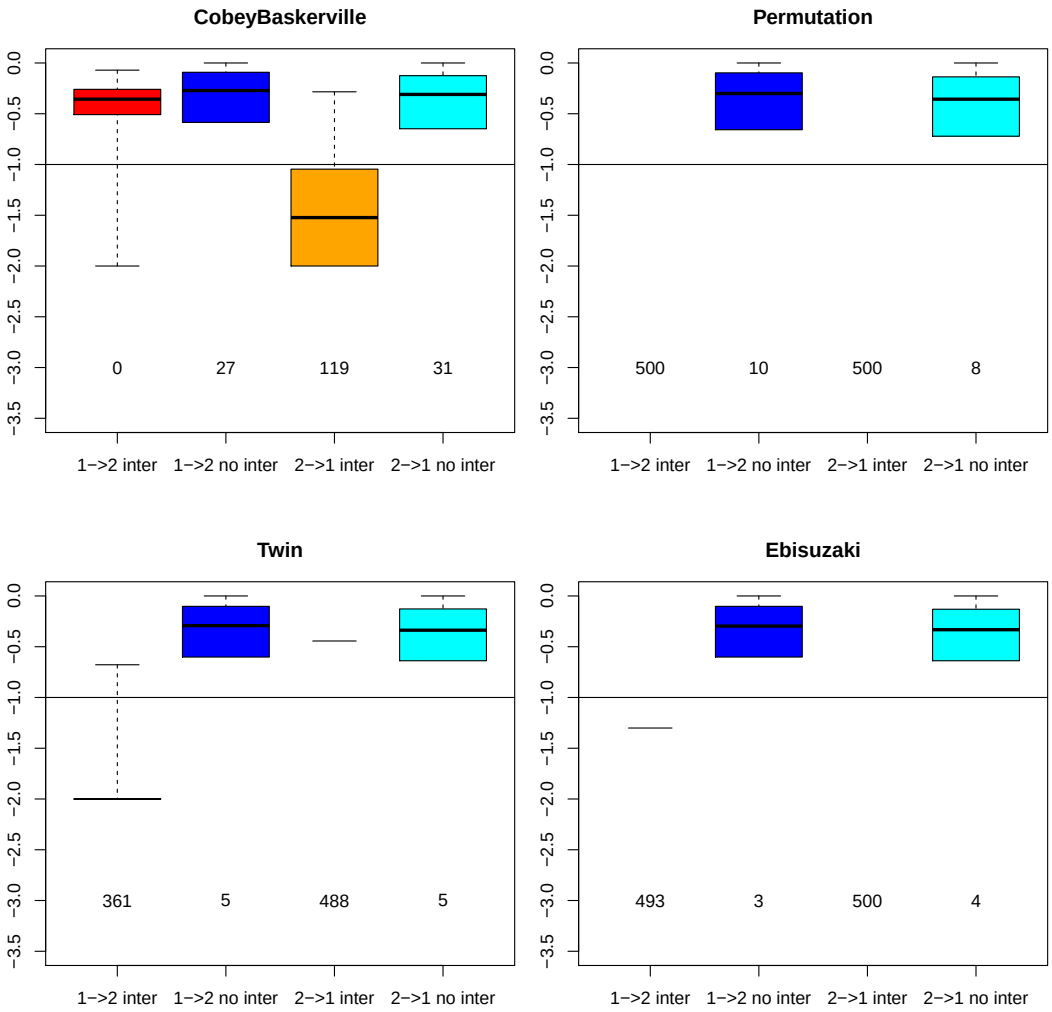

Figure 12: Comparison of different ways of computing p-values for the stochastic 2-species model.Permutation, twin and Ebisuzaki are different ways of computing surrogate time series for the p-vales. The number of p-values which are found to be 0, among the 500 simulations estimated, is written at the bottom of the p-value boxplot.

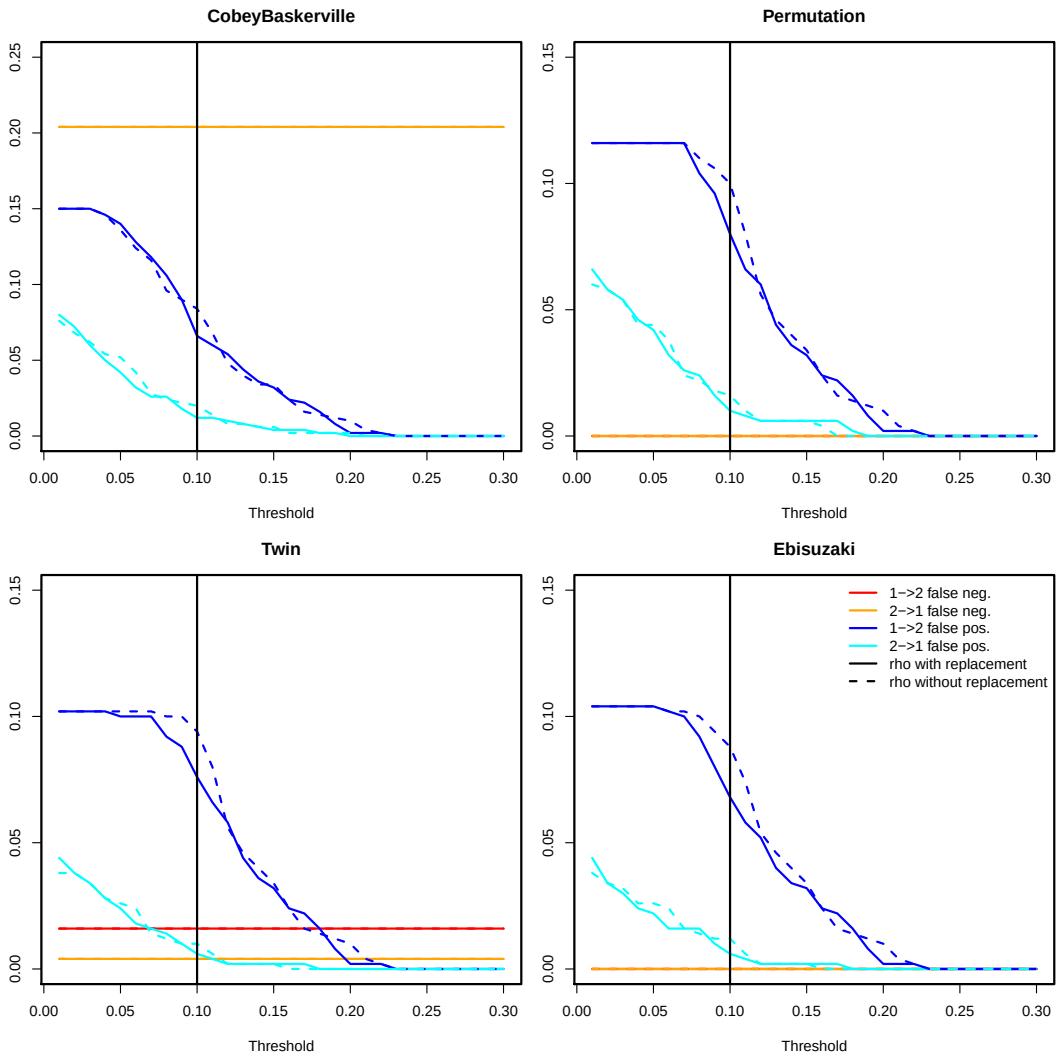

Figure 13: Comparison of false negatives and positives when combining p-values and thresholds on final ρ value.

Effect of log-transformation on CCM

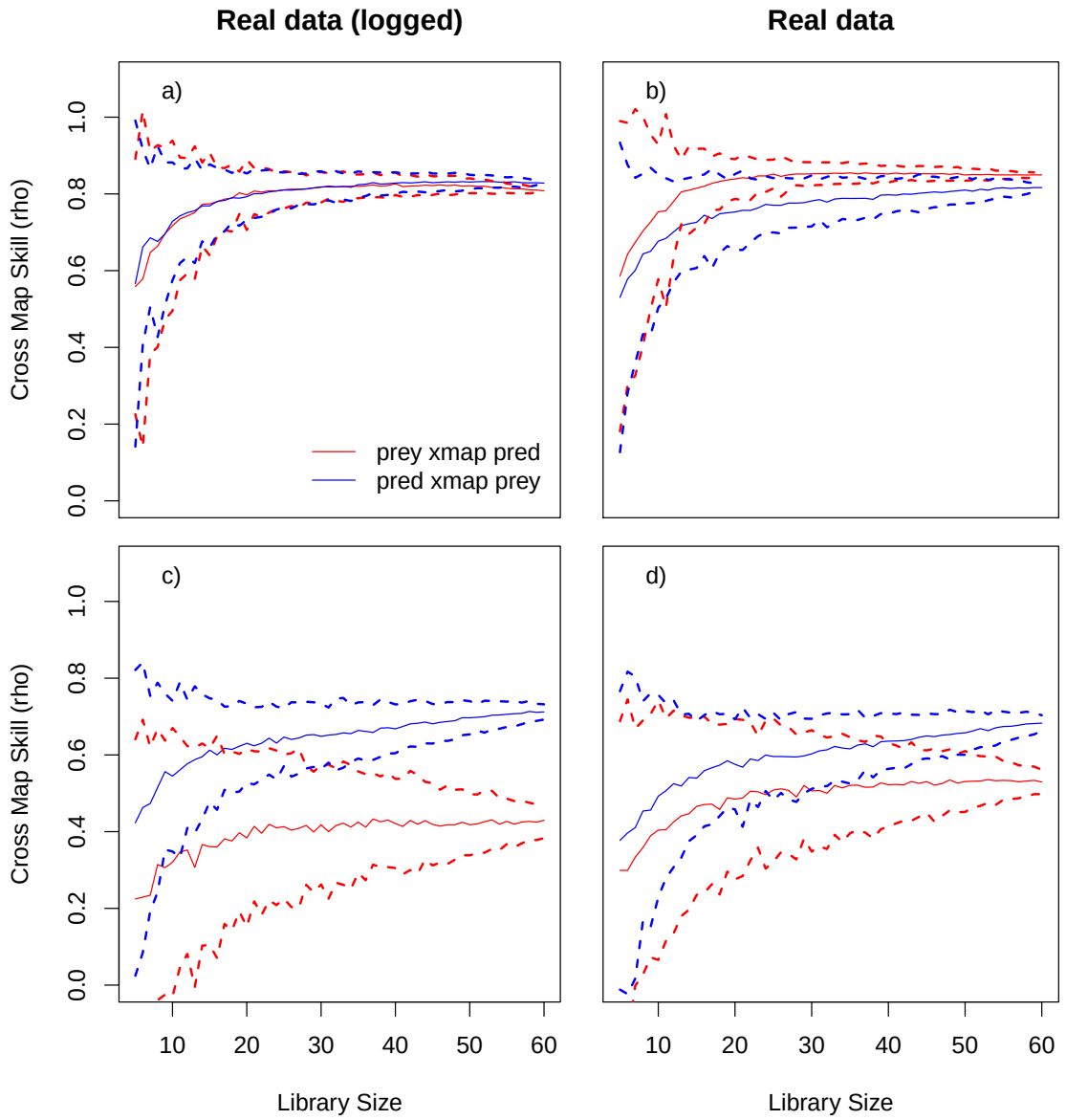

Figure 14: Convergent Cross-Mapping with (left) and without (right) ln-transform of the data for the Veilleux dataset

Lag order p selection in the MAR(p) framework, for the deterministic competition model

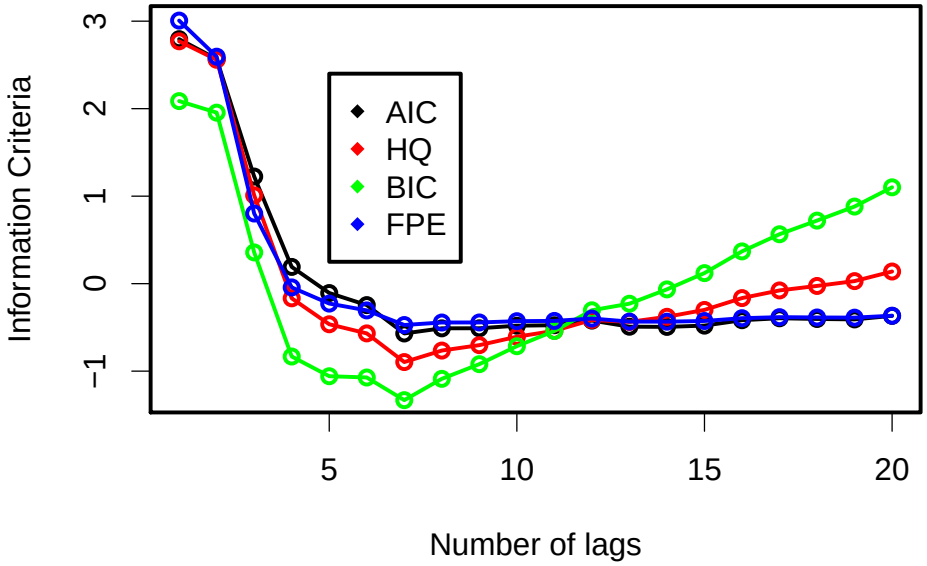

Figure 15: Results of model information criteria vs. lag order for the simulated deterministic competition model of eqs. XX

Spurious causality when including a confounding factor

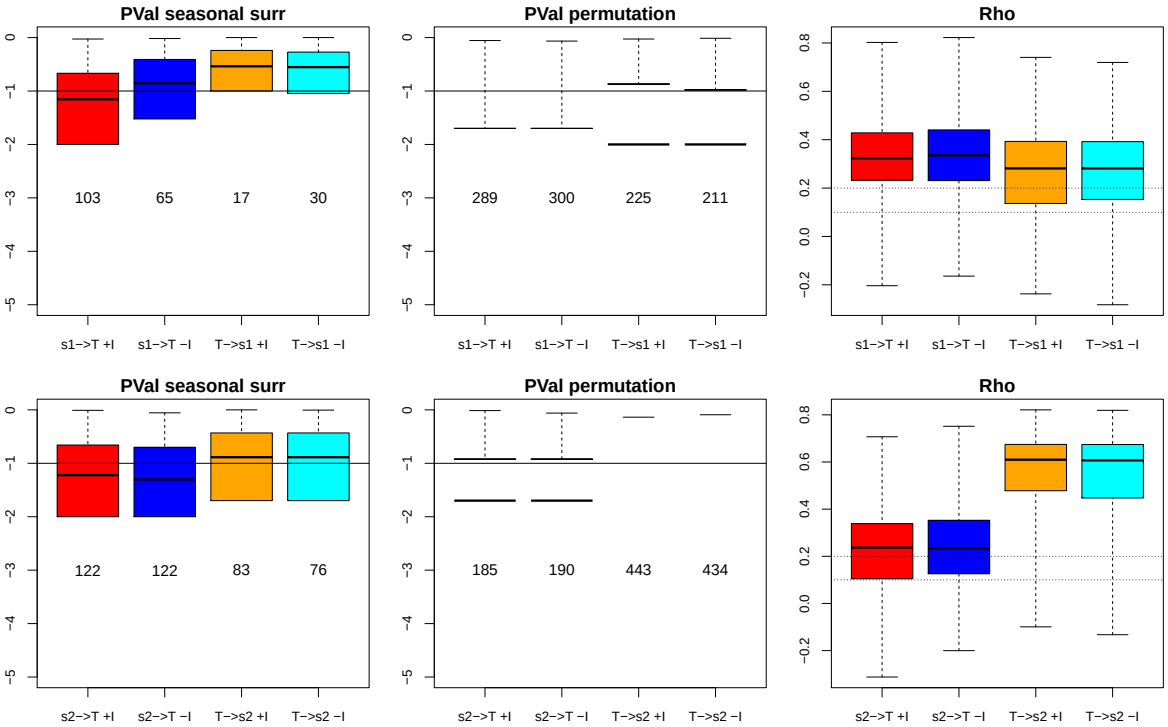

Figure 16: Comparison of log10(p-values) and CCM skill (rho) values to examine effects of temperature (T) on species 1 and 2 (s1 and s2), and the spurious reverse causality (species 1 or 2 causing temperature). Simulations were ran with (+I) and without (-I) interactions between species 1 and 2. Numbers below the boxplots indicate the number of times the p-value was 0, over 500 simulations. 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 (rho must be above the line for the causalityto be inferred)

Lag order selection for the 10- and 20-species model

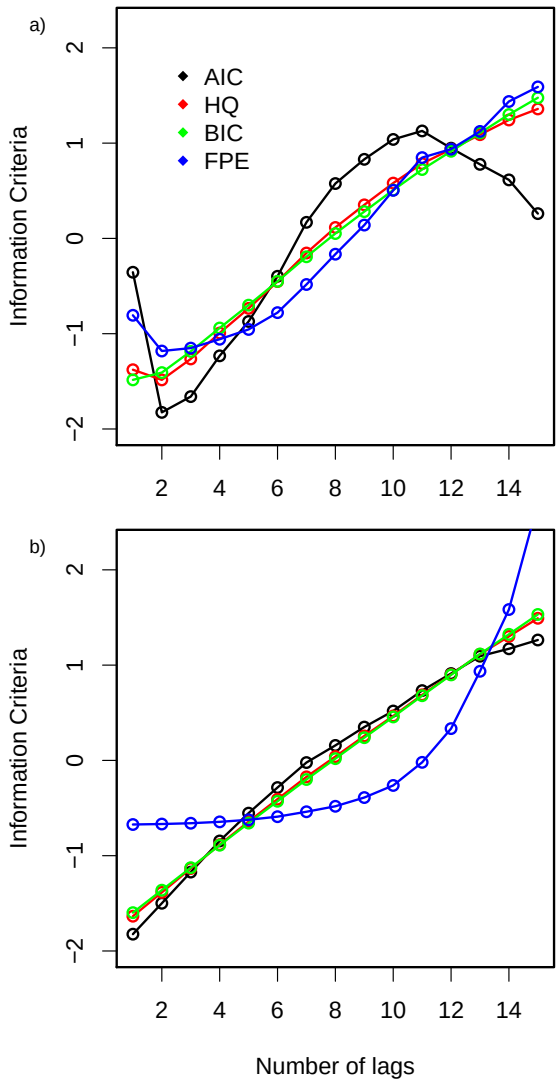

Figure 17: Lag order selection for (a) the 10-species and (b) one of the 20-species stochastic community model.

Interaction matrix for the 20-species model

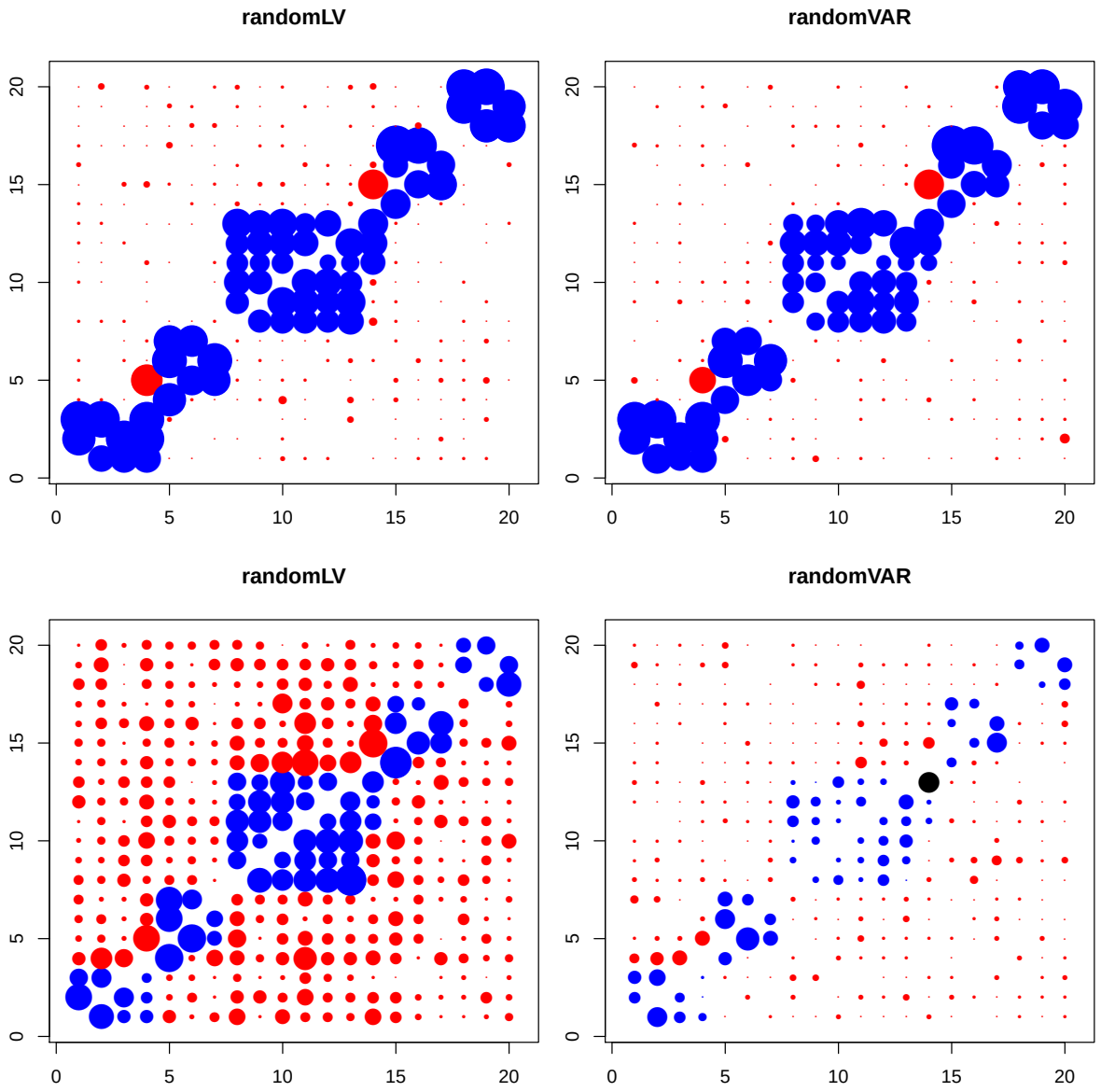

Figure 18: Interaction matrices obtained from pairwise GC (top) or 'shuffling' CCM (bottom) for 10-species communities. Blue circles are the true positives, red circles are false positives and black circles are false negatives. For the true and false positives, the size of the circles is proportional to the proportion of detection over 25 simulations.